СООБЩЕНИЯ
ОБЪЕДИНЕННОГО
ИНСТИТУТА
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ
ДУБНА

Р10 - 8622

Д.Д. Арнаудов

СТРАТЕГИЯ ПОИСКА В ИПС ОИЯИ

1975

P10 - 8622

Д.Д. Арнаудов

СТРАТЕГИЯ ПОИСКА В ИПС ОИЯИ

Арнаудов Д.Д. Р10 - 8622

Стратегия поиска в ИПС ОИЯИ

Описаны созданная стратегия поиска информации в проектируемой ИПС ОИЯИ, разработанный вариант трехэтапного поиска в системе дескрипторного типа, реализованная обратная связь с потребителем. Работа выполнена в ЛВТА.

Сообщение Объединенного института ядерных исследований
Дубна 1975

Arnaudov D.D. P10 - 8622

Search Strategy in the JINR Information Retrieval System

A created information search strategy in a projected JINR-IRS is given consideration as well as a constructed variant of three level search in a descriptoral type system and a realized user feedback. The work is performed in the Laboratory of Computing Techniques and Automatization.

**Communication of the Joint Institute for Nuclear Research
Dubna 1975**

Информационно-поисковая система (ИПС) - это система, предназначенная для информационного поиска, хранения и обновления информации. Под "информационным поиском" понимается область обработки данных, включающая совокупность ряда арифметических, логических операций, ввод -вывод , конечной целью которых является выявление по заданным признакам отбора всех данных (в виде документов, фактических справок), содержащих требуемую информацию и представляющих собой ответы на вводимую информацию.

Анализ работы поисковых систем /I/ показывает, что смысл выдаваемых ими документов определенным образом относится к смыслу запроса. В этой связи каждая ИПС ориентирована на определенный критерий смыслового соответствия. Для большинства ИПС этот критерий приблизительно одинаков. Наличие или отсутствие в документе какой-либо информации сверх того, о чем говорится в запросе, как правило, не влияет на решение вопроса о выдаче документа. Но если в документе отсутствует существенная часть информации, упоминаемой в запросе, то такой документ обычно не выдается. Не выдаются, как правило, и документы, в которых содержится информация более общая, чем та, о которой идет речь в запросе.

Уместно заметить, что чем менее формализована работа поисковой системы, тем больше творческого участия в ней принимает человек, тем менее определен критерий смыслового соответствия. Поэтому обычно в описаниях поисковых систем критерий смыслового соответствия не выделяется в качестве самостоятельной категории рассмотрения. Но в связи с постановкой задачи автоматизации основных поисковых процессов и, в частности, процесса установления смыслового соответствия, стало необходимым специальное рассмотрение средств, с помощью которых поисковая система устанавливает смысловое соответствие между документом и запросом.

Средства, которыми располагает ИПС для реализации критерия смыслового соответствия, на который она ориентирована, и методику использования этих средств будем называть стратегией ИПС, или, короче, стратегией поиска.

Таким образом, стратегия поиска включает в себя три основных элемента: логико-семантические отношения, правила сравнения и методику использования логико-семантических отношений и правил сравнения.

а) Логико-семантические отношения.

Они не зависят от конкретных текстов. Например, они могут быть между словами поискового языка (так, у INIS – это отношения подчиненности между дескрипторами, родово-видовые отношения, включая весовые коэффициенты дескрипторов, и т.д.)

Подобные отношения не являются обязательным элементом ИПС и в некоторых системах отсутствуют.

б) Правила сравнения.

Они определяют процедуру сопоставления отображений запроса и документов на поисковом языке. В результате сравнения определяются

документы, подлежащие выдаче. Во многих поисковых системах правила сравнения таковы, что выдаче подлежат те документы, поисковые образы которых полностью включают в себя поисковый образ запроса.

в) Методика использования логико-семантических отношений и правил сравнения.

Эта методика определяет последовательность работы с отдельными частями информационного массива при реализации правил сравнения.

