# On the optimization of library information retrieval systems

By A. M. Iványi and A. N. Sotnikow

## Introduction

In some information retrieval systems (e.g. in many library ones) the content of documents (books, papers, patents, computer programs etc.) and the content of user's questions are characterized by using a given set of key-words (so called descriptors). The aim of the retrieval is to find all the documents whose content is characterized by the descriptors in the question and, may be, by other descriptors too.

Of course, the sequential retrieval represents a slow method, which is satisfactory only in small systems.

A more effective method is proposed by G. Salton [1]. According to this method the set of documents is decomposed into disjoint subsets (so called clusters) consisting of "similar" documents, and the retrieval is organized in a hierarchical manner. Recently an excellent review was publicated by S. T. March on the different retrieval methods [2].

According to Salton's proposition [3] the efficiency of the retrieval system is characterized by the expected waiting time of the users, therefore the aim of the optimization is to minimize this time choosing suitable parameters (number of clusters, sizes of clusters, distribution of documents among the clusters).

At first in § 1 we describe a mathematical model of such systems [1], then in § 2 the mentioned expected time is computed for some values of parameters, later in § 3 this time is analysed as a function of the different parameters, and finally in § 4 some practical conclusions are made.

## § 1. The mathematical model

Let $m$, $n$, $r$ and $z$ be positive integers, $\delta=\{D_0, \dots, D_i, \dots, D_{n-1}\}$ — documents, $\varkappa=\{K_0, \dots, K_j, \dots, K_{m-1}\}$ — key-words, $\varrho=\{Q_0, \dots, Q_p, \dots, Q_{r-1}\}$ — possible questions of users, $\alpha=\{A_0, \dots, A_p, \dots, A_{r-1}\}$ — the corresponding answers of the information retrieval system, where $Q_p \subseteqq \varkappa$, $A_p \subseteqq \delta$ for $p=0, \dots, r-1$.

The content of documents is characterized by the document description matrix $\mathbf{D}=[d_{ij}]_{n\times m}$, where $d_{ij}=1$, if $K_j$ characterizes $D_i$, and $d_{ij}=0$ otherwise. The content of questions is characterized by using the question description matrix $\mathbf{Q}=[q_{pj}]_{r\times m}$, where $q_{pj}=1$, if $K_j$ characterizes $Q_p$, and $q_{pj}=0$ otherwise. The $i$-th row $\mathbf{d}_i$ of $\mathbf{D}$ represents the description of $D_i$, and the $p$-th row $\mathbf{q}_p$ of $Q$ represents the description of $Q_p$. In the case of the question $Q_p$ we have to find the documents $D_i$, for which $\mathbf{d}_i \geqq \mathbf{q}_p$.

The set $\delta$ is decomposed into disjoint clusters $C_0, \ldots, C_k, \ldots, C_{z-1}$. The decomposition is defined using the decomposition matrix $\mathbf{Y}=[y_{ik}]_{n\times z}$, where $y_{ik}=1$, if $D_i \in C_k$ and $y_{ik}=0$ otherwise. The content of clusters is described by using the characteristic matrix $\mathbf{U}=[u_{kj}]_{z\times m}$, where $u_{kj}=0$ if for any $D_i \in C_k$, $d_{ij}=0$, and $u_{kj}=1$ otherwise. The $k$-th row $\mathbf{u}_k$ of $\mathbf{U}$ is called the characteristic vector of $C_k$.

The retrieval consists of two levels: on the first level the description $\mathbf{q}_p$ of the question $Q_p$ is compared with the characteristic vectors $\mathbf{u}_k$ for $k=0, \ldots, z-1$. On the second level a sequential retrieval is realized in the relevant clusters: a cluster $C_r$ is relevant to $Q_p$ iff $\mathbf{u}_k \geqq \mathbf{q}_p$ $(u_{k,j} \geqq q_{p,j}$ for $j=0, 1, \ldots, m-1)$.

As a time unit (and as a cost unit, too) let us choose the time required to compare two $m$-dimensional vectors. Then in the case of a decomposition matrix $\mathbf{Y}$ the cost $S_{p\mathbf{Y}}$ of the answer $A_p$ to a question $Q_p$ is defined by

$$S_{p\mathbf{Y}} = z + \sum_{k=0}^{z-1} |C_k| L_{kp} + t \sum_{k=0}^{z-1} L_{kp},$$

where $L_{kp}=1$, if $C_k$ is relevant to $Q_p$ and $L_{kp}=0$ otherwise, $t$ is a nonnegative real parameter (the time, required to prepare a new cluster to the processing). The efficiency of a given decomposition algorithm $A$, resulting a decomposition matrix $\mathbf{Y}_A$ is denoted by $\mathbf{M}(\mathbf{Y}_A)$ and defined by the average cost of answers to all possible questions

