СООБЩЕНИЯ
ОБЪЕДИНЕННОГО
ИНСТИТУТА
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ
ДУБНА



Р11 - 10401

Д.Д.Арнаудов

# О ПРИМЕНЕНИИ МЕТОДА НЕГАТИВНЫХ СТРАТЕГИЙ В ИНФОРМАЦИОННО-ПОИСКОВОЙ СИСТЕМЕ С ИЕРАРХИЧЕСКОЙ ЧАСТИЧНО ИНВЕРТИРОВАННОЙ СТРУКТУРОЙ

1977

Р11 - 10401

Д.Д.Арнаудов

# О ПРИМЕНЕНИИ МЕТОДА НЕГАТИВНЫХ СТРАТЕГИЙ В ИНФОРМАЦИОННО-ПОИСКОВОЙ СИСТЕМЕ С ИЕРАРХИЧЕСКОЙ ЧАСТИЧНО ИНВЕРТИРОВАННОЙ СТРУКТУРОЙ

Арнаудов Д.Д.

Р11 - 10401

О применении метода негативных стратегий в информационно-поисковой системе с иерархической частично инвертированной структурой

Рассматривается применение метода негатнвных стратегий для поиска документов в большом информационном массиве. Данный метод дает возможность исключить из рассмотрения определенные множества с бесполезным перебором, что значительно уменьшает число зон исследуемого массива.

Работа выполнена в Лаборатории вычислительной техники и автоматизации.

Сообщение Объединенного института ядерных исследований. Дубна 1977

Arnaudov D.D.

P11 - 10401

A Method of Negative Strategies in the Work of a Hierarchical Partially Inverted Information Retrieval System

A method of negative strategies during the information retrieval is discussed. It is of a great importance when a large file is retrieved. This method gives the opportunity to annulate certain groups of documents which decreases the retrieved segments of the information file.

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

Communication of the Joint Institute for Nuclear Research. Dubna 1977



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 более общая информация, 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.

Сам процесс установления соответствия является неотъемлемой частью любого процесса обработки, выполняемого машиной. Особенность же этого процесса, как части информационного поиска, заключается в необходимости определения последовательности применения 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. Следовательно, одним из аспектов методики раскрытия критерия смыслового соответствия является установление правил последовательно усложняющегося анализа.

Одним из способов проведения подобного анализа является меметод негативных стратегий. Сущность этого метода заключается в предварительном отсеве нерелевантных документов, что дает возможность исключить определенные подмножества информационного массива с бесполезным перебором. В этой работе рассмотрим метод негативных стратегий, исходя из структурно-функциональных особенностей иерархической частично инвертированной ИПС, а также из возможностей внешних носителей на магнитных дисках, где в общем случае располагается информационный массив.

Пусть массив представляет некоторую последовательность из $F$ символов, и каждый отдельный элемент массива состоит из $N$ символов / следовательно, в массиве имеется $\frac{F}{N}$ элементов/. Предположим, что $y$ есть запрос, и соответствие между запросом и элементом массива устанавливается по заданным $n_t$ символам из $N$/ см.рис.I /.

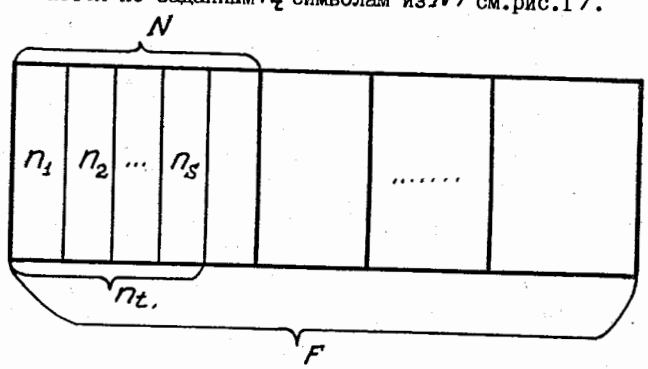


рис.I

Следовательно, для установления критерия смыслового соответствия при поиске необходимо просмотреть весь массив и совершить



соответствующие операции сравнения по $n_t$ символам. Если, однако, мы будем проводить сравнение не по $n_t$ символам, а выделим некоторую группу из $n_i$ символов таким образом, что $\sum_{i=1}^{S} n_i = n_t$, то тогда число сравнений будет значительно меньше, и сразу отсеются все фрагменты массива, где $n_3 \neq n_M$ / $n_3$ – соответствующие символы запроса, $n_M$ – символы элемента массива/. Таким образом, из рассмотрения исключаются те части массива, где уже априори ясно, что наверняка не встретится релевантный запросу документ.

Для тех элементов, которые пройдут данный отсев при одном значении $n_i$, производится дальнейший отсев при следующем значении $n_i$ и так далее, так что в конце концов останутся только те элементы, которые прошли проверку по всем $n_t$ символам.

