ОБЪЕДИНЕННЫЙ
ИНСТИТУТ
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ

ДУБНА

E10 - 10551

D.D.Arnaudov, N.N.Govorun

SOME ASPECTS OF THE FILE ORGANIZATION AND RETRIEVAL STRATEGY IN LARGE DATA-BASES

1977

E10 - 10551

D.D.Arnaudov, N.N.Govorun

SOME ASPECTS OF THE FILE ORGANIZATION AND RETRIEVAL STRATEGY IN LARGE DATA-BASES

Submitted to the Sixth Cranfield Conference
on Mechanised Informatin Storage
and Retrieval Systems

Арнаудов Д.Д., Говорун Н.Н. E10 - 10551

Некоторые аспекты организации массивов и стратегия поиска в больших информационных системах

В работе рассматриваются методы организации больших информационных систем. Специальное внимание уделяется организации массивов. Более подробно рассматривается адаптирующаяся структура массива. Данная методика дает возможность организовать структуру массива таким образом, чтобы при увеличении объема массива минимизировать время ответа.

В связи с этим рассматривается стратегия поиска, использующая частоты дескрипторов и частоты пар дескрипторов для прогнозирования числа релевантных документов.

На базе этих методов созданы программы, которые используются в ИПС ОИЯИ.

Работа выполнена в Лаборатории вычислительной техники и автоматизации ОИЯИ.

Препринт Объединенного института ядерных исследований. Дубна 1977

Arnaudov D.D., Govorun N.N. E10 - 10551

Some Aspects of File Organization and Retrieval Strategy in Large Data-Bases

The methods of organizing a big information retrieval system are described. A special attention is paid to the file organization. An adapting file structure is described in more detail. The discussed method gives one the opportunity to organize large files in such a way that the response time of the system can be minimized, when the file is increasing.

In connection with the retrieval strategy a method is proposed, which uses the frequencies of the descriptors and the couples of the descriptors to prognose the expected number of the relevant documents.

Programmes are made, on the base of these methods, which are used in the information retrieval systems of JINR.

Preprint of the Joint Institute for Nuclear Research. Dubna 1977

The problems of the organization of Information Retrieval Systems (IRS) are complex problems,which are connected with the retrieval of various information (not only scientific one,but economical, military,medical as well). In connection with this fact one can find a great deal of questions, related to storage,retrieval and transfer of informatin.In the process of solving these problems, we must keep in mind,that they are characterized by a big volume of information,and at the same time,by frequent accesses to the external devices.In this case the response time of the system is determined mainly by the access time.

We shall discuss two main problems,which are of great importance to the realization of the IRS.

The first problem concerns the file organization,and is especially important when a file is so big that it is impossible,or at least not economical,to retrieve each element of the file for checking the needed correspondence.This is the most difficult aspect of the problem,which causes the difficulties,connected with the coordination of the memory space and the response time.It is necessary for the practical solution of this problem to organize such a structure of the file,which will satisfy the specific needs of the retrieval,which comprises such aspects as response time, query logic,type of the external memory device and so on.

In connection with this appears the second problem,which is related to an efficient retrieval strategy.Here we must note the questions of a preretrieval prognose of the number of relevant documents,methods of archive strategy, etc.

We have already noted,that the problem of a file organization comes into view, when it is difficult to coordinate the memory space on the external devices and the response time.These difficulties may be solved on the base of constructing a selfadapting structure,where the levels of a hierarchy of the file's directory and the size of the segment (zona) can be dynamically changed in the process of file increasing according to a special criteria.Then the ratio between the time ,needed for retrieval of an element of the file,and the volume of the needed memory can be regulated.In the work /2/ we have shown, that improvement of the time characteristics of IRS may be achieved on the base of combination the characteristics of the inverted and the multilist structures.In this case a partially inverted multilevel hierarchical organization can be built.

In this paper we shall discuss the main features of this method, which gives the opportunity to create an adapting structure of a file by means of adapting two parametres of the structure; the size of the segment (zona) and the number of the hierarchical levels of the file directory.

In fig.I a partially inverted structure of the file is shown.