Необходимо отметить, что процесс установления соответствия является неотъемлемой частью любого процесса обработки, выполняемого машиной. Особенность же этого процесса, как части информационного поиска, заключается в необходимости определения последовательности правил сравнения. Под этим мы будем понимать последовательно усложняющийся анализ элемента с постоянным 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 массив поисковых образов документов, применять самый сложный анализ к каждому элементу оказывается слишком дорогостоящим делом /2/. Следовательно, одним из аспектов методики использования логико-семантических отношений и правил сравнения является установление правил последовательно усложняющегося анализа.

Особое внимание следует уделить вероятности пропуска нужного элемента, в связи с тем, что он может быть отброшен при грубом анализе, в то время как сложный анализ его бы принял. Этот вопрос усложняется также очень реальными трудностями при описании: а) вероятные ошибки; б) не все данные, составляющие полное описание, могут быть или будут включены; в) могут быть упущены важные взаимосвязи между дескрипторами. Обычно эти трудности устраняются на базе

использования логико-семантических отношений между дескрипторами, путем введения "весов", отражающих относительную важность или степень взаимосвязи между дескрипторами и т.п.

Имея ввиду основные теоретические положения стратегии поиска, а также принятые принципы организации ИПС ОИЯИ[3] и структурно-функциональную организацию информационных массивов [4], приведем основные характеристики стратегии поиска ИПС ОИЯИ.

I) Процесс поиска характеризуется своей иерархической структурой (три уровня).

2) На различных этапах поиска производится пакетная обработка запросов, состоящих из набора дескрипторов, связанных между собой операторами булевой алгебры ("и", "или", "нет").

3) Благодаря гибкости принятой структуры в процессе установления соответствия осуществляются параллельные поиски на основе совместной обработки информационных массивов и накопление поисковых требований.

4) За счет сегментации однородных списков производится сокращение их средней длины, а отсюда - и времени поиска.

5) В системе реализован гибкий критерий соответствия, позволяющий совершать поиск по набору дескрипторов, связанных операторами булевой алгебры, что дает возможность:

а) производить поиск по наборам дескрипторов на неполное совпадение;

б) производить поиск по наборам дескрипторов на полное совпадение;

в) производить поиск, учитывая значимость дескрипторов (с взвешиванием);

г) производить поиск с "уточнением", используя более "узкие" и более "широкие" дескрипторы;

д) производить поиск по характеристикам, которые не являются дескрипторами.

Все эти характеристики стратегии поиска реализуются как отдельные модули поисковой процедуры. Конечной целью процедуры поиска является выявление по заданным признакам отбора всех данных (в виде библиографии документов), содержащих требуемую информацию и представляющих собой ответы на вводимые запросы.

Фактически сама процедура поиска формально состоит из ряда логических, арифметических операций и операций ввода-вывода, совершаемых в основном над определенной структурно-функциональной организацией основных поисковых массивов. Существенная часть этих операций относится к работе с периферийными устройствами (магнитные диски, магнитные ленты и т.п.), где расположены соответствующие информационные массивы.

Стратегию поиска данной информационной системы нельзя рассматривать отдельно от структурно-функциональной организации массивов этой системы. Более того, эти два понятия органически связаны и представляют, в сущности, основу разрабатываемой поисковой системы. В информационном отношении стратегия поиска представляет некую процедуру, реализующую данный критерий смыслового соответствия между дескрипторами запроса и поисковыми образами документов.

В связи с этим основная цель разработчика состоит в том, каким образом реализовать эту процедуру, чтобы получить минимальное время на запрос или на группу запросов.

Большинство информационно-поисковых систем работают эффективно как в пакетном режиме, так и в режиме реального времени с дистанци-

онно-расположенных терминалов. Сам по себе режим реального времени не является противоречием пакетного режима, т.к. при наличии в данное время работающих терминалов, работа системы сводится к обработке некоторого списка (пакета) поступающих в систему запросов. Поэтому в дальнейшем при раскрытии стратегии поиска мы будем иметь в виду обработку группы запросов.

При наличии большого количества документов в системе (так, например, в системе ОИЯИ оперативный фонд ее около 5000000) основная цель стратегии поиска – это установление последовательности работы процедуры с отдельными частями массивов при раскрытии критерия смыслового соответствия и в то же время получение эффективного поиска. В связи с этим возникает задача разработки стратегии поиска, базирующейся на некотором алгоритме, ограничивающем область проведения поиска небольшими подмножествами исходного массива документов и в то же время не усложняющем обработку запросов различного типа.

