СООБЩЕНИЯ
ОБЪЕДИНЕННОГО
ИНСТИТУТА
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ
ДУБНА

Р11 - 10402

Д.Д.Арнаудов

СТРУКТУРНО-ФУНКЦИОНАЛЬНЫЙ ПОДХОД К ОРГАНИЗАЦИИ ИНФОРМАЦИИ В ИНФОРМАЦИОННО-ПОИСКОВОЙ СИСТЕМЕ

1977

P11 - 10402

Д.Д.Арнаудов

СТРУКТУРНО-ФУНКЦИОНАЛЬНЫЙ ПОДХОД К ОРГАНИЗАЦИИ ИНФОРМАЦИИ В ИНФОРМАЦИОННО-ПОИСКОВОЙ СИСТЕМЕ

Арнаудов Д.Д.

Р11 - 10402

Структурно-функциональный подход к организации информации в информационно-поисковой системе

Рассматриваются основные типы организации информации в поисковой системе. Особое внимание обращается на организацию информационного массива. Приведены основные функциональные характеристики информационного массива и классификационная таблица различных видов организации массива.

Работа выполнена в Лаборатории вычислительной техники и автоматизации ОИЯИ.

Сообщение Объединенного института ядерных исследований. Дубна 1977

Arnaudov D.D.

P11 - 10402

Some Aspects of the Functions and Structures in the Organization of the Information in the Information Retrieval System

Different types of the organization of the information in the IRS are described. Special attention is paid to the organization of the file. The basic characteristics and a scheme of the different types of the organization of the file are described.

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

Communication of the Joint Institute for Nuclear Research. Dubna 1977

В информационно-поисковой системе /ИПС/ имеющиеся данные являются по существу описаниями объектов, внешних по отношению к самой системе. Они могут быть описаниями людей, сигналов, идей, но они суть описания, а не сами объекты. Именно эти описания являются первичными носителями информации, обрабатываемой информационно-поисковой системой. Неотъемлемым свойством информацин о некоторой совокупности данных, которое конструктивно задается или существует в естественном представлении данных этой совокупности, является структура информации. Как отмечает Лефкович [1], проектировщик автоматизированной системы не контролирует эту первичную структуру, поскольку она является неотъемлемым и фундаментальным свойством информации как таковой. Структура информации задается извне системоаналитиком, проектирующим процедуру сбора и передачи информации. В связи с этим особое значение приобретает проблема организации информации в ИПС, которая должна удовлетворять специфическим требованиям поиска, включающим такие параметры, как время ответа, длина списка, логика запроса, тип запоминающего устройства и т.д. Проектировщик системы на базе этих требований выбирает форму записей, иногда называемую структурой данных, определяет структуру массива, взаимодействие между массивами внутри системы и т.д.

В этой работе мы рассмотрим организацию информации в ИПС на базе взаимодействия первичных структурных блоков, построение из них более сложных структур, выполняющих, соответственно, более сложные функции, и реализующих таким образом процесс хранения, обработки, и обновления больших информационных массивов.

Основные понятия структурно-функциональной организации информации

Общая блок-схема ИПС показана на *рис. 1*. Блоки 1,2,3 неразрывно связаны друг с другом и составляют