Из предложенной идеи ясно, что в таком случае весь массив будет состоять фактически из некоторого количества подмножеств, соответствующих конкретным $n_i$. Интуитивно ясно, что чем больше массив, тем больше количество последовательных групп $n_i$.

В связи с "отображением" структуры массива на физическое устройство [2] можно рассматривать данные $n_t$ символов для раскрытия смыслового соответствия как некоторую последовательность из $n_i$ групп, определяющих в тоже время и некоторые элементы принятого логического формата устройства, где располагается массив. В таком случае $n_t$ символы могут быть логически разбиты на части, по $n_i$ символов в каждой, причем любое $n_i$ может и не равняться $n_{i+1}$ *(для* $i=1,2,...S$*)*. В общем случае, при расположении информационного массива более чем на одном устройстве можем иметь следующее разбиение данных $n_t$ символов:

$$n_t = \{ND, NZ, NE\}, \text{ где}$$

$ND$ – количество символов, указывающих номер дискового устройства;

$NZ$ – количество символов, указывающих номер зоны данного дискового устройства;

$NE$ – количество символов, указывающих номер элемента данной зоны.

Разбиение массива на подходящие подмножества дает возможность применить теорию негативных стратегий для поиска релевантных документов. Особенно эфективно эти стратегии могут быть исполь-

зованы при реализации поиска в ИПС с частично инвертированной структурой.

Необходимо заметить, что, анализируя методику организации структур информационных массивов, которые характерны для поисковых систем, где в общем случае поиск осуществляется по набору дескрипторов /ключей/, Лефковиц /3/ отмечает, что можно выделить два полюса в организации массивов – инверсную организацию и мультисписковую /ассоциативно-адресную/. Все другие организации можно рассматривать как варианты этих двух. При этом имеем спектр различных частично инвертированных схем.

На рис.2 графически изображена эта тенденция.


Рис.2

Как инверсная, так и мультисписковая организации представляют различные варианты структуры массива, реализация которых приводит к различным особенностям программирования. Однако обе организации могут представлять одну и ту же информационную структуру: иерархическую или мультисписковую. Мы показали в работе /4/, что, комбинируя инверсную и мультисписковую организации, применяя различные варианты прямого доступа /5/, можно создать эффективно функционирующую ИПС с частично инвертированной структурой. В основе методики построения такой структуры лежит классическая мультисписковая организация /6/. Она может модифицироваться в сторону "инверсности". Это необходимо, когда информационный массив велик и дескрипторные списки достаточно длинны, отчего поиск значительно замедляется /напр., когда информационной массив занимает больше одного дискового устройства/, см.рис.3.



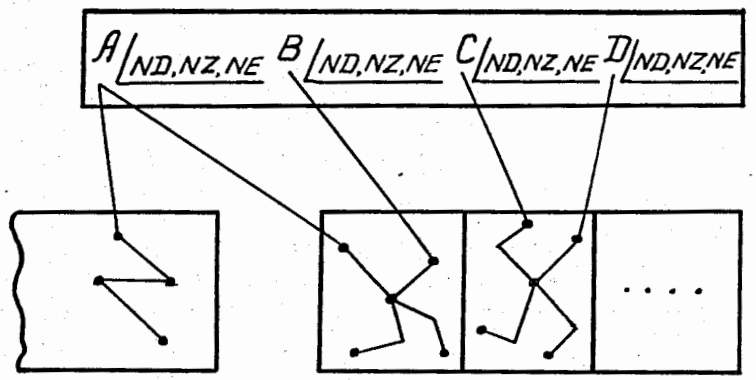


рис.3

В таком случае, с целью уменьшения времени ответа /4/, необходимо произвести некоторые изменения в уровнях структуры, которые сделали бы ее более "инверсной". Эти изменения связаны с введением понятия "зона" в логическом формате устройства, и сегментацией списков в пределах данной зоны. В таком случае конкретный дескриптор может иметь уже не один, а множество списков, находящихся в различных зонах дискового устройства. Тогда в управляющей части массива появятся заголовки соответствующих списков. Возможная структура массива показана на рис.4. Цифры в скобках указывают на адреса {ND,NZ,NE} начал соответствующих узловых списков.

Введение сегментации информационного массива по дисковым устройствам /МД/, а в пределах данного диска – по зоном /о выборе величины зоны подробно рассказано в работе /4/, дает возможность эффективно применить метод негативных стратегий. В этом случае при работе процедуры поиска с конкретным запросом /с группой запросов/ сначала "отбрасываются" такие "диски" и "зоны", где наверняка не встретятся релевантные запросу документы, а потом, уже в ограниченном множестве массива, производится сравнение с дескрипторами поисковых образов наикратчайших узловых списков /I/. Проведенные исследования показывают /3,4/, что при достаточно большом массиве исследуются в среднем/после применения метода негативных стратегий/ не более, чем I0 % от общего количества зон, принадлежащих цеп-