В ИПС ОИЯИ это достигается на базе параллельной обработки цепных списков, расположенных в зонах основных массивов при иерархическом трехуровневом поиске.

Иерархический (трехуровневый) поиск экспериментирован нами еще в 1969 году /5/. Как показали наши исследования, главная цель иерархии поиска – это постепенное сужение величины основного массива поисковых образов документов при раскрытии критерия смыслового соответствия. С другой стороны, модульная структура трехуровневого поиска дает возможность эффективно проводить поиск по гибкому критерию соответствия (не только на полное совпадение дескрипторов запроса с дескрипторами поисковых образов документов), включающему поиск на неполное совпадение, введение отрицательных дескрипторов с "взвешиванием" и т.п.

Более конкретно остановимся на отдельных этапах трехуровневого поиска.

Как уже отметили в /4/, в структурно-функциональной организации имеется три поисковых массива – ОМПОД, МЗД, ОМД. С точки зрения теории файла /6/ эти три массива можно рассматривать как один информационный массив (ОМПОД) с соответствующей ему управляющей частью (массивы МЗД, ОМД). Сущность поиска состоит в исполнении правил сравнения при последовательном трехэтапном движении по определенным массивам управляющей части и основного информационного массива. При этом на всех трех этапах производится поиск на полное совпадение положительных дескрипторов (дескрипторы без знака отрицания). После нахождения множества релевантных поисковых образов документов на третьем этапе производится "отсеивание" тех документов, в которых находится отрицательный дескриптор (дескриптор со знаком отрицания). В данном случае имеется в виду, например, что если А отражает свойство "дисперсный", то ¬А ("не" А) означает свойство "недисперсный". Подобное использование булевского оператора "отрицания" не сопровождается никакими неблагоприятными последствиями, независимо от возможных двух способов реализации:

а) Отрицательное свойство принимается независимым от любого другого соответствующего свойства, так что свойство "недисперсный", не является логическим дополнением свойства "дисперсный", а рассматривается просто как другое свойство.

в) Отрицательное свойство рассматривается в качестве логического дополнения соответствующего положительного свойства, и всем запросам в каждом конкретном случае приписываются либо положительные, либо отрицательные дескрипторы.

В первом из указанных случаев некоторые запросы могут быть снабжены дескрипторами A, некоторые ¬A , однако дополнением множества, отмеченного A, является не множество, соответствующее ¬A , а множество, не имеющее дескриптора A. Другими словами, если известно множество, соответствующее A, ничего определенного не может быть сказано относительно множества, отмеченного ¬A . Аналогично, при известной поисковой эффективности одного свойства нельзя сделать вывода об эффективности другого. Дескриптор со знаком ¬A в этом случае может быть заменен некоторым новым дескриптором, скажем B, который будет рассматриваться как независимый термин в дескрипторном словаре.

Это, на наш взгляд, недостаточно удобно, т.к., имея уже некоторые статистические характеристики положительного дескриптора, нельзя их использовать для соответствующего ему отрицательного дескриптора.

Поэтому мы рассматриваем булевский оператор отрицания в смысле операции дополнимости в отношении множества A и его дополнения ¬A. При этих условиях множество, соответствующее свойству "дисперсный" или "недисперсный", эквивалентно множеству образов документов со свойствами, скажем, "реактивный" или "нереактивный" (так как $A \vee \bar{A} = B \vee \bar{B}$ и $A \wedge \bar{A} = B \wedge \bar{B}$). Это, в свою очередь, имеет большое значение на допоисковом этапе предварительной оценки, когда, используя накопленную статистику о дескрипторах запроса, можно прогнозировать ожидаемое количество релевантных запросу документов.

На третьем этапе поиска реализуется и обратная связь с потребителями. Она сводится фактически к налаживанию взаимосвязи потребителя с поисковой системой и улучшению поисковой эффективности путем непосредственного использования предоставляемой потребителем информации. Это предполагает наличие первоначальной формулировки

поисковых запросов и предусматривает на их основе новые более совершенные формулировки. Получаемые при этом более совершенные переформулировки будут более соответствовать нуждам потребителей и, следовательно, приводить к лучшим поисковым результатам. Пересмотренные запросы составляются самими потребителями вручную с использованием разнообразных вспомогательных средств или же системой автоматически, на основе получаемой от потребителей информации.