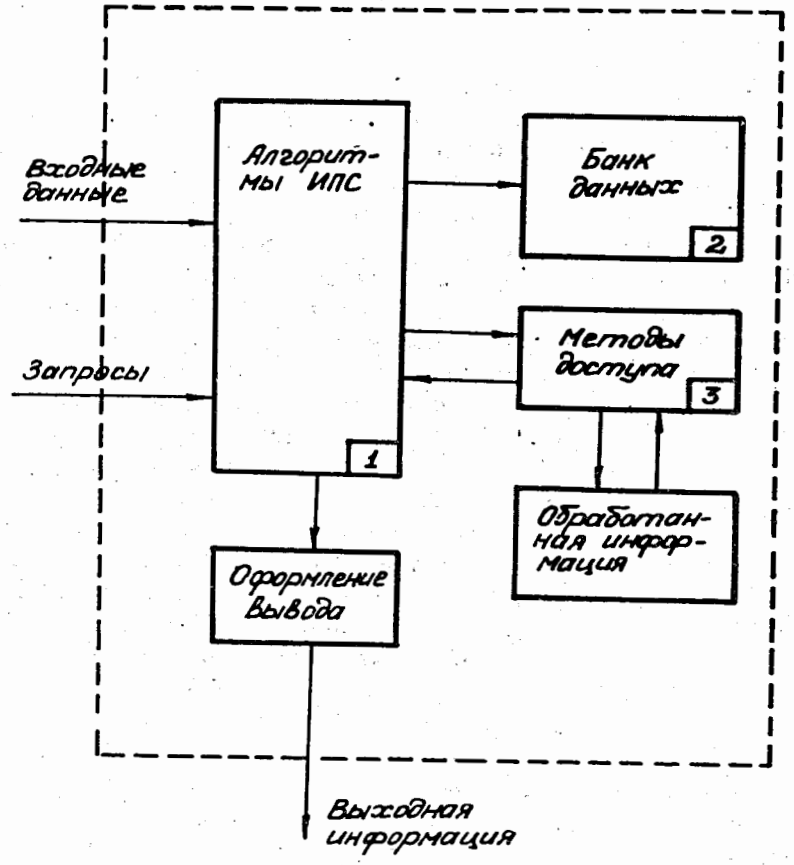

Рис. 1

сущность любой ИПС. Сформулируем теперь основные функциональные характеристики информационно-поисковой системы:

1. Осуществление быстрого доступа к хранимой информации.
2. Ввод данных различной структуры в систему "ручным" способом или автоматически.
3. Быстрая обработка и анализ хранимых данных.
4. Вывод искомой информации по заказу потребителя.
5. Редактирование и модифицирование выводимой информации.

Следовательно, каждая ИПС функционально состоит из следующих элементов:

а/ запись входных данных,
б/ вывод данных,
в/ банк данных,
г/ обработка данных,
д/ доступ к данным.

В связи с этим при разработке ИПС необходимо иметь в виду:

1. Возможности системы для построения логических структур данных и связи между ними.
2. Стратегию "отображения" логических структур на физическую структуру и вопрос управления свободной памятью /2/.
3. Стратегию доступа к логической структуре на внутренних и внешних носителях /3/.

Эти возможности системы реализуются, главным образом, на основе банка данных и методов доступа к данным, которые являются основными элементами ИПС.

Информация в системе фактически состоит из чисел, имен, кодов, символов. Банк данных может быть определен как интегральное множество доступных взаимосвязанных данных. Эти данные бывают элементарными информационными единицами, которые обычно называются атрибутами или полями, и сложными структурами /адреса связи, записи, массивы/.

Установление взаимосвязей между отдельными массивами банка данных, а также между элементами данного массива приводит к тому, что банк данных фактически несет информацию, характеризующую не только

его содержание /например, библиографические сведения о документах/, но также и, порой даже в большей степени, его структуру. Именно эта "структурная" информация является основным звеном тех средств и методов, с помощью которых организуется доступ к отдельным элементам информационной структуры. Методы доступа к информации находятся в прямой зависимости от типа этой структуры. В ИПС обычно осуществляются сложные поисковые процедуры, и доступ к отдельным элементам банка данных осуществляется с помощью массивов. Они как бы составляют коммуникационную субстанцию ИПС. В связи с этим необходимо иметь в виду, что главная задача ИПС - это обработка информации по "содержанию", что приводит к еще большему усложнению информационных структур. Обычно для достижения этого осуществляется некоторое функциональное разделение информационных элементов на множества, а сами эти множества организуются таким образом, чтобы уменьшить время поиска /4/. Все это, конечно, достигается с помощью определенной структуры массивов.

Информационный массив является групповой, составной структурой, состоящей из некоторого количества записей. Это, по существу, "форматная" организация записей, служащая для определения и распределения их на устройствах, с целью сохранения и дальнейшей обработки. Сложность такой организации заключается не в коммуникации между отдельными элементами массива, а в создании структуры массива как главной субстанции, с помощью которой необходимо осуществить информационно-поисковый процесс.

На *рис. 2* показана схема "вертикальной" организации информации. Основные понятия "вертикальной" структуры - это элементарная информационная единица, запись, массив, банк данных.

