Multistage Information Retrieval System Based upon Researcher Files

Arikawa, Setsuo
Research Institute of Fundamental Information Science, Kyushu University

Kitagawa, Tosio
Honorary Professor of Kyushu University, Director of International Institute for Advanced Study of Social Information Science

https://hdl.handle.net/2324/10153

出版情報：RIFIS Research Report. 51, pp.1-34, 1975-03. Research Institute of Fundamental Information Science, Kyushu University
バージョン：
権利関係：

KYUSHU UNIVERSITY

Research Institute of Fundamental Information Science

Research Report No. 51
March 1975

Multistage Information Retrieval System Based upon Researcher Files

By

Setsuo Arikawa

and

Tosio Kitagawa

Fukuoka, Japan

Multistage Information Retrieval System Based upon Researcher Files*

By

Setsuo Arikawa**

and

Tosio Kitagawa***

* The work is done as members in the research group "C-2: The structure of data base and theory of information retrieval" belonging to the Special Research Project "Advanced Information Processing of Large Scale Data over a Broad Area" supported by the Research Grand of the Ministry of Education of Japan.

** Research Institute of Fundamental Information Science, Kyushu University.

*** Honorary Professor of Kyushu University, Director of International Institute for Advanced Study of Social Information Science.

ABSTRACT

A new viewpoint on the information storage and retrieval systems mainly for the use of researchers is discussed. A mathematical formulation of the terms in the information retrieval systems is presented. Then a practical information storage and retrieval system, so called MIR-RF system, is constructed in accordance with the new viewpoint and the formulation. As raw materials for MIR-RF system, 1) the researcher files are prepared by individual researchers or groups of researchers, and 2) the lists of documents are picked up by researcher themselves from their interest. 3) From the titles of the listed documents, the key words or key phrases are extracted by a newly designed automatic indexing system REKWEST. 4) The retrievals are carried out by the multistage uses of the several kinds of associations between the two files prepared in 1) and 2).

In this system, the researcher's own knowledges, views and insights about their fields of studies can be reflected with minimal efforts. As a byproduct, some materials for constructing the general thesaurus are accumulated.

1. Introduction

In nearly all research activities, it has already become a fixed routine to publish the results of researches as printed documents and then restart the next stage of researches by using the previously published documents. The published documents are rapidly increasing in number and in vareity, and researchers' requests assume continually varying, developing and evolving aspects.

Hence an information storage and retrieval system should be designed so as to be adapted to this variety and quantity of aspects. In constructing an information retrieval system, the most essential and labouring work is, as is well-known, to make and organize data files such as document lists, key words, thesaurus and other dictionaries. A new information retrieval system should also give a solution to this problem.

In the present paper, we present a new viewpoint of information retrieval systems and their uses and report some results on constructing the information retrieval system based upon the viewpoint. We shall call the information retrieval system together with the viewpoint as MIR-RF system, which is an abbreviation of 'Multistage Information Retrieval system based upon Researcher Files'.

Our fundamental attitude and requirements for the information retrieval system mainly for use of researchers are as follows:

1). The system should and can reflect the knowledges, views and insights of the individual researchers and/or researcher groups.

2). In spite of the flexibility in 1), the system should not become a burden to the researchers, in preparing and making raw materials for the input data files.

Readers may feel something strange to our attitude and requirements. The authors, however, believe that the information searching activity of the researcher is one of the most subjective activities among all the research activities, and the success or failure of an information retrieval system depends on the capability of reflecting the subjectivity or individuality.

In implementation of our MIR-RF system, the authors restrict the field to be dealt with by the system to fundamental information science, which is the authors' study field and consists mainly of information theory in a narrower sense, mathematical neuron science, control theory including game theory, stochastic processes, pattern recognition including speech recognition and statistical multivariate analysis, and language and automaton theory.

In Section 2 we illustrate an outline of MIR-RF system. In Section 3 we introduce a new automatic indexing system REKWEST. In Section 4 we point out the necessity and sufficiency of researcher files, give a mathematical formulation of terms concerning the information retrievals and discuss the function of the multistage retrievals.

2. An outline of MIR-RF system

As stated in introduction, MIR-RF system has been designed so as to be able to reflect the researchers' knowledges, views and insights with minimal efforts. To keep the researchers effort minimal is vital for the continuous life of document retrieval system, since, for the researchers, the document searching activities are everyday jobs, but not, as a matter of fact, all of their research activities.

2.1. Researchers' jobs

In MIR-RF system, there is not much to be done on the side of the researchers. Their duties are restricted only to the following:

1). To list the documents, in which the researchers feel interest, in such forms of the references or bibliographies as in scientific papers. In substitution for these lists may be adopted some existing lists of documents such as the bibliographies at the end of books or papers surveying a field of study. Moreover the researchers can be allowed to pick up only the names of journals. All the requirements for each entity of the lists of documents are to contain a) title, b) author(s), c) name of journal, or publisher (in case of book), d) year of publication, e) volume, number, page and language used. These lists of documents constitute the bases of data files of MIR-RF system.

2). To extract key words or key phrases from (and on) the books or papers read carefully and organize them if necessary. It should be done at each time when the researchers have finished reading. The key words are organized in tree structures as shown in Fig. 2.1, although, concepts are usually comprehended as a kind of graphs or networks. The collection of these data constitutes the researcher files (RF).

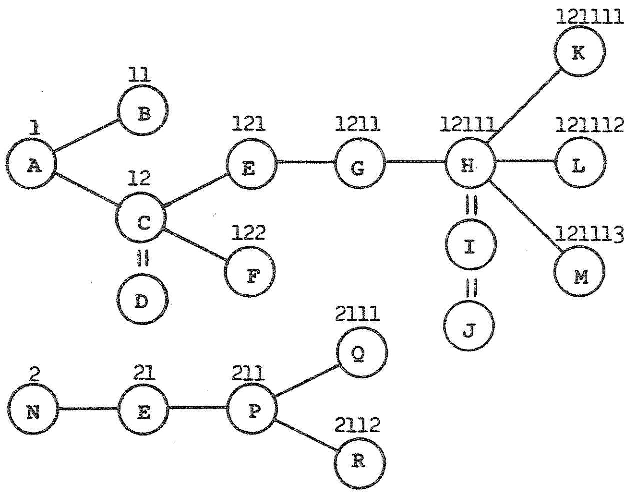
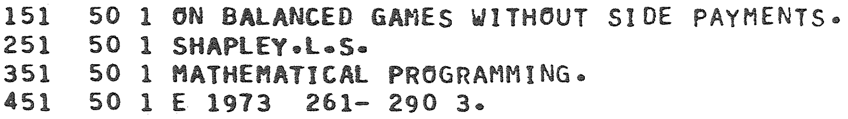


Fig. 2.1. Tree structures of an entity of RF

In Fig. 2.1, A, B,..., R in the circles denote the key words or phrases, 1, 11,..., 2112 above the circles indicate the representations of the tree structures in the file RF, the left key words or phrases are higher concepts than the right ones and the vertical equals ॥ denote the relation of equivalences of concepts.