В первом случае, когда переформулировка запросов осуществляется потребителем, система выступает в качестве автоматического словаря, обеспечивая потребителей информацией словарного типа о дескрипторах, используемых в формулировках запросов. В системе ОИЯИ даются сведения о "более узких" или "более широких" терминах, соответствующих данному дескриптору.

Необходимо добавить, что на основе найденного на третьем этапе множества документов можно выдать потребителю дескрипторы, встретившиеся одновременно в нескольких из них, но не присутствовавшие в исходной формулировке запроса. Кроме того, можно указать и степень поисковой эффективности каждого из дескрипторов; на наш взгляд, это абсолютное число индексируемых им документов. В этой связи следует предположить, что поисковая сила дескриптора обратно пропорциональна частоте его приписывания документам массива, так как в большинстве случаев специфичность высокочастотных терминов в соответствующих им контекстах невысока, тогда как многие низкочастотные термины могут отражать основное содержание документа.

В этих условиях является логичным требование такого словарного обеспечения потребителя, при котором наряду с представлением ему соответствующих дескрипторов указывается частота их приписывания документам массива. Тогда, повышая веса терминов, предполагаемых

как менее полезные с точки зрения их поисковой силы, потребитель может использовать эту информацию для совершенствования поискового запроса. В ИПС ОИЯИ допускается и автоматическое вычисление веса данного дескриптора на базе накопленной статистики. Тогда, по желанию потребителя на основе некоторого уже полученного по запросу множества документов система может провести поиск, учитывая значимости дескрипторов, приписывая им веса, получаемые автоматическим путем. При этом в ответе все документы будут ранжированы от самого релевантного "до "менее релевантного".

В этом отношении в дальнейшем стратегию поиска можно обогатить, используя метод обратной связи по релевантности /7,8/. Сущность метода заключается в проведении первоначального поиска и предоставлении потребителю определенной части найденной информации. Потребитель изучает некоторые из найденных документов и в соответствии со своей целью поиска оценивает каждый документ как релевантный или нерелевантный. Оценки потребителя о релевантности возвращаются потребителю и автоматически используются для совершенствования первоначального поискового предписания путем повышения роли дескрипторов повышением их весовых коэффициентов, присутствующих в релевантных документах, и аналогичного понижения роли дескрипторов, встречающихся в документах, определенных как нерелевантные.

Измененные запросы могут быть снова введены в систему для проведения вторичного поиска по новому поисковому предписанию. При условии нормальной работы системы получают дополнительное число релевантных документов или, по крайней мере, более высокую степень их корреляции с измененным запросом по сравнению с исходным. Вновь найденные документы снова изучаются потребителем с последующим использованием результатов изучения в отношении степени их релевантности

для повторного переформулирования запросов. Таким образом может быть выполнено несколько итераций до тех пор, пока потребитель не будет удовлетворен результатами поиска. Упрощенная схема такого процесса приведена в /I/.

В стратегии поиска системы ОИЯИ предусмотрена и процедура поиска по характеристикам, которые не являются дескрипторами (например, год издания, имя автора, предметная категория и т.п.). Имеются четырнадцать таких характеристик. Эти характеристики являются логическими полями в массиве документов. Но проводить поиск по этим характеристикам в массиве документов - означает расходовать большое количество времени, так как оперативное хранилище этого массива содержит 400000 документов. Поэтому для более эффективного поиска предусмотрена комбинация поиска по дескрипторам и характеристикам. Пользователь описывает несколькими дескрипторами сущность своего требования и тем самым ограничивает поиск по характеристикам в данном подмножестве информационного массива.

Выделение характеристик как особых средств для ведения поиска имеет большое значение для повышения его эффективности. Поскольку нахождение определенных образов документов по дескрипторам требует "совпадения" дескриптора поискового образа с дескриптором запроса, то над дескриптором запроса в процессе поиска нельзя совершать других операций, кроме операций сравнения, в то время как характеристика может подвергаться арифметическим и логическим тестам (таким как равенство, больше, меньше, в пределах, включение во множество и т.д.). Эти же операции, конечно, можно заложить и в выражение для запроса и использовать данные характеристики как дескрипторы, но это усложняет проведение поиска на первых двух этапах.