Организацию информации необходимо рассматривать, однако, не только в вертикальном, но и в горизонтальном направлении. Это требует введения таких понятий, как строка, таблица, дерево, список. Об этом подробно написано в работе /17/.

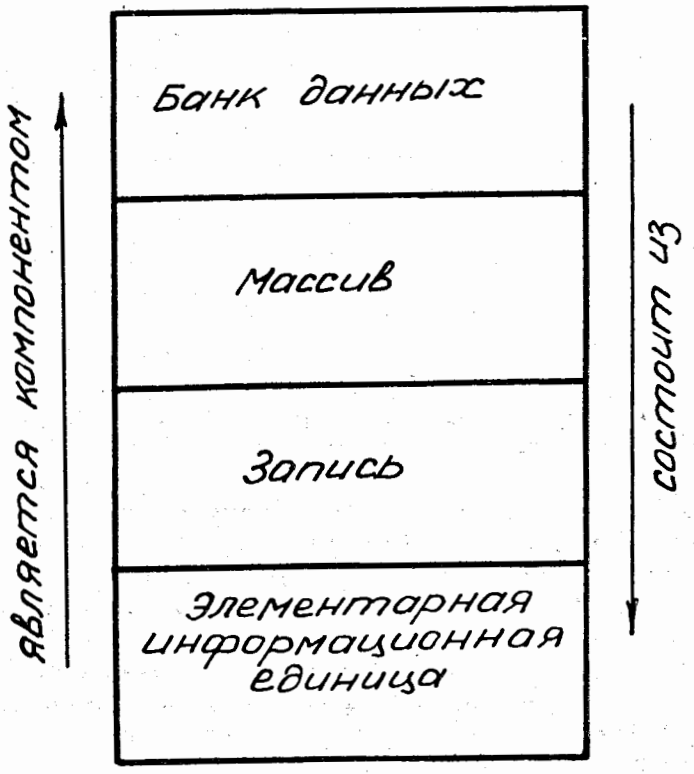

Рис. 2

Общая схема организации информационного массива

Гносеологическое развитие понятия массива описано в работе /7/. Авторы отмечают, что бросок от процессорной ориентации к ориентации банка данных начался с выделения массива как части программы. Имеется в виду, что более естественно /и не так громоздко/ связать описание записей /данных/ с конкретным массивом. Например, инструкция COPY в КОБОЛе /5/ отражает наглядно эту тенденцию. Это подчеркива-

ется и т.н. "форматными" массивами системы, где описание данных является свойством конкретного массива. Именно эта "форматная" организация записей является определяющей в структуре массива.

Разные авторы предлагают различные основы для классификации организаций массивов. Так, например, в работе /8/ предлагается классифицировать организации массивов в зависимости от их структуры и типа. В зависимости от структуры предлагают различать линейные, иерархические и степенные массивы. В зависимости от типа - последовательные, индексные и цепные массивы. Эта классификация, на наш взгляд, является в большей степени "структурной" и никак не затрагивает функциональные характеристики массива.

Хернер /9/ отмечает два основных пути при организации массива: первый путь раскрывает концепцию "единой" записи массива, а второй - концепцию "инверсного массива". Положительные и отрицательные стороны этих концепций были в дальнейшем рассмотрены Прентисом /10/, который отмечает, что с точки зрения поисковых характеристик массива эти концепции имеют свои недостатки и их нельзя принять как основу для классификации.

Другие авторы /11/ предлагают рассматривать в качестве основы организации массивов ориентированный граф. Эта структура, конечно, является более удобной, т.к. "дерево" или другую, более простую структуру можно рассматривать как специальный случай ориентированного графа. Вся проблема в данном случае состоит в представлении графа как одномерной строки в памяти ЭВМ. Описанный метод включает в себя:

A. "Единый уровень", где граф представлен как строка, состоящая из пар в виде узлов, показывающих внутренние связи.

B. "Двойной уровень", где внутренние связи показаны дважды, т.е. AB и BA.

C. "Список", где граф вначале трансформируется во множество деревьев, а сами деревья представляют собой списковую структуру.