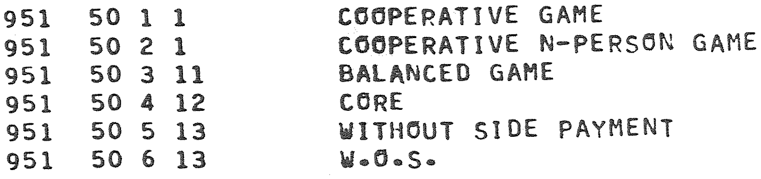
2.2. File construction

Now these raw materials prepared in 1) and 2) are punched in cards in designated formats as shown in Fig. 2.2, and Fig. 2.3.



```
151   50 1 ON BALANCED GAMES WITHOUT SIDE PAYMENTS.
251   50 1 SHAPLEY.L.S.
351   50 1 MATHEMATICAL PROGRAMMING.
451   50 1 E 1973  261- 290 3.
```

Fig. 2.2. Contents of document list



```
951   50 1 1         COOPERATIVE GAME
951   50 2 1         COOPERATIVE N-PERSON GAME
951   50 3 11        BALANCED GAME
951   50 4 12        CORE
951   50 5 13        WITHOUT SIDE PAYMENT
951   50 6 13        W.O.S.
```

Fig. 2.3. Contents of RF

These collections of punched cards are stored in a magnetic tape as a master file (MF). In this storing process key words or phrases are extracted fully automatically by an indexing system so called REKWEST, to which we refer more in detail in the next section.

From the master file (MF), four kinds of files are constructed. 1) Document file (DF) each entity of which constitutes of title, author(s), name of journal or publisher and other items on the document given in Fig. 2.2, and which is used as outputs in real retrievals. 2) Key words (phrases)-from-title file (TKF) each entity of which is an output of the REKWEST. 3) Researcher file (RF). 4) Author file (AF) which is arranged in lexicographic order. And then two kinds of inverted files $(TKF)^{-1}$ and $(RF)^{-1}$ are constructed from TKF and RF, respectively. The ordering is again lexicographic. The whole flow diagram of file construction procedure is given in Fig. 2.4.

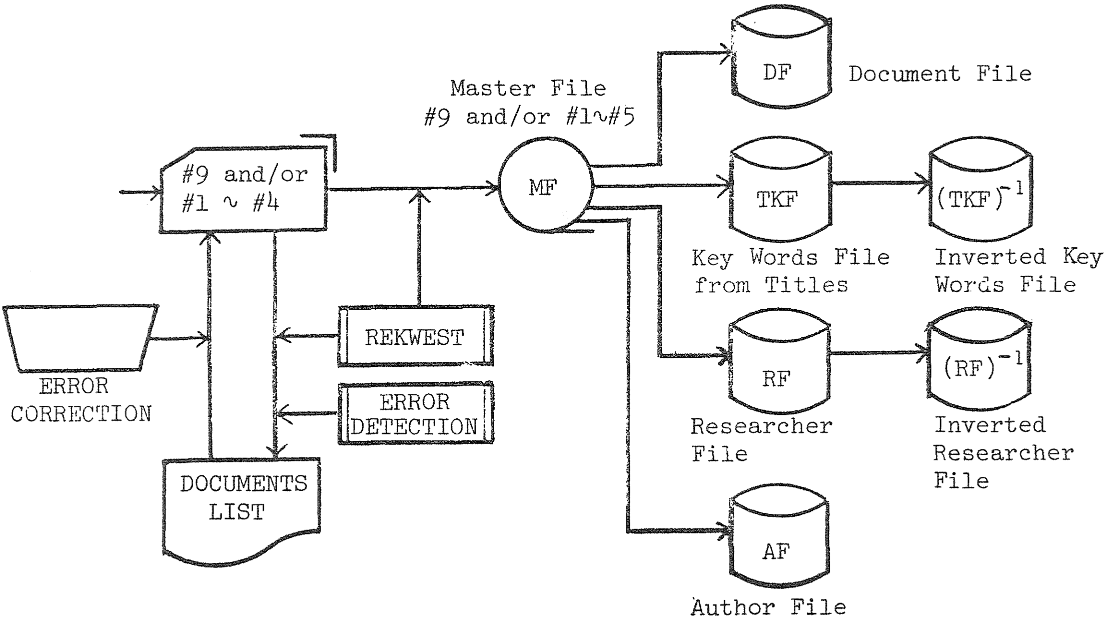


Fig. 2.4. File construction procedure

2.3. Error detection

In MIR-RF system, some kinds of errors are detected as shown in Fig. 2.4. The detectable errors are format errors and lacks of entities of document list. And at the same time, from the first file updating, MIR-RF system give a warning to errors or ambiguous letters. The ambiguities occur in processing titles of papers or book by REKWEST. The errors to which MIR-RF system gives a warning are the misspellings of words and names of author. The principle of the detection lies in the experimental fact that the same misspelling rarely occurs two or more times. This fact is widely accepted. In an experiment on about 700 documents, we could not find any error in those words that occurred two or more times. Usually this fact is applied to the input data construction, in the method that the same materials are punched by two different punchers. In MIR-RF system, however, in order to save human power, we adopt the method of using the previously obtained inverted files. If a word occurs at first, that is, if the word is not found in the inverted file, MIR-RF system gives a warning to it.

2.4. Questions for MIR-RF system

Users of MIR-RF system can question on the documents by 1) author(s), 2) name of journal, 3) logical connections of key words as in the usual retrieval systems. And also users can question by a title expected as an answer from the retrieval system. In that case the title is resolved into key words (phrases) by REKWEST just as in the indexing, and the set of such words or phrases is processed by automatically assuming a logical formula connected with logical 'and'.