Так как в библиографических системах при поиске особое значение имеет имя автора, в стратегии поиска системы ОИЯИ предусмотрен

поиск в специальном автоматическом авторском указателе. После нахождения методом прямого доступа имени автора, в авторском указателе по адресу связи воспроизводится доступ в соответствующий подмассив, где расположены адреса библиографических записей трудов данного автора.

Все до сих пор рассмотренные процедуры (за исключением поиска в автоматическом авторском указателе) совершаются как дополнение к основному поиску на полное совпадение дескрипторов запросов с дескрипторами поисковых образов документов, и основная их цель - - повышение эффективности поиска с точки зрения полноты и точности. Они осуществляют обратную связь с потребителем на третьем этапе иерархического поиска.

Основная процедура поиска в системе - это поиск на полное совпадение. При этом принятая стратегия трехуровневого поиска по группе запросов характеризуется способностью генерировать разбиения по мультисписку /9/. Это понятие имеет более сложную структуру, чем простое разбиение. В случае одного признака (ключа) разбиение зависит только от структуры массива, в то время как разбиение по мультисписку генерируется в процессе поиска, согласно булевому выражению запроса. Поэтому работа процедуры поиска на полное совпадение вносит в разработку стратегии критический элемент - время ответа.

Именно, минимизация времени ответа представляет самую большую трудность при разработке стратегии. Остановимся более подробно на этом вопросе.

В /4/ описаны значения логических полей записей массивов ОМД, МЗД, ОМПОД. Там отмечено, что поисковый образ документа представляет набор дескрипторов. Для каждого дескриптора организованы цеп-

ные списки в пределах данной зоны. Величина зоны - половина цилиндра дискового устройства. Тем самым получаем модификацию мультисписковой организации, описанной нами в /9/, при этом эффективно используется архитектура внешнего запоминающего устройства, так как, если считывающая головка установлена на данный цилиндр, то поиск по некоторому списку производится без дальнейших передвижений головок.

На рис.I показан фрагмент основных массивов. В МЗД накапливаются заголовки соответствующих цепных списков ОМПОД. Экспериментально проверено (на массиве из IOOOO документов), что для данного дескрипторного словаря и тематического фонда системы ОИЯИ отношения МЗД: ОМПОД =I:6. Следовательно, в одной зоне МЗД собираются заголовки из шести зон ОМПОД. Кроме того, в каждой зоне МЗД все заголовки, принадлежащие данному дескриптору, соединены в цепочке. В массиве ОМД для каждого дескриптора указаны все зоны МЗД, где он имеет заголовки. Следовательно, отношение ОМД:МЗД =I:6. Так как ОМПОД заполняется поисковыми образами документов в хронологическом порядке, то и соответствующие заголовки в ОМД и МЗД расположены также в хронологическом порядке.

Стратегия поиска на полное совпадение состоит в том, что на первом этапе поиска в массиве ОМД находятся общие зоны для всех дескрипторов запроса в массиве МЗД. Общие зоны - это такие зоны, в которых все дескрипторы данного запроса имеют заголовки. Если хоть один из дескрипторов запроса не имеет заголовка в данной зоне, то она не считается "общей". Если нет ни одной общей зоны, то запрос аннулируется еще в первом этапе поиска.