$$\mathbf{M}(\mathbf{Y}_A) = \frac{1}{r} \sum_{p=0}^{r-1} S_{p\mathbf{Y}}.$$

For example, let $m=2$, $n=r=4$, $z=2$, $t=0$,

$$\mathbf{D} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{Q} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Then the corresponding answers are $A_0=\{D_0, D_1, D_2, D_3\}$, $A_1=\{D_1, D_3\}$, $A_2=\{D_2, D_3\}$ and $A_3=\{D_3\}$.

Let us assume that we construct two different decompositions defined by the matrices

$$\mathbf{Y}_1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{Y}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Then the corresponding characteristic matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ are as follows:

$$\mathbf{U}_1 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{U}_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

If all the four possible questions are put once, then we have eight comparisons on the first level for both decompositions; the second decomposition requires $4\times 4=$ $=16$ ones on the second level, while the first needs only $2\times 4+2\times 2=12$, therefore, $\mathbf{M}(\mathbf{Y}_1)=\frac{20}{4}=5$, while $\mathbf{M}(\mathbf{Y}_2)=\frac{24}{4}=6$.

## § 2. Efficiency of a decomposition algorithm

Let $m$ be fixed, $n=r=2^m$. Let us suppose, that the descriptions of the documents (the rows of the matrix $\mathbf{D}$) and the descriptions of the questions (the rows of the matrix $\mathbf{Q}$) are different, that is $0\leqq i<I\leqq n-1$ implies $\mathbf{d}_i\neq\mathbf{d}_I$, and $0\leqq p<P\leqq$ $\leqq r-1$ implies $\mathbf{q}_p\neq\mathbf{q}_P$. In this case we have $2^m$ possible different documents and also $2^m$ different questions. For the simplicity let us suppose, that the descriptions as binary numbers give the corresponding indices, that is

$$i=\sum_{j=0}^{m-1} d_{ij}2^{m-1-j},\quad p=\sum_{j=0}^{m-1} q_{pj}2^{m-1-j}\quad (i, p=0,\dots,2^m-1).$$

Let us suppose that $x$ is a nonnegative integer and $|C_k|=2^x$ for $k=0,1,\dots$ $\dots,z-1$. In this case we have $z=2^{m-x}$ and $0\leqq x\leqq m$.

For this special values of parameters the decomposition algorithm $S$ [4] is defined as follows:

$$C_k=\{D_{k\cdot 2^x}, D_{k\cdot 2^x+1},\dots,D_{k\cdot 2^x+2^x-1}\}\quad (k=0,1,\dots,2^{m-x}-1).$$

Now we can give the efficiency of $S$ as a function of parameters $x$, $m$ and $t$ [4].

**Theorem 1.** If $x, m$ are positive fixed integers, $n=r=2^m$, $t$ is a nonnegative real number, $0\leqq i<I<n$ implies $\mathbf{d}_i\neq\mathbf{d}_I$, $0\leqq p<P<r$ implies $\mathbf{q}_p\neq\mathbf{q}_P$, then

$$\mathbf{M}(\mathbf{Y}_S)=2^{m-x}+2^{2x-m}3^{m-x}+t\cdot 2^{x-m}3^{m-x}.\quad\Box\qquad (1)$$

*Proof.* According to the definition $\mathbf{M}(\mathbf{Y}_S)$ we have

$$\mathbf{M}(\mathbf{Y}_S)=\frac{1}{2^m}\Big[2^{2m-x}+2^x\sum_{k=0}^{z-1}\sum_{\mathbf{q}_p\leqq\mathbf{u}_k}1+t\sum_{k=0}^{z-1}\sum_{\mathbf{q}_p\leqq\mathbf{u}_k}1\Big].$$

In this expression the sum $\sum_{\mathbf{q}_p\leqq\mathbf{u}_k}1$ is equal to the number of questions for which $(q_{p,0},\dots,q_{p,m-1})\leqq(u_{k,0},\dots,u_{k,m-1})$.

Considering $\mathbf{u}_k$ as a binary number $b_k$ according to the grouping rule of $S$ we get

$$b_k=\sum_{j=0}^{m-1}u_{k,j}2^{m-j-1}=(k+1)2^x-1$$

(the first $m-x$ digits determine $k$, the last $x$ digits are 1's).