3. REKWEST indexing system

In this section we propose a new automatic indexing system REKWEST. REKWEST is an abbreviation of REformed KWEST. And KWEST is an abbreviation of Key Word Extracted as a String of Terms. The underlying system KWEST of our REKWEST has been recently introduced by T. Watanabe [1], in order to decrease the redundancy of the traditionally well-used indexing systems such as KWIC and KWOC. His main idea is to extract index terms as forms of word strings by deleting appropriately designated stop-words from titles of documents. His idea of extraction of index terms as forms of word strings is accepted into our system not only because his proposition is agreeable but also because it can satisfy another point of view of ours. That is, we know and recognize by everyday experience that there is a trend that the more familiar the concepts become, the more frequently are they expressed in the forms of word strings without prepositions. And also it may be said that questions in retrievals become more similar to sentences of natural language. We believe that the total redundancy in indexing and retrieval should considerably decrease.

The language to be dealt with by the present singularizer is restricted to English. For our purpose, at the starting point, it is sufficient to deal only with English, since we can say that about 95 or more percent of documents in our field are written, or at least have titles, in English.

Now let us explain the outline of our REKWEST system. The flow of jobs in REKWEST is given in Fig. 3.1.

From now on we use a term 'key string' for 'key word', 'string of key words' or 'key phrase'.

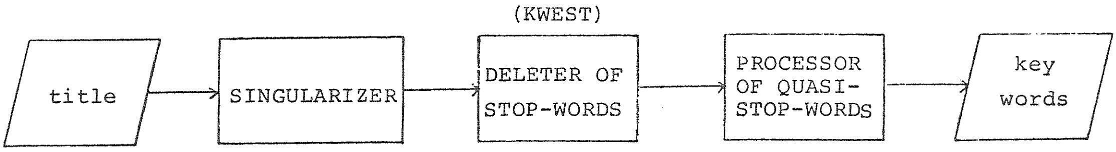


Fig. 3.1. REKWEST indexing system

As in the above figure, REKWEST system has two more stages of processing in addition to the original KWEST system. One is a singularizer of all noun form words which constitute a title. And the other is a processor of quasi-stop words, which is a notion newly introduced in MIR-RF system.

3.1. Singularizer

We adopt a singularizer to add to our indexing system from two purposes: The first is to decrease the number of index terms and the second is to compress the size of the dictionary, i.e., the list of stop-words and quasi-stop-words. In fact, by the adoption of a singularizer, we have seen about 32% decrease of index terms in the extracted samples which consists of 700 documents collected by Wood [2]. The decreasing rate of size of the dictionary is not worth mentioning, since the total size is kept within a very small size. At any rate, by this simple singularizer we have achieved considerable decrease of redundancy in document retrieval systems, at the stage of preprocessing.

English nouns are divided into those with regular plurals and those with irregular plurals. Hence a kind of a dictionary is necessary to process the latter. We have selected and summarized 30 suffixes of nouns by using the Inverted English Dictionary [4]. These 30 suffixes play a role of a dictionary in our singularizer.

This singularizer with such a small dictionary can be said to be sufficient for our present purpose, i.e., for processing of the field of fundamental information science. In fact, we could not find any essential example of misprocessing or underprocessing in a practical experiment of the singularizer with about 2,000 documents. Detected misprocessings were on words other than common nouns such as 'towards', 'does', Gries (a name of author) etc.

The practical application of the singularizer is restricted to words with 4 or more length, in order to prevent strings of abbreviated notation from being overprocessed.

3.2. Stop-words

The stop-words in the usual automatic indexing systems such as KWIC are selected by taking into account of the frequency of occurrences of words. Roughly speaking, words with high frequency or low frequency are picked up as stop-words. In our system, however, we do not adopt such a statistical criterion. The reason is that we have not yet sufficient data to adopt such a criterion, and we are afraid to lose meaningful words. We should, in our retrieval system, take care not to lose informations in order to answer varieties of researchers' requests.

In REKWEST system, from these viewpoints, we have picked up about 80 words below:

1) prepositions except 'up', 'down', 'to', and, 'without',
2) articles,
3) co-ordinate conjunctions,
4) auxiliary verbs,
5) numerals from one to ten,
6) personal pronouns,
7) interrogative except 'how', 'what', 'which', and 'when',
8) pronominal adjectives,
9) some other adverbs and adjectives concerning degrees,

and some ing-forms working participially and prepositionally.

Note that some prepositions such as 'to' and 'up' and interrogatives are excluded from the list. These excluded words are to be processed at the next stage as the quasi-stop-words.

3.3. Quasi-stop-words

The essence and difficulty of indexing systems such as KWIC, KWOC and KWEST lie, in fact, in setting up a criterion for the stop-words. Under a loose criterion a trivial redundancy occurs and under a rigid criterion a vital information loss occurs.

In order to improve these systems without appealing statistical methods, it is necessary to develop at least syntactical considerations on titles. In general it is necessary for the more syntactical consideration to construct some large and complicated dictionary. In our system, however, it has been attempted to use the notion of quasi-stop-words to make a step toward a syntactical consideration by keeping the size of a dictionary as small and simple as possible.

A word is called a quasi-stop-word if it is deleted under a contextual condition. The following three kinds of words are selected as quasi-stop-words.

1) 'to', 'when', 'how', 'what' and 'which'.

These words are not deleted if they occur in forms 'when to', 'how to', etc, otherwise deleted.

2) words ending with 'ed' or 'ble'.

These words are deleted if they occur in the right end of key strings, otherwise not deleted.

3) 'kind', 'note', 'study', 'remark', 'result', 'case',etc.

These words are deleted if each of them occurs alone in the outputs of the deleter.

By means of this simple processor of the quasi-stop-words, the redundancy of index terms decreases to a certain degree, without losing meaningful words in some other contexts.

Remark 3.1 A main difference of titles from the usual sentences lies in the fact that in almost all titles verbs are transformed into nouns or neglected, and subjects are also neglected. In return, some prepositions are supplied. Consequently all the prepositions in titles play key roles to decide the syntax of titles. Hence we can say that some simple syntax of titles has been taken into consideration when deleting stop—words, since the stop-words mainly consist of prepositions (See also [5], [7]).