Некоторые авторы /12/ предлагают использовать информационную структуру как основу для общей организации массива, в которой для каждой характеристики записи /для каждого дескриптора, например/ выделяется поле /одна позиция/ в некоторой таблице, где отмечается, содержится ли данная характеристика в записи или нет. Эта методика подобна позиционной строке, используемой в "Сетке - 5" /13/.

Более общий подход к использованию некоторой логической структуры как основы организации массива рассмотрен в работах /14,15/, где описывается некая иерархическая структура доступа к элементам массива. В работе /15/, например, дискутируется индексно-последовательный способ на IBM-360; рассматриваются его положительные и отрицательные стороны. В работе /16/ Оплер предлагает некую иерархическую структуру, но отмечает, что нельзя подчинять все уровни структуры главному уровню. Он предлагает главному уровню подчинить только следующий за ним уровень иерархии, то же проделать со вторым уровнем и т.д.

Мало, однако, в имеющейся литературе сказано о концепции представления организации массива одного уровня. Эта концепция предполагает, что информация о данных /записях/ расположена на одном уровне, в отдельных ячейках памяти, которые одинаковым способом доступны программе. Это фактически представляет собой некоторую разновидность манипулирования с таблицами /о работе ИПС с таблицами мы писали подробно в монографии /17/ /. Такую методику можно было бы эффективно использовать в качестве общей схемы для организации массивов, если бы ячейки памяти имели переменную длину, и если бы можно было так представить структуру информации, чтобы существовали адреса данных, которые в то же время отражали бы и взаимоотношения между отдельными данными.

Дод /18/ предлагает рассматривать все типы организации массива на базе трех известных организаций: последовательной, прямой и списковой. Лефковиц /1/ рассматривает списковую организацию как основную, из которой можно образовать всевозможные виды мас-

сивов. На основе списковой организации массивов создана и некая общая модель их организации /19/.

Мультисписковая структура как организация массивов разработана Прайсом и Греем /20/, усовершенствована дальше Прайсом /20,21/ и Китовым /22/ и известна как ассоциативно-адресный метод организации массивов.

Как уже говорилось выше, информация в массиве, с точки зрения семантики, может быть двух видов: информация, характеризующая содержание массива /"содержательная" информация/, и информация "структурная". Имеются случаи, конечно, когда в массиве содержится только "структурная" информация, т.е. для этого массива она совпадает с его "содержательной" информацией /например, массив из наборов определенных адресов связи/. В общем случае, однако, определенные функциональные требования накладываются как на "структурную", так и на "содержательную" информацию массива, что вызывает появление сложной структурно-функциональной организации информационного массива. Для пояснения этого рассмотрим связь между функциональными требованиями, с одной стороны, и "структурной" и "содержательной" информацией массива, с другой стороны, на базе основного массива поисковых образов документов ИПС ОИЯИ /23/. Первое функциональное требование /см. *табл. 1*/ узко связано со вторым, и они требуют такой мультисписковой структуры, которая позволяет вызвать и выявить соответствующие записи по набору дескрипторов, связанных булевскими операторами "и", "или", "нет". При этом минимизация времени поиска может быть достигнута только при некотором оптимальном разбиении массива на вызываемые с диска подмассивы /23/ /сегменты списков/.

Пакетный режим обработки запросов и режим реального времени требуют оптимального функционирования не только при работе со списками определенного массива, но и при совместной обработке нескольких массивов одновременно. Это ведет к созданию структурных связей между отдельными массивами.

Следующее функциональное требование непосредственно связано со сравнением дескрипторов поискового образа документа и запроса. Это ряд операций, проводимых в общем случае последовательно над отдельными элементами узла, что, со своей стороны, требует выделения узла как логической структуры /23/ при помощи логических полей.

Интерес представляет функциональное требование - поиск по характеристикам, не являющимся дескрипторами. Это подразумевает введение в структуру узла дополнительных полей, чьи признаки будут соответствовать отдельным характеристикам. Они могут подвергаться как арифметическим, так и логическим тестам, таким как: равенство, больше, меньше и т.д. Различие понятий "дескриптор" и "недескриптор" ведет к повышению эффективности работы с массивом.