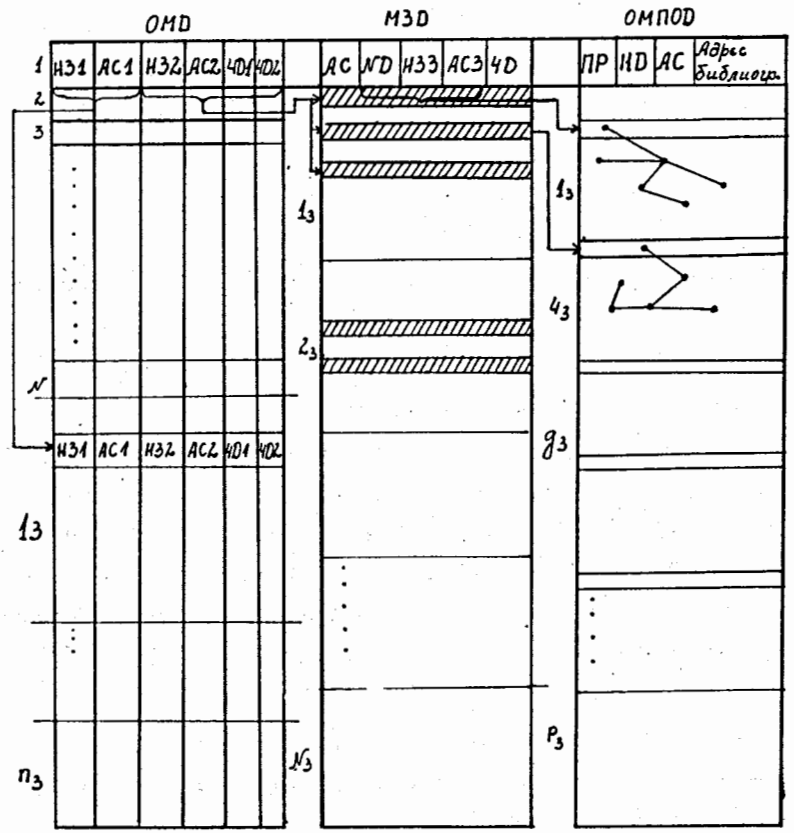

Рис.I.

На втором этапе поиска среди заголовков уже выбранных общих зон для каждого запроса ищутся общие диски и общие на диске зоны в массиве ОМПОД. В случае нахождения таких общих дисков и зон, для

каждой зоны (для каждого запроса) выбираются наикратчайшие цепные списки.

На третьем этапе обследуются для каждого запроса в каждой общей зоне данные наикратчайшие списки на полное совпадение в каждом узле всех дескрипторов запроса с дескрипторами поискового образа документа.

При совпадении данный документ выдается потребителю, происходит выборка по адресу библиографии документа в массиве документов.

Для получения минимального времени ответа при использовании данной стратегии поиска необходимо минимизировать число доступов к магнитному диску. В /4/ показано, что это сводится к минимизации общего времени доступа, которое определяется по формуле:

$$t_в = t_п + t_о + t_{транс.} \qquad (I)$$

При чтении целой зоны в ОП это время равняется:

$$t_в = t_п + 30t_о = 75 + 30x25 = 825 \text{ мс.} \qquad (2)$$

При чтении только отдельного элемента зоны (для ОМД и МЗД - величиной в 2 слова, а для ОМПОД - около 20 слов) величина этого времени гораздо меньше:

$$t_{в_1} = t_п + \tfrac{1}{2}t_о + t_{транс.} \qquad (3)$$

При чтении данного элемента $t_{транс.}$ достаточно мало.

Следовательно,

$$t_{в_1} = t_п + \tfrac{1}{2}t_о \approx 87{,}5 \text{ мс, т.е.} \qquad (4)$$

приблизительно в десять раз меньше, чем $t_в$.

Имея в виду выражения (2) и (4), в стратегии поиска необходимо определить, в каких случаях требуется выборка всей зоны в ОП, а в каких случаях - выборка отдельных элементов при обследовании соот-

ветствующих цепных списков. Для каждого отдельного момента в процессе поиска, когда совершается шаг по цепи списка, поисковая система принимает решение о выборе одной из двух существующих альтернатив.

Это решение в простейшем случае базируется на числе логических доступов к диску и заключается в подсчете этого числа и его сравнении с критическим числом. После этого принимается соответствующее решение. Критическое число определяет максимальное число доступов к отдельным элементам массива.

Для одного случая это число (К) определяется выражением:

$$t_g = K t_{g_1}; \quad K = \frac{t_g}{t_{g_1}} \approx 10.$$

Следовательно, если необходимо производить больше чем IO доступов в данный момент реализации процедуры поиска, то более эффективно записать всю зону в оперативную память, а потом обрабатывать соответствующие цепные списки.