3.4. Examples and remarks

Finally we present some examples of outputs by REKWEST. An original title (A) is decomposed into (D) by REKWEST:

(A) ON THE LANGUAGES DEFINED BY SENTENTIAL FORMS OF
CONTEXT-FREE GRAMMARS

$\Downarrow$ (by the singularizer)

(B) ON THE LANGUAGE DEFINED BY SENTENTIAL FORM OF
CONTEXT-FREE GRAMMAR

$\Downarrow$ (by the deleter of stop-words)

(C) LANGUAGE DEFINED
SENTENTIAL FORM
CONTEXT-FREE GRAMMAR

$\Downarrow$ (by the processor of quasi-stop-words)

(D) LANGUAGE
SENTENTIAL FORM
CONTEXT-FREE GRAMMAR.

In the above example, the underlined words are processed at the next stages. Some other examples are given in Fig. 3.2.

```
1  A STUDY ØN SPEECH SYNTHESIS SYSTEM USING ZERØ-CRØSSING WAVE.
11 SPEECH SYNTHESIS SYSTEM
12 ZERØ-CRØSSING WAVE

2  AN EXPERIMENTAL STUDY ØN SPEECH RECØGNITIØN AND LINGUISTIC INFØRMATIØNS.
21 EXPERIMENTAL STUDY
22 SPEECH RECØGNITIØN
23 LINGUISTIC INFØRMATIØN

3  ØN LENGTH FUNCTIØNS ØF LANGUAGES RECØGNIZABLE BY LINEAR BØUNDED AUTØMATA.
31 LENGTH FUNCTIØN
32 LANGUAGE
33 LINEAR BØUNDED AUTØMATØN

4  WHEN TØ STØP..A ZERØ-SUM GAME MØDEL.
41 WHEN TØ STØP
42 ZERØ-SUM GAME MØDEL
```

Fig. 3.2. Input-output examples of REKWEST

In Fig. 3.2, the n denotes an input and the nm denotes an output for the input n.

Remark 3.2. In punching cards, some manual processings are added by secretaries (punchers).

1) Two kinds of hyphens, the one to compound words and the other to indicate subtitles, are distinguished. The latter is replaced by a hyphen with two spaces before and behind it.

2) Numerals to denote series or part numbers are punched in Arabic numerals and deleted in REKWEST. Numerals not to be deleted in REKWEST are punched in Roman numerals.

3) Parentheses meaningful are punched without spaces, and ones meaningless are punched with one or more spaces.

4) Non-English letters are punched according as their English pronounciations.

Remark 3.3. From the first updating, key strings with too many characters or with too many words are decomposed into shorter strings by using the inverted file $(TKF)^{-1}$. For example, a 'key string' with 5 words

AVERAGE REWARD MARKOVIAN DECISION PROCESS

is decomposed into two key strings

AVERAGE REWARD,

MARKOVIAN DECISION PROCESS,

if these two key strings are in the file $(TKF)^{-1}$. Otherwise the system gives a warning to it.

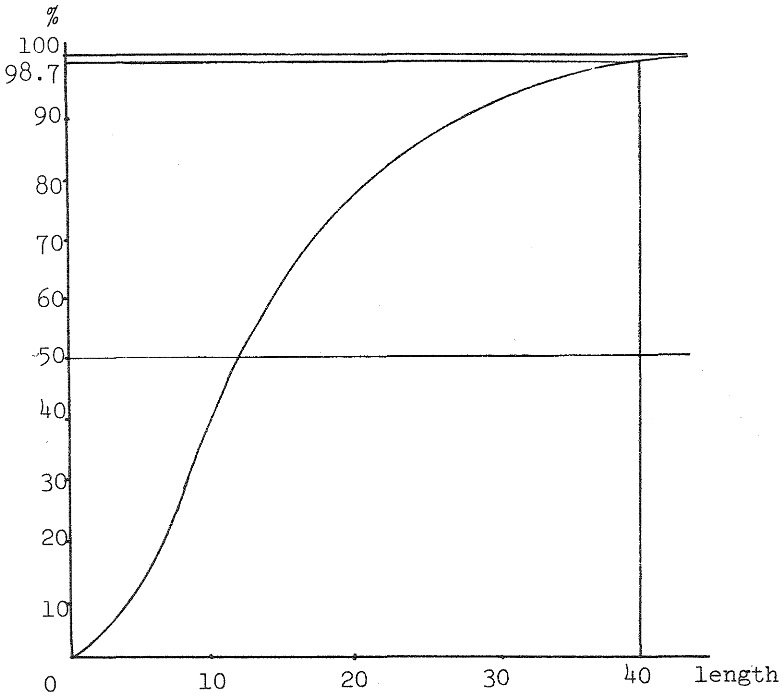


Fig. 3.4. Distribution of lengths of key strings

Number of words	1	2	3	4	5∿9
%	38.2	41.7	15.5	3.5	1.1

Table 3.1. Frequency of numbers of words in key strings

According to the experiments on about 2,000 documents summarized in the above Table 3.1 and Fig. 3.4, the critical values to apply these additional processings are set at 40 and 5 in length and in number of words, respectively.

4. Multistage retrievals using researcher files

In our MIR-RF system, the researcher files are used for two purposes. The one is to supplement key strings extracted from titles by REKWEST. And the other is to reflect researchers' knowledges, views and insights as stated in Introduction.

The basic retrieval procedures by MIR-RF is illustrated in Fig. 4.1. Each part of the procedures is explained in detail in the later sections.

Before that we want to discuss the necessity of RF.

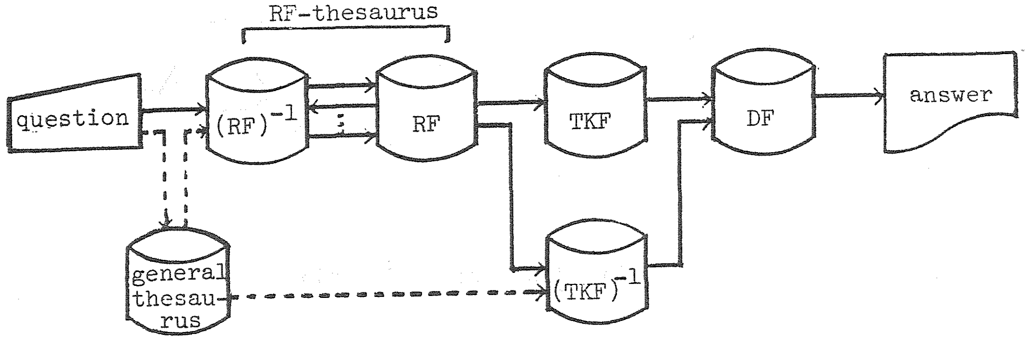


Fig. 4.1. Basic flow of MIR-RF retrievals

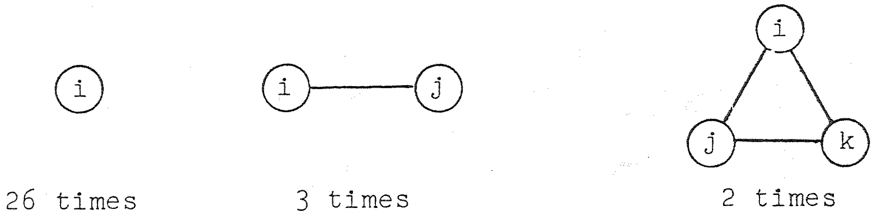
4.1. Necessity of researcher files

In principle it is possible to retrieve only by using the files TKF and/or $(TKF)^{-1}$. In such retrievals it has been observed that the precision rate becomes very high but the recall rate becomes extremely low by our experiments below.