If $\mathbf{u}_k$ contains $w$ 1's on the first $m-x$ places, then we have $2^w\cdot 2^x$ questions with $\mathbf{q}_p\leqq\mathbf{u}_k$. Therefore, using the equality

$$\sum_{w=0}^{m-x}\binom{m-x}{w}2^w=(2+1)^{m-x}$$

we get

$$\mathbf{M}(\mathbf{Y}_S)=2^{m-x}+(2^{x-m}+t\cdot 2^{-m})\sum_{w=0}^{m-x}\binom{m-x}{w}2^w 2^x=$$

$$=2^{m-x}+2^{2x-m}3^{m-x}+t\cdot 2^{x-m}3^{m-x}. \quad \square$$

## § 3. Analysis of the influence of the continuous parameters

According to the Theorem 1 the cost $\mathbf{M}(\mathbf{Y}_S)$ depends on $x$, $m$ and $t$ as

$$\mathbf{M}=\mathbf{M}(x, t, m)=e^{mb}e^{-xb}+e^{m(c-b)}e^{x(2b-c)}+t\cdot e^{m(c-b)}e^{-x(c-b)},$$

where $b=\ln 2\approx 0{,}69314718$ and $c=\ln 3\approx 1{,}09861229$, $t$ is a nonnegative real parameter, $m$ is a positive integer parameter, and $x$ represents the independent variable being integer and $0\leqq x\leqq m$.

In the practice the number of key-words $m$ equals 10—1000 [4, 5], the time of preparation of a new cluster to the processing $t$ is about 0—$10^9$ time units (may be, it is necessary to put new disc or magnetic tape).

In the analysis of the function $\mathbf{M}=\mathbf{M}(x, t, m)$ its continuous variant $\mathbf{R}=\mathbf{R}(x, t, m)$, plays an important role, where $x$, $t$ and $m$ are real variables with $x\in(-\infty, +\infty)$, $t\in(-\infty, +\infty)$, $m\in(-\infty, +\infty)$.

**Lemma 1.** If $t\geqq 0$ and $m$ are fixed real numbers, then the function $\mathbf{R}=\mathbf{R}(x, t, m)$ as a function of $x$ is convex, it has one local minimum and this minimum represents an absolute one too. □

*Proof.* Differentiating $\mathbf{R}(x, t, m)$ by $x$ we get

$$\mathbf{R}_x=-b\cdot e^{mb}e^{-bx}+(2b-c)e^{m(c-b)}e^{x(2b-c)}-(c-b)t\cdot e^{m(c-b)}e^{-x(c-b)}.$$

Since $\mathbf{R}_x$ is a monotone increasing continuous function of $x$ (its members are increasing) and its limit is $+\infty$ for $x\to+\infty$ and $-\infty$ for $x\to-\infty$, so $\mathbf{R}_x$ has a unique zero-place $x_0$ and $\mathbf{R}_x<0$ for $x<x_0$, $\mathbf{R}_x\Rightarrow 0$ for $x>x_0$. Therefore $\mathbf{R}(x, t, m)$ is convex, and it has an absolute minimum at $x_0$. □

**Lemma 2.** For a fixed $m$ the function $x_0=x_0(t, m)$ is a monotone increasing function of $t$. □

*Proof.* From $\mathbf{R}_x=0$ we get

$$t=t(x_0, m)=\frac{2b-c}{c-b}e^{bx_0}-\frac{b}{c-b}e^{m(2b-c)}e^{-x_0(2b-c)}.$$

For a fixed $m$ the function $t(x_0, m)$ is a monotone increasing one of $x_0$ for $x_0\in(-\infty, +\infty)$, so it has a monotone increasing inverse $x_0=x_0(t, m)$. □

**Lemma 3.** For fixed $t\geqq 0$ the function $x_0=x_0(t, m)$ is a monotone increasing function of $m$. □

*Proof.* It follows from $\mathbf{R}_x=0$ that

$$m=\frac{1}{2b-c}\ln\left[\frac{2b-c}{b}\,e^{x_0(3b-c)}-t\,\frac{c-b}{b}\,e^{x_0(2b-c)}\right].$$

According to Lemma 1 for given $t$ and $m$ there exists a corresponding $x_0$, therefore the expression in the square brackets has to be positive (otherwise $m$ has no meaning).