На первом этапе поиска (в массиве ОМД) имеется доступ к первой части ОМД для выборки первых членов цепных списков дескрипторов. Это производится автоматически по кодам дескрипторов, которые совпадают с индексами элементов в первой части массива. В зависимости от числа различных дескрипторов запросов и от их распределения по зонам (как первая, так и вторая части ОМД в информационном отношении состоят из определенного числа зон) при сравнении с критическим числом принимается решение о выборке всей зоны в целом или только отдельных элементов зоны.Все запросы обрабатываются параллельно. Для подсчета числа доступов к членам цепей списков в зонах П части массива ОМД используется число, показывающее количество членов цепи в данной зоне (ЧДI). Решение о доступе принимается на базе общего числа элементов списков дескрипторов всех запросов, которые имеют элементы в данной зоне.

В результате I-го этапа поиска получаются все общие зоны в массиве МЗД, где имеются заголовки данных дескрипторов запросов. В каждой такой зоне один и тот же дескриптор может иметь не по одному, а по несколько заголовков.

На П-ом этапе поиска необходимо просмотреть все эти заголовки и найти общие зоны в ОМПОД. Для этого нужно сделать выборки этих заголовков из зон МЗД, где они связаны в цепочку. В начале П-го этапа, т.к. имеются сведения о длине каждой цепи в данной общей зоне, можно вычислить число доступов и принять решение о методе доступа в каждой зоне, имея в виду параллельную обработку всех запросов. После этого производится сравнение выбранных заголовков для каждого запроса и выбираются самые короткие списки в каждой зоне.

На третьем этапе производится обследование наикратчайших списков. Так как имеется информация о длине каждой цепи, на этом этапе может быть принято решение о методе доступа при обработке этих списков. Рассмотрим более подробно сущность поиска на третьем этапе.

Пусть на П-ом этапе, например, для пакета из трех одновременно обрабатываемых запросов (Q_1, Q_2 и Q_3) выбраны следующие зоны, где все дескрипторы данного запроса имеют общие зоны.

Таблица I

запрос	зоны
Q_1	I,5,I2
Q_2	4,I2,I9,25
Q_3	I,9

После этого происходит сортировка по номеру зоны, причем для каждой определен самый короткий список для данного запроса.

Таблица 2

зона	запросы
I	Q_1 , Q_3
4	Q_2
5	Q_1
I2	Q_1 , Q_2
I9	Q_2 , Q_3
25	Q_2

Например, для запросов Q_1, Q_2, Q_3 имеем такую картину.

Таблица 3

Запрос	Зона	Наикратч. список	Длина списка	Адрес первого элемента
Q_1	I	B	3	4
Q_3	I	F	8	I
Q_2	4	C	2	3
Q_1	5	A	4	2

Из таблицы 3 видно, что при доступе к первой зоне необходимо обработать Q_1 и Q_3 с общей длиной (8+3) - одиннадцать элементов. Следовательно, необходимо записать всю зону в память ЭВМ, а потом обрабатывать соответствующие списки.

Число элементов зоны 4 и 5 меньше критического, поэтому необходимо осуществлять доступ к отдельным элементам, а не к целой зоне.

При этом, как видно из табл.2,3, идет последовательная обработка каждой зоны (каждого цилиндра), пока не будут обработаны все

списки в данных зонах. Время обработки пакета запросов на третьем этапе ограничивается просмотром общего количества членов наикратчайших списков в определенных зонах.

Из описанной стратегии поиска ясно, что по обращению из запроса полный поиск по всем спискам дескрипторов запроса, в общем случае, не производится.Кроме того, поиск всегда перемешан со списковыми поисками других запросов в соответствии с заложенной в системе стратегией поиска.

Модульный характер процедуры поиска и гибкость структурно-функциональной организации поисковых массивов позволяют осуществить поиск в реальном времени по принципу разделения времени на двух уровнях. На наиболее высоком уровне можно произвести разделение рабочего времени применительно к работе оперативной памяти ЭВМ.

В устройстве памяти могут находиться на исполнении до 50 запросов, а если их общее количество превышает 50, то запросы, являющиеся избыточными, будут циркулировать между определенной памятью и дисковым буферным устройством. Кроме того, обработку запросов в текущем порядке можно разделить во времени, используя три этапа поиска. Пока идет обработка на третьем этапе, новые запросы могут быть приняты и параллельно обрабатываться на I и П этапах поиска.