We have selected all the papers on the automaton and language theory from the Journal of ACM from 1968 to 1970. There are 31 documents. These papers have key strings selected by their authors. Let RK(i) be a set of key strings extracted from the title of the i-th document by REKWEST. Let $R=(r_{ij})$ be an incidence matrix defined by

$$r_{ij} = \begin{cases} 1 & \text{if } RK(i) \cap RK(j) \neq \phi, \\ 0 & \text{otherwise.} \end{cases}$$

Then we have Fig. 4.2 as its graphic representation.



26 times 3 times 2 times

Fig. 4.2. The incidence matrix R

The Fig. 4.2 shows that at most three documents are able to be retrieved by a question which consists of a key string extracted by REKWEST. This fact is not due to the independence of these documents. In fact, we have observed the following. Let AK(i) be a set of key strings of the i-th document selected by its author(s), and $A=(a_{ij})$ be an incidence matrix defined by

$$a_{ij} = \begin{cases} 1 & \text{if } AK(i) \cap AK(j) \neq \phi \\ 0 & \text{otherwise.} \end{cases}$$

Raising A third power, we have Fig. 4.3, where $A^3=(a_{ij})$ is normalized by if $a_{ij} > 0$ then $a_{ij}:=1$ else a_{ij}:=blank. This asserts that the 31 documents except the two may be assumed identical.

As a cause making the recall rate extremely low, it may be taken that the key strings are too long or the number of words constituting the strings is too large. But this is not necessarily the case. In fact, we have observed the following fact, by our larger scale experiment on about 700 documents in [2]. The list of documents are carefully selected by D. Wood from the field of automaton and language theory. Let w be

```
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
                                    1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
                                            1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   1 1 1   1 1 1 1 1 1 1 1
```

Fig. 4.3. The matrix A^3

a singular noun and D(w) be a set of documents in the above list each of which contains a noun w or its plural form. Then we have observed that

$$|D(\text{automaton})| = 144,$$
$$|D(\text{language})| = 209,$$
$$|D(\text{automaton}) \cup D(\text{language})| = 341,$$

where $|D|$ denotes the number of elements in D. It seems that the values are large enough. However the ideal value of the last expression should be near to 700, i.e., the total number of documents (See also Section 4.3.3.).

From the above observations we conclude that it is realy necessary to support REKWEST system. In general, such an indexing system is supported by a large scale thesaurus. In our MIR-RF system, however, the thesaurus is replaced by the researcher files. By this reason we call the process of going back and forth between $(RF)^{-1}$ and RF an RF-thesaurus, as shown in Fig. 4.1. This is a necessary use of RF for our retrieval system.

4.2. Simulation of researchers' document searching activities

Another use of RF is to simulate document searching activities of researchers. Every researcher has a look at the contents of journals and checks the documents in which he finds an interest. The decision of the choice is done from his own motive using his own brain, i.e., his accumulated knowledges, views and insights, and associations. This sort of searching is the simplest and the most fundamental one in all searching activities. This is the very aspect that is to be simulated by MIR-RF system. In our MIR-RF system, the brain is taken as $(RF)^{-1}$ and RF with back and forth motions between them. The motive is taken as a question to MIR-RF system. By using his brain $(RF)^{-1}$ and RF with some associations. MIR-RF system searches documents from the document file DF according to the question.

RF is continuously updated just as its owner's brain. RF is growing with its owner. In this sense we can say that in MIR-RF system every researcher or every user can have a newest copy of his brain with a minimal effort.

4.3. Basic algorithm of multistage retrievals

Our system is of multistage retrievals from the two aspects.

One is seen in RF-thesaurus in Fig. 4.1. That is, in retrieval, the original question is appropriately augmented through multistage uses of $(RF)^{-1}$ and RF. And the other is seen, as in Fig. 4.4 below, in the multistage uses of two or more pairs of different $(RF)^{-1}$ and RF.

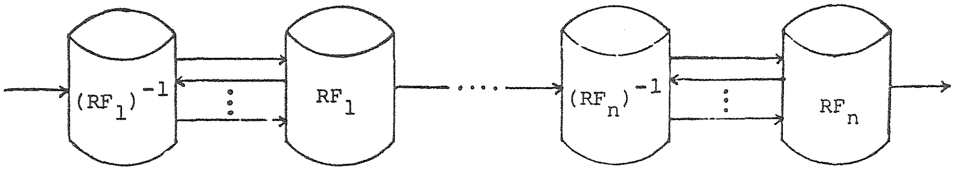


Fig. 4.4. A part of multistage retrievals using different RF's

4.3.1. A formalism on files and retrievals

In this Section we attempt to formulate the basic multistage retrievals including these two modes.

Definition 4.1. Let X, Y and Z be sets. For two relations $R_1 \subset X \times Y$ and $R_2 \subset W \times Z$, an operation τ is called a composition of R_1 and R_2 if $R_1 \tau R_2 \subset X \times Z$.

Notation 4.1. Let $\circ$ denote a composition of $R_1 \subset X \times Y$ and $R_2 \subset Y \times Z$ such as

$$R_1 \circ R_2 = \{(x,z) \ ; \ (\exists y) \ [(x,y) \in R_1 \ \& \ (y,z) \in R_2]\}.$$

For $R \subset X \times Y$, let $R(x) = \{y \ ; \ (x,y) \in R\}$. Let N be the set of all positive integers, V be a finite vocabulary and V* be the set of all finite strings over V. It is assumed that each positive integer is identified with a document.

Definition 4.2. A mapping $K : N \longrightarrow 2^{V^*}$ with $|K(n)| < \infty$ for any n in N is called an indexing.