Одна из главных задач при работе с массивами заключается в выборе способа обновления. При режиме обновления в реальном времени некоторые способы организации массива предпочтительнее других. В результате возникают ограничения на оптимизацию процесса поиска. В большинстве существующих поисковых систем требование реального времени относится к поиску, а не к обновлению. Поэтому комплектование блоков свободной памяти происходит в некоторых интервалах работы системы; создаются необходимые структурные связи для работы с этими блоками /например, адрес связи к первому "свободному" блоку/.

В *таблице 1* показаны используемые методы доступа к информации при определенной структуре для выполнения данного функционального требования. Эти методы доступа нельзя рассматривать отдельно от структуры и функции, потому что все это образует структурно-функциональную организацию массива. Это становится особенно ясным из того факта, что организации массива в абстрактном смысле не существует; она существует, как мы уже отметили, в связи с тем, что массив располагается на некотором устройстве памяти /оперативная память, расширенная память, память на магнитной ленте и магнитном диске/. Особенно актуальной при этом становится проблема обращения и

Таблица

Функциональное требование	Структурная информация	Содержательная информация	Метод доступа
1)Выполнение критерия смыслового соответствия на базе булевских комбинаций дескрипторов.	Мультисписковая ассоциативно-адресная структура	Набор поисковых образов документов	Произвольный, последовательный
2)Осуществление оперативного поиска (минимизация времени)	Сегментирование списков	"	Прямой, последовательный
3)Пакетный режим обработки запросов и реальное время	Осуществление структурных связей между отдельными списками и другими массивами (адреса связи)	"	Прямой
4)Непосредственное сравнение дескрипторов ПОД и ПОЗ	Структурирование в виде узлов	"	Последовательный
5)Поиск по характеристикам, не являющимся дескрипторами	Выделение отдельных логических полей узла	Признаки логических полей в составе узла	Прямой, последовательный
6) Обновление массива	Связи между структурой и свободной памятью	Набор поисковых образов документов и свободная память	Произвольный, последовательный

доступа к массиву, а также к его отдельным элементам, расположенным на различных видах памяти.

Исходя из этого, мы говорим об общей организации массива как понятии комплексном, т.е. любой массив представляет собой комбинацию из двух частей: управляющей части массива и основного массива записей. Эти две части являются определяющими для описания любой из существующих в настоящее время классификаций. И независимо от того, какие составные элементы входят в структуру /последовательные элементы, узлы, списки, индексные таблицы и т.п./ этих двух частей, организация такой структуры для выполнения определенных функциональных требований в конечном счете сводится к некоторому взаимодействию между ее элементами и элементами других массивов и структур банка данных, что реализуется либо произвольным, либо последовательным доступом. Поэтому в качестве самой общей классификации разных организаций можно рассматривать разделение массивов на организации с последовательным и произвольным доступом. Общая классификационная схема различных организаций массива показана на *рис. 3*. Необходимо отме-

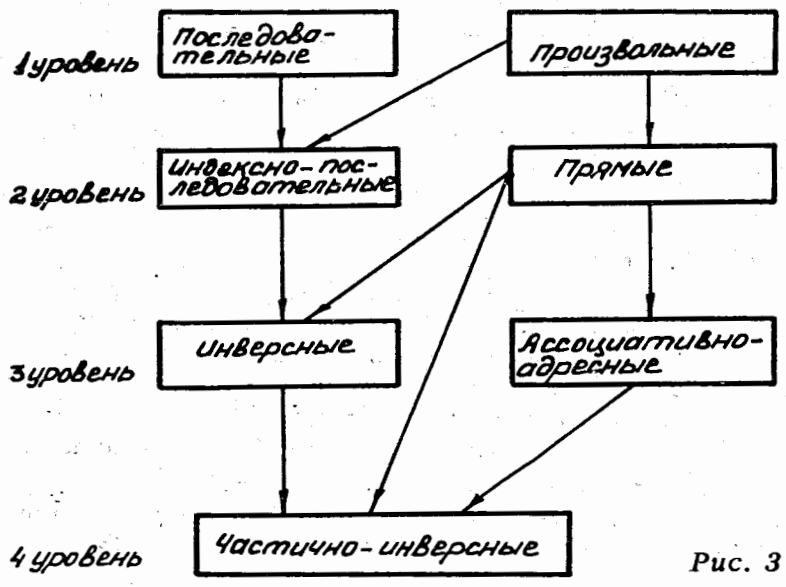

Рис. 3

тить, что отдельные элементы любого уровня с меньшим номером можно встретить в любой организации более высокого уровня /независимо от того, что это не отражено стрелками на рисунке/. Так, например, в структуре управляющей части инверсной организации элементы могут быть организованы последовательным способом /см., напр., работу [1] /.