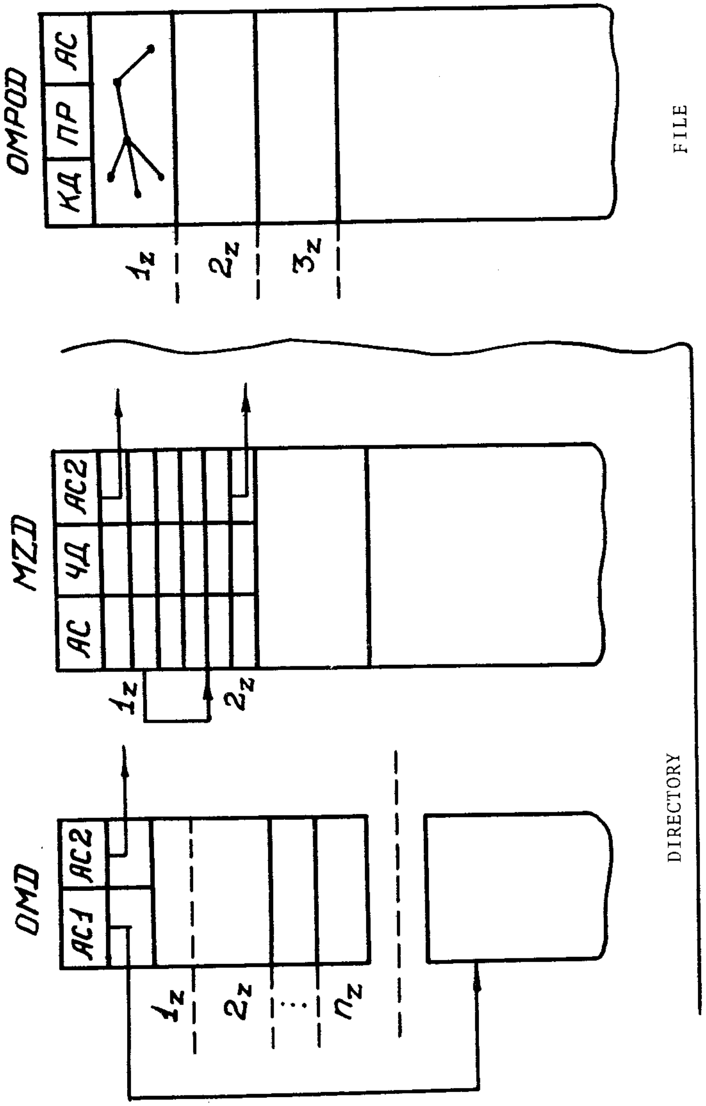

Fig. 1

The disk memory of files OMPOD,MZD,OMD is divided into logical partitions , called "zonas".In particular,the list nodes in OMPOD file, for instance,are positioned in such partitions.It means,that each list is localized in such partitions,i.e.,in zonas.The directory of the file (files OMD and MZD) has hierarchical three leveled structure. In the MZD file are gathered the list heads of the OMPOD file,and in the OMD file are collected the list heads of the MZD file.This file structure is well described in the work / 5 /,so we shall discuss the expressions,needed for a calculation of the response time,connected with the access time of the structure.For the better explanation of this fact a table is given (Table I),where the main characteristics of the file,connected with the query,computer and the file itself are represented.

The average retrieval time of a file structure with one level of the directory (a pure multilist structure),two levels, and three levels,can be calculated according to expressions (I - 3).

$$T_{od} = L\,(t_n + \tfrac{1}{2} t_o)\ ; \qquad (1)$$

$$T_{dv} = C_{K\,OMD}\, T_{z_1} + a_{OMPOD}\, C_{K\,OMPOD}\, T_{z_2}\ ; \qquad (2)$$

$$T_{tr} = C_{K\,OMD}\, T_{z_1} + a_{MZD}\, C_{K\,MZD}\, T_{z_2} + {}$$
$$+\ a_{OMPOD}\, C_{K\,OMPOD}\, T_{z_3}\ . \qquad (3)$$

Parametre T_z expresses the access time of a zona.

In fig.2 the function L = f (N_r) is shown.