Definition 4.3. Let $D \subset N$ be a finite set of documents. A finite relation $F = \{(n,w) ; n \in D \ \& \ w \in K(n)\}$ is called a key strings file (on D by K). And an inverse relation $F^{-1} = \{(w,n) ; (n,w) \in F\}$ is called an inverted file (on D by K).

Definition 4.4. Let Q be a set defined as follows:

(i) if $q \in V^*$, then $q \in Q$.

(ii) if $q_1, q_2 \in Q$, then $(q_1 \ \& \ q_2), (q_1 \vee q_2), (\sim q_1) \in Q$,

(iii) only the elements follow from (i) and (ii) are elements of Q. Each element of Q is called a question.

Definition 4.5. A logical retrieval is a function $L:Q \longrightarrow 2^D$ such that

(i) $L(w) = F^{-1}(w)$ for $w \in V^*$,

(ii) $L((q_1 \ \& \ q_2)) = L(q_1) \cap L(q_2)$,

(iii) $L((q_1 \vee q_2)) = L(q_1) \cup L(q_2)$,

(iv) $L((\sim q)) = D - L(q)$

for q, q_1, q_2 in Q.

Now let us define the multistage retrievals as extensions of the usual logical retrievals in Definition 4.5.

Definition 4.6. Let D_1 and D_2 be finite subsets of N, K_1 and K_2 be indexings, and $F_1 \subset D_1 \times K_1(D_1)$ and $F_2 \subset D_2 \times K_2(D_2)$.

(1) A composition π of F_1 and F_2^{-1} is called a document-association on D_1 and D_2,

(2) A composition τ of F_1^{-1} and F_1 is called a word-association on $K(D_1)$.

That is, the document-association is an operation to construct an intermediate file $G \subset D_1 \times D_2$, and the word-association is, in turn, an operation to construct an intermediate file $H \subset K_1(D_1) \times K_1(D_1)$.

Definition 4.7. Let $D_1, \ldots, D_n$ be finite subsets of N, $F_1, \ldots, F_n$ be key string files on $D_1, \ldots, D_n$, respectively, and $K_1, \ldots, K_n$ be indexings. For document-association $\pi_1, \ldots, \pi_{n-1}$ on $D_1, \ldots, D_n$, respectively, and $w \in V^*$, let

$$(4.1) \quad L^{(n)}(w) = F_1^{-1} \circ (F_1 \pi_1 F_2^{-1}) \circ (F_2 \pi_2 F_2^{-1}) \circ \cdots \circ (F_{n-1} \pi_{n-1} F_n^{-1})(w),$$

and for word-association $\tau_1, \ldots, \tau_{n-1}$ on $K_1(D_1), \ldots, K_{n-1}(D_{n-1})$, resepectively, let

$$(4.2) \quad L^{(n)}(w) = (F_1^{-1} \tau_1 F_1) \circ (F_2^{-1} \tau_2 F_2) \circ \cdots \circ (F_{n-1}^{-1} \tau_{n-1} F_{n-1}) \circ F_n^{-1}(w).$$

Then an *n-stage retrieval using document-associations* $\pi_1, \ldots, \pi_{n-1}$ is a function $L^{(n)}: Q \longrightarrow 2^{D_n}$ which satisfies (4.1) and (ii), (iii) and (iv) in Definition 4.5. And an *n-stage retrieval using word-associations* $\tau_1, \ldots, \tau_{n-1}$ is a function $L^{(n)}: Q \longrightarrow 2^{D_n}$ which satisfies (4.2) and (ii), (iii) and (iv) in Definition 4.5.

4.3.2. Multistage retrievals using document-associations

In this Section we illustrate the retrieval modes which use a file TKF in Fig. 4.1. For the sake of simplicity we deal only with the case of multistage retievals on a single pair of $(RF)^{-1}$ and RF.

Definition 4.8. An *n-stage retrieval using RF and TKF* is a function $L_d^{(n)}: Q \longrightarrow 2^{D_T}$ such that

(4.3) $L_d^{(n)} = (RF)^{-1} (\circ\ RF\ \pi_r\ (RF)^{-1})^n \circ (TKF\ \pi_t\ (TKF)^{-1})$,

where $D_T = \{x\ ;\ (x,y) \in TKF\}$, $RF \subset TKF$ and π_r, π_t are document-associations. L_d is also written as

(4.4) $L_d^{(n)} = L_d^{(n)}[\pi_r/RF,\ \pi_t/TKF]$

$= L_d^{(n)}[\pi_r,\ \pi_t]$.

In our MIR-RF system, users can select or define any one of the document-associations. Present MIR-RF system is provided with the following document-associations $\pi_1(\varepsilon)$, $\pi_2(\varepsilon)$ and $\pi_3(\varepsilon)$. Before defining these associations, we introduce three auxiliary coefficients.

(4.5) $r_1(i,j) = \max\{1/|K(i)|,\ 1/|K(j)|\} \cdot |K(i) \cap K(j)|$,

(4.6) $r_2(i,j) = |K(i) \cap K(j)| / (|K(i)| + |K(j)| - |K(i) \cap K(j)|)$

(4.7) $r_3(i,j)$ is a coefficient defined with a distance between two key strings taken into account (The details are omitted here, see [5]).

Now by using these coefficients, $\pi_k (k=1,2,3)$ are defined as follows:

(4.8) $F\pi_k(\varepsilon)F^{-1} \equiv \{(i,j)\ ;\ (\exists x)(\exists y)[(i,x) \in F\ \&\ (y,j) \in F^{-1}]$

$\&\ r_k(i,j) \geq \varepsilon\}$.

Remark 4.1. There exist other document-associations than the above. One of the most familiar is one depending on

Pearson's correlation coefficient which is more complex than ours. Since in our MIR-RF system the intermediate files are flexible and constructed in each retrieval, it seems to be better to use the simpler ones.

By our small scale experiments it has been observed that, even when $n=0$, the increases of the recalls are from 1.5 to 3 times as large as those in the logical retrievals with no use of thesaurus. In the experiment, D_T is selected at random from the field of the automaton and language theory and other branches of fundamental information science in ratio about 1 : 1. D_R consists of documents used in Section 4.1 and RF is one selected by the authors of the documents. And $|D_T|=203$, $|D_R|=31$ and $\varepsilon=1/2$ or $1/3$. On the precisions, however, it has been observed that the rates do not much decrease in retrievals by higher level concepts, but the rates heavily decrease in retrievals by lower level ones. (See also [5]). This is in the nature of the case. If a lower level concept is extracted as a key string together with a higher level concept, then it may happen that the higher level concept brings many other unexpected key strings. This possibility is due to the sensitivity of our document-associations, more exactly, due to the extreme difference of sizes of indexings $RF(D_R)$ and $TKF(D_T)$. In fact, for almost all n in D_T and m in D_R, we have seen that $|TKF(n)| \leq 4$ and $|RF(m)| \geq 7$ (See also Table 3.1.).