С учетом предложенной модели последовательная организация характеризуется нулевой управляющей частью. Различные виды произвольных организаций представляют либо организацию массива с произвольным доступом, либо, в общем случае, некую комбинацию произвольного и последовательного доступов [3].

Управляющая часть массива является "средством доступа" к записям массива, находящимся в основном информационном массиве. Так, например, у последовательного массива не имеется управляющей части, и доступ к любой записи осуществляется в общем случае последовательно, т.е. не имеется предварительно никакой информации о местонахождении данной записи, если массив не упорядочен.

При массивах с произвольной организацией управляющая часть всегда несет некую информацию, связанную с размещением записей в массиве.

Сам основной информационный массив, в общем случае, представляет некоторое множество записей. Его структура может быть любой, т.е. записи могут быть расположены в последовательных ячейках памяти /упорядочены или не упорядочены по некоторому признаку/ либо нметь списковую структуру /быть связанными при помощи адресов связи/ и т.п.

Вообще говоря, организация основного информационного массива и вид организации управляющей части могут быть совершенно разными. Однако при конкретной реализации имеется некоторая взаимосвязь между этими структурами. Если, например, имеется некоторая разновидность произвольной организации массива, известная как индексно-последовательная, то управляющая часть массива представляет некоторый набор последовательных индексов, определяющих номер блока

в основном информационном массиве, а сам основной информационный массив разбит логически на блоки, в которых упорядочены соответствующие записи.

Литература

1. *Лефковиц Д. Структуры информационных массивов оперативных систем . М., "Энергия", 1973.*
2. *Арнаудов Д.Д. ОИЯИ, Р1О-10368, Дубна, 1977.*
3. *Арнаудов Д.Д. ОИЯИ, 1О-7553, Дубна, 1974.*
4. *Арнаудов Д.Д. ОИЯИ, Р1О-9178, Дубна, 1975.*
5. *Control Data 6000 Computer System, COBOL, Reference Namual, USA, 1973.*
6. *Canniny R.G. Data Management: File Organization. EDP Analyser, v. 16, No. 12, 1967, p. 14, USA.*
7. *Senko M.E. e.a. IBM Systems Journal, 1973, v. 12, No. 1, p. 30-94.*
8. *Warheit I.A. AFIPS Proc., 1963, v. 24, p.167-172.*
9. *Herner S. Methods of Organising Information for Staorage and Searching. American Doc., No. 13, Jan. 1962.*
10. *Prentice D. The Combined File Search System. IBM, San Jose, 1965.*
11. *Dzubac B. Comm. ACM, 1963, v. 8, p. 446-452.*
12. *Davis D.K. Comm. ACM, 1965, v. 14, No. 4, p. 243-246.*
13. *Китов А.И. Программирование экономических и управленческих задач . "Сов. радио", М., 1971.*
14. *Scarro G. Proc. IFIP, 1965, p. 137-141.*
15. *Poland C. Proc of IFIP, 1965, v. 1, p.249-154.*
16. *Opler A. Proc. IFIP, 1965, v. 1, p. 273-276.*
17. *Арнаудов Д.Д. Автоматизирани ИТС. Техника, София, 1975.*
18. *Dodd G. J. ACM, 1969, v. 1, No. 2.*
19. *Hsiao D. Comm. ACM, 1970, v. 13, No. 2.*
20. *Prywes N., Gray H. Proc. IFIP, Munich, 1962.*
21. *Prywes N. Proc. IFIP, 1965, v. 2.*
22. *Китов А. Программирование информационно-логических задач. "Сов. радио", М., 1967.*
23. *Арнаудов Д.Д. ОИЯИ, Р1О-8621, Дубна, 1975.*

Рукопись поступила в издательский отдел 28 января 1977 года.