Now we have

$$\frac{\partial m}{\partial x_0}=\frac{\frac{2b-c}{b}(3b-c)\,e^{(3b-c)x_0}-\frac{t(c-b)(2b-c)}{b}\,e^{x_0(2b-c)}}{(2b-c)\left[\frac{2b-c}{b}\,e^{x_0(3b-c)}-\frac{t(c-b)}{b}\,e^{x_0(2b-c)}\right]}.$$

The expression in the numerator is greater than the one in the square brackets of the denominator (the positive member is multiplied by $3b-c$, the negative one by $2b-c$), therefore $\frac{\partial m}{\partial x_0}>0$ and so the derivative $\frac{\partial x_0}{\partial m}$ also is everywhere positive and $x_0$ is an increasing function of $m$. □

**Theorem 2.** If $t\geqq 0$ and $m$ are fixed real numbers, then

$$x_0\begin{cases}=\beta m+\gamma, & \text{if } t=0,\\ >\beta m+\gamma, & \text{if } t>0,\end{cases}$$

where $\beta=\frac{\ln 4/3}{\ln 8/3}\approx 0{,}29330495$ and $\gamma=\frac{\ln\frac{\ln 2}{\ln 4/3}}{\ln 8/3}\approx 0{,}89657440$. □

*Proof.* Lemma 1 and the equation $\mathbf{R}_x=0$ at the condition $t=0$ imply $x_0=\beta m+\gamma$, and then Lemma 2 gives the remaining part of the theorem. □

## § 4. Conclusions

In the previous paragraph we considered $x$, $t$ and $m$ as continuous parameters. In the practice $x$ and $m$ have positive integer values.

According to Lemma 1, $\mathbf{R}(x, t, m)$ as a function of $x$ is a convex function, therefore Theorem 2 implies for the optimal size-parameter $x_{\text{opt}}$ the assertion: if $t=0$, then

$$\lfloor\beta m+\gamma\rfloor\leqq x_{\text{opt}}\leqq\lceil\beta m+\gamma\rceil.$$

If $t>0$ then we have a lower bound $x_{\text{opt}}\geqq\lfloor\beta m+\gamma\rfloor$ due to Lemma 2.

Analysing the function $\mathbf{M}(x, t, m)$ one can see that for small values of $t$ the third member of (1) has a relatively weak influence on the value of $x_{\text{opt}}$, and so in the first approximation is neglectable.

We propose the following approximate formula:

$$x_{\text{opt}}\approx\lfloor 0{,}29330495m+0{,}89657440\rfloor. \tag{2}$$

Table 1. Summary of the numerical results

| $m$ | $t$ | $x_{opt}$ | $x_a$ | $m$ | $t$ | $x_{opt}$ | $x_a$ |
|---|---|---|---|---|---|---|---|
| 20 | $10^0$ | 7 | 7 | 80 | $10^0$ | 24 | 24 |
| | $10^2$ | 8 | 7 | | $10^2$ | 24 | 24 |
| | $10^4$ | 14 | 7 | | $10^4$ | 24 | 24 |
| | $10^6$ | 20 | 7 | | $10^6$ | 24 | 24 |
| 40 | $10^0$ | 13 | 13 | 100 | $10^0$ | 30 | 30 |
| | $10^2$ | 13 | 13 | | $10^2$ | 30 | 30 |
| | $10^4$ | 14 | 13 | | $10^4$ | 30 | 30 |
| | $10^6$ | 20 | 13 | | $10^6$ | 30 | 30 |
| 60 | $10^0$ | 19 | 19 | | | | |
| | $10^2$ | 19 | 19 | | | | |
| | $10^4$ | 19 | 19 | | | | |
| | $10^6$ | 21 | 19 | | | | |

The following table shows some numerical results of a BASIC-program, running on a personal computer Commodore-64.

In this table $x_{opt}$ represents the optimal integer value of $x$ and $x_a$ is the value resulted by the approximate formula (2).

According to the presented numerical data formula (2) results good approximations of the optimal size parameter for $m > 20$ and $t < 1\,000\,000$.

The authors are indebted to Dr. Péter Simon at Eötvös Loránd University of Budapest for the consultation and valuable remarks.

## References

[1] Salton, G., Automatic information organization and retrieval. McGraw Hill, New York, 1968.

[2] March, S. T., Techniques for structuring databese records. Computing Surveys, *15* (1) (1983), 45—79.

[3] Salton, G., Dynamic information and library processing. Prentice Hall Inc., Englewood Cliffs, 1975.

[4] Ивани А. М. и Сотников А. Н. Об оптимизации дескрипторных АИПС с зонно-иерархической организацией поискового множества. Вестник Московского Университета, Сер. 15. Вычислительная Математика и Кибернетика, (2) (1984) 53—57.

[5] Iványi, A. M. and Nyirádi, L., Library retrieval systems INF and FARMDOC. In: Proc. of Conf. for information-servicing of professionally linked computer users (Varna, 1977), 307—313.

*(Received March 13, 1985)*