На более низком уровне можно производить разделение рабочего времени поиска после обращения к магнитному диску. Для этого нужно использовать управляющие списки , подобно тем, что показаны в табл. I,2,3, которые обеспечивают реализацию поиска по запросам по принципу разделения времени.

Применительно к каждому запросу (в табл.I рассматриваются три запроса) формируется список всех зон, в которых должна быть произведена выборка, для того чтобы обеспечить обслуживание этого процесса. Для обслуживания, например, запроса Q_1 должна быть произведе-

на выборка в зонах I, 5 и I2. Списки всех таких зон для всех запросов объединяются (см. табл.2) для образования объединенного списка выборки зон, определяющего план поиска, в соответствии с которым осуществляется разделение рабочего времени в целях обеспечения наиболее эффективного поиска информации, хранящейся в ЗУ на магнитном диске. Таким образом, из таблицы следует, что сначала осуществляется выборка зоны I путем подведения считывающей головки к определенному цилиндру. Массив в зоне имеет списковую структуру (см.рис.I) и для запроса Q_1 в этой зоне обследуется и обрабатывается специальный список (наикратчайший). После этого в зоне I аналогичным образом исполняется запрос Q_3 . После этого происходит перемещение считывающей головки к следующему цилиндру для исполнения запроса 2 (зона 4). Затем аналогичный процесс осуществляется по отношению к зонам 5, I2, I9 и 26, после чего на все запросы будут получены полные ответы. Если во время этого процесса поступил новый запрос, то производится немедленное формирование для него объединенного списка выборки зон и исполнение этого запроса начинается сразу же по достижении считывающей головкой ближайшей зоны.

В заключение отметим, что процесс поиска может быть усложнен с добавлением требования поиска по приоритету. Если, например, все подлежащие исполнению запросы имеют одинаковый приоритет, то тогда применяется указанная выборка зон (табл.2).

Если один или несколько запросов имеют более высокий приоритет, то обеспечивается сначала выборка зон, относящихся к исполнению этих запросов, а затем других зон в указанной последовательности.

ЛИТЕРАТУРА

I. Г.Сэлтон. Автоматическая обработка, хранение и поиск информации. М., I973, "Советское радио".

2. Meserovic M.D. "Multi-Level Systems and Information Problems", Preprints; First Congress on the Information Systems Sciences, Nov.,19,1962, USA.

3. Д.Д.Арнаудов. Выбор оптимальной структуры основного информационного массива ИПС для размещения на магнитном диске. Сообщение ОИЯИ, I0-7949, Дубна, I974.

4. Д.Д.Арнаудов. Структурно-функциональная организация основных информационных массивов ИПС ОИЯИ. Сообщение ОИЯИ PIO-862I, Дубна, I975.

5. Д.Д.Арнаудов. Об одном способе организации информационного массива и дескрипторного словаря в библиографической ИПС. Сб.Цифровая вычислительная техника и программирование, вып.7. I970, Сов.радио, М.

6. Д.Д.Арнаудов. Анализ методов доступа информации к внешним ЗУ на магнитных дисках при работе ИПС, реализованной на ЭВМ третьего поколения, препринт ОИЯИ, I0-7953, Дубна, I974.

7. Rocchio J.J. "Document Retrieval Systems-Optimization and Evoluction", Report ISR-10 to National Science Harvard -Computation Laboratory,March 1966,USA.

8. Salton G. "Search Strategy and the Optimization of Retrieval Effectiveness", FID/IFIP Congress on Mechanized Documentation. Rome 1967, p(151-173).

9. Д.Д.Арнаудов. Организация мультисписочных узловых структур и их программная реализация на КОБОЛе, сообщение ОИЯИ, Р0-7587, Дубна, 1973.

10. Д.Д.Арнаудов. Выбор оптимальной структуры основного информационного массива ИПС для размещения на магнитном диске, сообщение ОИЯИ Р0-7949, Дубна, 1974.

Рукопись поступила в издательский отдел
19 февраля 1975 г.