From these expriments, it can be concluded that the multistage retrievals using document-associations should be used mainly in case of retrieval with questions of highter level concepts.

4.3.3. Multistage retrievals using word-associations

The retrieval modes used in the last Section depend on a numerical nearness of documents. In the present case, though we can also use such a numerical nearness, we use here only retrieval modes depending on a qualitative nearness of words. Our present aim is mainly to illustrate the retrieval modes which use a file $(\mathrm{TKF})^{-1}$ in Fig. 4.1. For the sake of simplicity we deal again only with the case of multistage retrievals using a single pair of $(\mathrm{RF})^{-1}$ and RF.

Definition 4.9. An n-stage retrieval using RF and $(\mathrm{TKF})^{-1}$ is a function $L_W^{(n)}: Q \longrightarrow 2^{D_T}$ such that

$$(4.9) \quad L_W^{(n)} = ((\mathrm{RF})^{-1} \, \tau \, \mathrm{RF} \circ)^n \, (\mathrm{TKF})^{-1},$$

where τ is a word-association, and $L_W^{(n)}$ is also written as

$$(4.10) \quad L_W^{(n)} = L_W^{(n)}[\tau/\mathrm{RF}, \mathrm{TKF}]$$
$$= L_W^{(n)}[\tau].$$

Before giving specifications of τ, we discuss briefly thesauruses.

Definition 4.10. A thesaurus S is a finite collection of finite relations $P \subset V^* \times V^*$.

In a real retrieval, a subthesaurus $P \in S$ is selected and used.

Definition 4.11. Let $\tau_1, \ldots, \tau_k$ be word-associations provided in MIR-RF system. Then

$$(4.11) \quad P_i^{(n)} = ((RF)^{-1} \tau_i \; RF \; \circ)^n$$

is called an n-stage subthesaurus by τ_i. If the series $\{P_i^{(n)}\}$ converges, then

$$(4.12) \quad P_i = \lim_{n \to \infty} P_i^{(n)}$$

is called a subthesaurus by τ_i. If every $\{P_i^{(n)}\}$ converges to P_i, $(i=1,\ldots,k)$, then the collection

$$(4.13) \quad S = \{P_1, P_2, \ldots, P_k\}$$

is called an RF-thesaurus of MIR-RF system.

Now we have precise definitions of thesaurus and RF-thesaurus, and have a reason why we have called so in Fig. 4.1. Let us return to $L_w^{(n)}$.

In order to specify τ in Definition 4.9, we prepare some definitions and notations. Let U be a finite set of symbols, and let $T : N \times V^* \longrightarrow 2^{U^*}$ such that $T(i,x)$ is the set of tree structures written in RF. For example, in Fig. 2.3, $T(50,\text{CORE}) = \{12\}$. (Note that, strictly speaking, the file RF should be taken as a finite relation $RF \subset D_R \times U^* \times V^*$. Until now we have neglected the middle terms, since no necessity has occurred.) We write $T(i,x) < T(j,y)$ if there exist $u \in T(i,x)$, $v \in T(j,y)$ and $w \in U^*$ such that $uw=v$.

In our MIR-RF system, the following four word-associations

τ_0, τ_1, τ_2, τ_3 are provided:

(4.14) $\tau_0 = \circ$,

(4.15) $(RF)^{-1}\ \tau_1\ RF = \{(x,y);(\exists i)[(x,i) \in (RF)^{-1}\ \&\ (i,y) \in RF\ \&\ T(i,x) \cap T(i,y) \neq \phi]\}$,

(4.16) $(RF)^{-1}\ \tau_2\ RF = \{(x,y);(\exists i)[(x,i) \in (RF)^{-1}\ \&\ (i,y) \in RF\ \&\ T(i,y) < T(i,x)]\}$,

(4.17) $(RF)^{-1}\ \tau_3\ RF = \{(x,y);(\exists i)[(x,i) \in (RF)^{-1}\ \&\ (i,y) \in RF\ \&\ T(i,x) < T(i,y)]\}$.

From the definitions, it is obvious that

(4.18) $P_1^{(n)} \subseteq P_k^{(n)} \subseteq P_0^{(n)}$ $(n \geq 0,\ k=2,3)$,

(4.19) $P_i^{(n)} \subseteq P_i^{(n+1)}$ $(n \geq 0,\quad 0 \leq i \leq 3)$,

and for each i, the series $\{P_i^{(n)}\}$ converges. Hence the collection of subthesaurus P_i $(0 \leq i \leq 3)$ constitutes the RF-thesaurus of MIR-RF system.

Theoretically the ordering of recall rates corresponding to the uses of τ_0, τ_1, τ_2, and τ_3 coincides with the ordering in (4.18), and the ordering of such precision rates is a reverse ordering of (4.18). And in the multistage retrieval using the word-association τ_1, we can increase the precision rates with keeping recall rates constant. This property has been observed in a real computer experiment. For example, in the larger scale experiment in Section 4.1, even by virtue of a simple RF shown in Fig. 4.5, we have

$$|L_W^{(1)}[\tau_1](\text{automaton})| = 260$$

$$|L_W^{(1)}[\tau_1](\text{language})| = 304$$

$$|L_W^{(1)}[\tau_1](\text{automaton} \vee \text{language})| = 551,$$

which are comparatively larger than ones in Section 4.1.

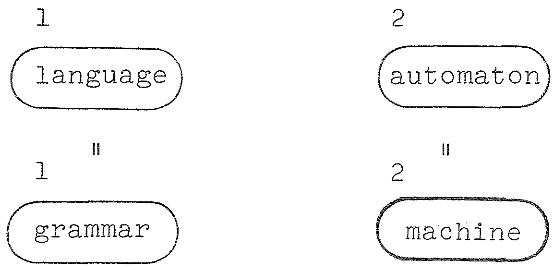


Fig. 4.5. The simple RF

Finally we stress again the fact that the most typical difference of our multistage retrievals using word-associations, namely n-stage subthesaurus, from the usual retrievals lies, in the point that, in the former, the thesaurus is taken as a flexible one and is being constructed in the course of the retrievals, while in the latter it is taken as a fixed one and is already constructed before the retrievals. Although because of the flexibility it is obliged to spend a longer time and a wider space of working memory in every retrieval, we can have, in return, a lot of possibilities to have our own subjectivities and individualities reflected in the system.