In fig.3 is shown the variation of the size of the sona (Z),

Table I

N – Number of descriptors (keys) in Vocabulary
Z – Size of segment (zona) in physical records
V – Number of distinct descriptors in file
N_r – Number of Records (documents) in System
N_K – Number of Keys/Record (AV)
L – Average List Length $\left(= \frac{N_r N_K}{V}\right)$
C_K – Zona/key $\left(= \begin{matrix} C_{K\,OMD} \text{ – in OMD file} \\ C_{K\,MZD} \text{ – in MZD file} \\ C_{K\,OMPOD} \text{ – in OMPOD file} \end{matrix}\right)$
$r3_q$ – Number of Zonas per Query $\left(= \begin{matrix} r3_{q\,OMD} \\ r3_{q\,MZD} \\ r3_{q\,OMPOD} \end{matrix}\right)$
N_t – Number of Terms im a single Query (AV)
N_p – Number of Nonnegated Terms in a single Query (AV)
L_s – Shortest List Length in Query
ρ – Ratio of Query Response to L_s (AV)
p – Ratio of average number of zonas in Response per Query in OMPOD file to $C_{K\,OMD}$ $\left(= \frac{r3_{q\,OMPOD}}{C_{K\,OMD}}\right)$
$\delta = \frac{r3_{q\,MZD}}{C_{K\,OMD}}$
$a \begin{Bmatrix} a_{OMD} \\ a_{MZD} \\ a_{OMPOD} \end{Bmatrix}$ – Ratio of average Number of Query Zonas in Response to C_K $\left(a_{OMD} = \frac{r3_{q\,OMD}}{C_{K\,OMD}};\ a_{MZD} = \frac{r3_{q\,MZD}}{C_{K\,MZD}};\ a_{OMPOD} = \frac{r3_{q\,OMPOD}}{C_{K\,OMPOD}}\right)$
A – Number of File Record Addresses per DASD
$t_s = \begin{cases} t_n \text{ – Random Access time of DASD} \\ t_o \text{ – Rotation time of DASD} \\ t_{mp} \text{ – Transfer Rate of DASD} \end{cases}$

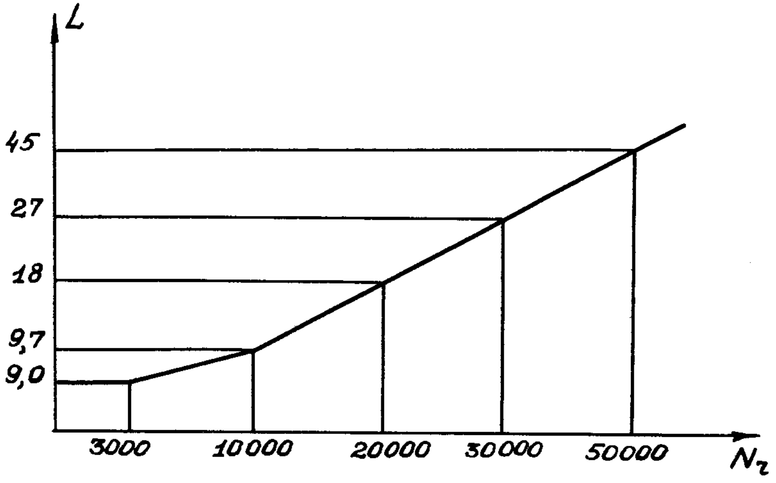

Fig.2

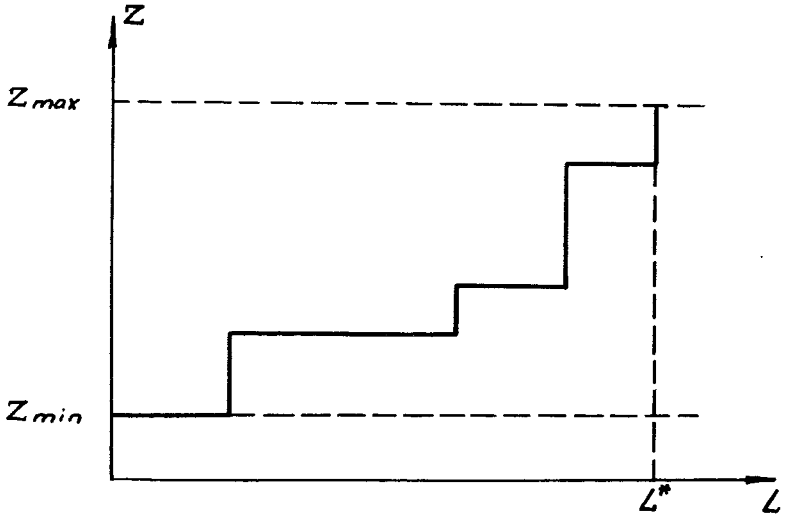

Fig.3

when the average length of the list is increasing.The point L* expresses the maximal length of the list (respectively the file),when Z_{max} is achieved.In practice,when a certain computer is concerned always there is a moment,due to the memory space,when one cannot increase Z anymore.Then the charactiristic ρ is getting to be very small (IO^{-2} - IO^{-3}).In this case a file structure with one level more must be created.