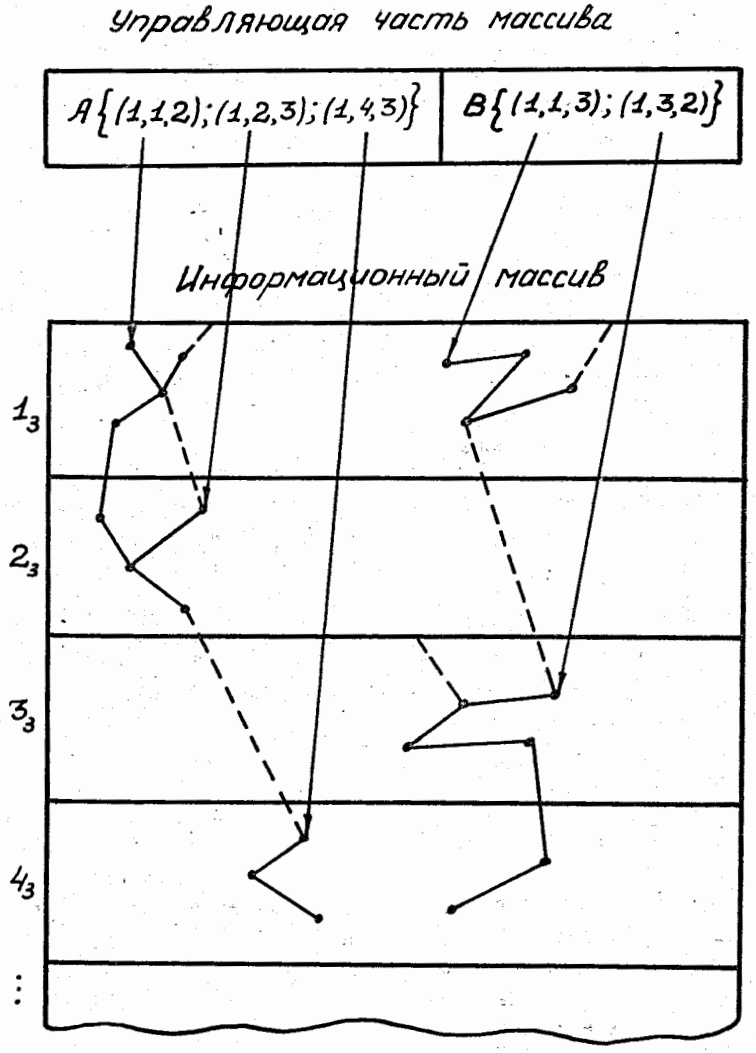


Рис. 4



ному списку дескрипторов запроса /запрос представляет конъюнкцию из 3-4 дескрипторов/. Это указывает на то, что предложенный метод дает возможность в значительной степени уменьшить время поиска в частично инвертированной ИПС. Более того, за счет изменения числа уровней иерархии и величины зоны всегда можно организовать такую частично инвертированную структуру, время ответа в которой /в случае применения метода негативных стратегий/ будет заведомо меньше, чем время ответа в любой мультисписковой структуре[4]. С другой стороны, поиск в этой структуре может быть оптимизирован таким способом, что время ответа будет меньше, чем в инверсной структуре. Этого можно достичь на базе уплотнения узловых списков, находящихся в зонах массива. Это, конечно, не приводит к тому, что система будет отвечать на запросы с определенной комбинацией дескрипторов более эффективным образом, чем на запросы с любой другой комбинацией дескрипторов, а скорее ведет к тому, что любая комбинация дескрипторов в запросе становится равновероятностной. В этом случае поиск, основанный на произвольной взаимосвязи дескрипторов в запросе, оптимизируется. Наши исследования[4,7] показывают, что для информационного массива объемом в 300000 документов достаточно иметь степень уплотнения узловых списков, равную 17, чтобы поиск в этой структуре, с применением метода негативных стратегий, совершался быстрее, чем в инверсной и мультисписковой структурах.

Литература

1. Арнаудов Д.Д. Стратегия поиска в ИПС ОИЯИ. Сообщ. ОИЯИ, Р10-8622, Дубна, 1975.

2. Арнаудов Д.Д. Некоторые вопросы организации больших информационно-поисковых систем. Препринт ОИЯИ. Р10-10368, Дубна, 1977.

3. Лефковиц Д. Структуры информационных массивов оперативных систем. М., "Энергия", 1973.

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

5. Арнаудов Д.Д. Анализ методов доступа информации к внешним ЗУ на магнитных дисках при работе информационно-поисковой системы, реализованной на ЭВМ третьего поколения . Препринт ОИЯИ, 10-7953, Дубна, 1974.

6. Китов А.И. Программирование экономических и управленческих задач. "Сов. радио", М., 1971.

7. Арнаудов Д.Д., Янев Н.И. Об одном способе применения частично-целочисленного линейного программирования для оптимизации поиска в многоуровневых ИПС. Препринт ОИЯИ, Р11-9770, Дубна, 1976.

Рукопись поступила в издательский отдел
28 января 1977 года