4.4. General thesaurus

Let us return again to Fig. 4.1. There we can see another way of retrievals, which is shown by broken lines. This is a plan of retrievals to be carried out at a time when our MIR-RF system has reached to a stable state and is intended for the use of visitors who have no RF in MIR-RF system. Essentially the general thesaurus is a summarized and edited accumulation of all researchers' RF's and it will be constructed by using some consistent parts of RF-thesauruses.

Our MIR-RF system has a service to present such an edited hard copy of RF and DF, as in Fig. 4.6*, to the owner of the file. These filed copies will serve as materials not only in constructing the general thesaurus but also in updating and correcting RF's of the researchers. In the figure, the digits 9's denote key strings (with tree structure) selected and organized by the researcher, which are printed in tree forms by computer. The colons in the trees indicate identical key strings.

5. Conclusions

A viewpoint on information storage and retrievals and their systems has been presented, and an outline and function of the MIR-RF system based upon it are illustrated. In addition, a mathematical formulation of files, retrievals and thesauruses

*) The computer program for constructing trees is due to Mr. S. Takeya, a colleague of the first author (S.A).

```
13A   4 1 ON THE LANGUAGES DEFINED BY SENTENTIAL FORMS OF CONTEXT-FREE GRAMMARS.
23A   4 1 ARIKAWA,S.
33A   4 1 RIFIS-RR., KYUSHU UNIV.
43A   4 1 E 1970    1-   12 17.

93A   4 1 1            FORMAL LANGUAGE
93A   4 2 1            LANGUAGE
93A   4 3 11           CONTEXT-FREE LANGUAGE
93A   4 4 11           TYPE II LANGUAGE
93A   4 5 12           SENTENTIAL FORM
93A   4 6 12           QUASI LANGUAGE
93A   4 7 111          QUASI CONTEXT-FREE LANGUAGE
93A   4 8 112          REGULAR SET
93A   4 9 112          REGULAR EVENT
93A   4 A 112          REGULAR LANGUAGE
93A   4 B 121          QUASI CONTEXT-SENSITIVE LANGUAGE
93A   4 C 1211         QUASI CONTEXT-FREE LANGUAGE
93A   4 D 2            DECISION PROBLEM
93A   4 E 21           EMPTINESS PROBLEM
93A   4 F 22           COMPLETENESS PROBLEM
93A   4 G 3            CLOSURE PROPERTY
93A   4 H 3            NON-CLOSURE PROPERTY
```

```
FORMAL          --+-- CONTEXT-FREE     --+-- QUASI
LANGUAGE          |   LANGUAGE           |   CONTEXT-FREE
  :               |     :                |   LANGUAGE
LANGUAGE          |   TYPE II            |
                  |   LANGUAGE           +-- REGULAR SET
                  |                            :
                  |                          REGULAR EVENT
                  |                            :
                  |                          REGULAR
                  |                          LANGUAGE
                  |
                  +-- SENTENTIAL       -----  QUASI               -----  QUASI
                      FORM                    CONTEXT-SENSIT             CONTEXT-FREE
                        :                     IVE LANGUAGE               LANGUAGE
                      QUASI LANGUAGE


DECISION        --+-- EMPTINESS
PROBLEM           |   PROBLEM
                  |
                  +-- COMPLETENESS
                      PROBLEM


CLOSURE
PROPERTY
  :
NON-CLOSURE
PROPERTY
```

Fig. 4.6. An edited copy of RF and DF

is attempted.

In the real implementation of MIR-RF system, we have used FACOM U-200, the small scale computer with 32kw core memory, two units of drums and three units of magnetic tapes. The present MIR-RF system is written by a higher level FORTRAN language. A question-answering mode is adopted. The response time is rather long, which is mainly due to slowness of the access speed of the magnetic tapes used.

We have a plan to proceed with our study and to realize MIR-RF system in computer networks by using a TSS terminal connected with the larger scale computer FACOM 230-75 in Computer Center of Kyushu University. There our accumulated materials on documents will serve as a kind of common resources for researchers of field. And every researcher will be able to search necessary documents from the resources by using his own researcher files.

In the present paper, we have stressed the philosophy and functions of MIR-RF system, but we have not dealt so much with the problem of numerical evaluations. A new evaluation scheme should be established for our MIR-RF system, since our system is very sensitive to the individual researcher file. We have some ideas for a scheme of evaluations [9].

ACKNOWLEDGEMENT

The authors would like to express their sincere thanks to Professor S. Kanō, director of the first author's Institute, for his enthusiastic discussions, and suggestive and valuable advice in the whole course of designing and constructing MIR-RF system.

The authors express their thanks also to Professor C. Asano and to Messrs. S. Takeya and Y. Kai for their kind presentations of materials on documents and discussions.

REFERENCES

[1] Watanabe, T.: KWEST index system — A new attempt of automatic indexing, Dokumen Kenkyu 1974-06 (1974), 195-200 (in Japanese).

[2] Wood, D.: Bibliography 23, Formal language theory and automata theory, Comp. Rev. (1970), 417-430.

[3] Luhn, H. P.: Keyword-in-context index for technical literature (KWIC index), Ameri. Documentation 11 (1960), 288-295.

[4] Gunshi, T.: A pocket inverted English dictionary, Kaibundo (1967).

[5] Arikawa, S.: A study on structures of documents and analysis of title sentences, Report of Special Research Project "Advanced Information Processing of Large Scale Data over a Broad Area", No. 8 (1974), 39-44 (in Japanese).

[6] Wagner, R. A., and Fischer, M. T.: The string-to-string correction problem, JACM 21 (1974), 168-173.

[7] Kitagawa, T.: The efficiency of information storage and retrieval system for scientific research activities, Res. Rept. Funda. Inform. Sci., Kyushu Univ., No. 38 (1974), 1-21.

[8] Stiles, H. E.: The association factor in information retrieval, JACM 8 (1961), 271-279.

[9] Arikawa, S., Takeya, S. and Kai, Y.: Numerical evaluations of MIR-RF system (in prep.).