All these problems are analyzed in the work /2/.We have experimented such structures,using the INIS database on CDC 6500,and we have determined the optimal sizes of the zonas and the number of the levels of the hierarchy.The file was situated on the 84I CDC disc device. The experiments showed that a three level structure is good for files with more than 300000 documents.

Next question which we would like to discuss concerns the optimization of a retrieval strategy in large data-bases.We shall describe a method which gives the opportunity to prognose the number of the expected relevant documents in IRS,without proceeding the actual search.This enables the user to reformulate his query,if he is not satisfied with the number of the answers he is expected to receive.

Let D stands for a definite set of documents (quantity of documents - K),which are used to serve Q queries.Documents are indexed by descriptors from the vocabulary ,which size is N.The elements Q are any strings of elements from the set N,connected by conjunction, disjunction and negation.

Let R stands for a function with values in 2^D (where 2^D means a set of all subsets D). Then each query $q \in D$ is put in correspondence with an element from 2^D , which is called an answer to the query q.

Keeping in mind all that,an IRS can be formally formulated as a set of a definite set of documents D,descriptor's vocabulary N,function R,and a query language Q.An information retrieval system is partially arranged,if in the set Q a transitive and unsymetrical relation $\geqslant$ is determined. In this case,if $q \in Q$ and $S \in Q$,then $q < S$, if $q \wedge S = q$. Such relations are valid for the set $\mathcal{P}^D$ too.

We say,that an information retrieval system is including,if function R for $q < S$ gives $R(q) > R(S)$,where $(q,S) \in Q$.It means, that a characteristic "inclusion" guarantees the finding of all documents for a query S,which were found for a query q.It means,that q is equal to S but only the descriptor included in S, is excluded in q.

In this case,using a query q,more documents may be found,which were not found for S; then R(q) is bigger than R(S).

These characteristics of the system were used for the solution of our problem.We must add two more identifications: $\bar{R}(q)$ and φ , where $\bar{R}(q)$ means the expected number of relevant documents (we must remind,that R(q) is the exact number of relevant documents, found after the actual searching was done).

Identificator φ stands for a threshold number of the expected documents.It means, that if as a result of the prognose $\bar{R}(q)$ documents are expected to be found, and $\bar{R}(q) > \varphi$,then reformulation of the query is needed.

In this case the prognose problem may be defined by two unequations:

$$\bar{R}(q) \leqslant \varphi \quad ; \qquad (1')$$

$$\bar{R}(q) > \varphi \quad . \qquad (2')$$

When (I') is true,then a query is considered to be "correct",and it is retrieved by the system.But ,if (2') is true,then a query must be reformulated.

We must note,that as a first step in the solution of the problem, as an answer $\bar{R}(\alpha_i^*)$,the frequency of the descriptors in the database f(i*) is used.By the way,in a table are gathered only those f(i*), for which the uneouation (3') is true.

$$f(i^*) > \varphi \; . \qquad (3')$$

If after investigating all of the f(i*) for the descriptors in a concrete query q,is found,that among them is at least one descriptor, whose

$$f(i^*) \leqslant \varphi \quad \text{for } i^*=1..S \qquad (3'')$$

then a query is considered to be correct, and it must be retrieved by the system.In this case for f(i*) in(3") the minimal frequency of the descriptors in the query is used.But if (3") is not true, then all the possible couples of the descriptors in the query must be investigated.

Our experiments[6] show, that for queries,containing about four descriptors,it is enough to use the couples of the descriptors for the prognose.

Using this method a programme was written for the information system of JINR,which prognoses the number of relevant documents.

References

I.D.D.Arnaudov et al. JINR Dep.publ., EI-II-8553,Dubna,I975, (in Russian).

2.D.D.Arnaudov .JINR Comm.,PIO-9I78,Dubna,I975,(in Russian).

3.D.D.Arnaudov .JINR Comm.,PIO-8622,Dubna,I975,(in Russian).

4.D.D.Arnaudov .JINR Comm.,PIO-862I,Dubna,I975 (in Russian).

5.D.D.Arnaudov,N.N.Govorun.JINR Prepr.,EIO-8855,Dubna,I975,(in Russian).

6.D.D.Arnaudov et al. JINR Comm.,PIO-905I,Dubna,I975,(in Russian).

Received by Publishing Department
on April 4,I977.