C. NEGOIȚĂ

Strategy of Probabilistic-Type Searching for Information Retrieval Systems

Under the conditions of the present technical scientific revolution, when science is growing into a direct productive force and the quantity of information carriers increases at an impressive rate, the matter of information retrieval acquires an exceptional importance. By using modern mathematics, science offers a number of strategies meant to secure the harmonizing of questions with the stored documents.

Information retrieval is defined as an act of retrospective searching. This act may directly deliver the information (data retrieval) or may deliver the document containing the information (document retrieval). Reference is made below only of the document retrieval systems.

To the input questions, these systems feed out sets of documents or lists of documents providing the requested information. Under these circumstances the tasks of the retrieval system involve the marking of documents — namely to index them, to store them and to memorize the procedure used to recover them easily.

Thus, by retrieval process we shall understand the concordance between a *question* — expressed in the form of finite sets of identifiers — and a collection of *documents* (stored under the form of a set of identifiers of the documents content). The common property of the archives elements belonging to the reply is called relevance as against the question. The set of questions documents makes up the archives and the subset of the archives, which agrees with the question, makes up a *reply*. The concordance of questions with the stored documents proceeds now currently upon several strategies: matrix, associative, of grouping or of probabilistic type.

In the event of probabilistic strategies, each question put to the system specifies a distribution of probabilities for the entire archives. The system

has to find an optimum reply to a question posed on the basis of its distribution of probabilities.

Definition I

A measure is a function of sets with real values, fully additive and non-negative, determined on a set M, so that $\mu(0) = 0$.

In a given set of sets M, for each $N \in M$ a real number $\mu(N)$ is established which complies with the following conditions:

1. for each $N \in M, \mu(0) = 0$
2. for each N_1 and $N_2 \in M$ so that $N_1 \leqslant N_2$ when

$$\mu(N_1) \leqslant (N_2)$$

3. for each $N \in M$ its complement N' is in M and

$$\mu(N) + \mu(N') = \mu(M)$$

4. if $N_i (i = 1, 2, ..., k)$ is a series of disjoint subsets of M, then

$$s = \sum_{i=1}^{k} N_i$$

is an element of set M and

$$\mu(S) = \sum_{i=1}^{k} \mu(N_i) = \mu \sum_{i=1}^{k} N_i$$

Definition II

The relevance is a measure of the information carried by a document x in connection with a certain question q.

In a given S archives and a question q, we say that $x \in S$ is relevant to q if

$$\mu(x) > \xi_0$$

in which ξ_0 is a real, non-negative number, equal or below one.

$\mu(x)$ may be considered as a measure of probability without being deprieved of generality.

Strategy I

If we take a set Q, a set S and a transformation E of set Q into set $P(S)$ of the sections of set S, the members of set Q represent questions, those of set S represent documents and transformation E is a function of a definite estimate on set $P(S)$ in the terms of a measure of relevance $\mu(x)$ on S.

The correspondence between $q \in Q$ and subset $A(q) \in P(S)$, called reply to q, is established by maximizing E.

$A(q)$ consists of all the members $x \in S$ whose measure of the relevance

$$\mu(x) > \xi_0$$

since E is defined as a function of monotonously growing sets, so that $E = 0$ for $\mu(x) = \xi_0$

The valuation function may be of the form

$$E(X) = \sum_{x \in X} V_q(x)$$

in which $V_q(x)$ is of the form

$$V_q(x) = \alpha F_q(x) - \beta[1 - F_q(x)]$$

$F_q(x)$ is the probability that $x \in S$ should be relevant as against a question q;

α, β are non-negative constants of the system.

In this circumstance

$$\xi = \frac{\beta}{\alpha + \beta}$$

is called critical probability.

We mark the answer $A(q)$ with an arbitrary sequence

$$A(q) = (x_0, x_1, \ldots, x_n).$$

Definition III

Two elements x_{k-1}, x_k are informational if

$$p_{k-1,k} \geqslant r_{k-1,k}$$

$$p_{k,k-1} \geqslant r_{k,k-1}$$

in which $p_{k-1,k}$ is the conditional probability for x_k to be relevant to q if x_{k-1} is relevant to q.

$r_{k-1,k}$ is the conditional probability for x_k to be relevant to q if x_{k-1} is non-relevant to q.

THEOREM I

The necessary and sufficient condition requested for the elements of a set $\bar{A} \leqslant A$ to be connected to q, irrespective of the sequence in which they are surveyed, is that these elements be informational pairs.

Definition IV

A set X is closed when no $y \in X$ is informational with each $x \in X$.

The closure $\bar{A}$ of x is the smallest closed set containing element x, whose members are informational pairs.

Strategy II.

The best reply consists of the closure of that member of A provided with the highest probability F.

We mark this set by $\bar{A}_1$.

The next good reply is the closure of $x \in A - \bar{A}_1$ having the highest probability F.

We mark this set by $\bar{A}_2$.

Continuing this process, an ordering of set A in disjoint subsets is reached, each of them having elements whose relevance is independent of the elements sequence.

Hence, with a function $F(x)$ defined on S, which shows the relevance measure of $x \in S$ with regard to q, irrespective of the relevance of other members of S, the following conditions are verified:

(a) $F(x) > \xi$ for each $x \in X$

(b) $P_{k-1,k} \geqslant r_{k-1,k}$ $P_{k,k-1} \geqslant r_{k,k-1}$ and for each x_k, x_{k-1}, in X

The optimum reply $A(q)$ to a question q is obtained by determining a set $\bar{A} \leqslant A$ closed as against property (b) and which contains those $x \in A$ for which $F(x)$ is maximum.

Definition V

An equivalence relation between the elements x_k, x_{k-1} of a set is a binary relation (between each two elements); it is *reflexive* ($x_k R x_k$ exists), *symmetrical* ($x_k R x_{k-1}$ implies $x_{k-1}\ R x_k$) and *transitive* ($x_k R x_{k-1}$ and $x_{k-1} R x_{k-2}$ implies $x_k R x_{k-2}$).

Definition VI

Two elements x_k, x_{k-1} are converse if

$$p_{k-1,k} \geqslant \xi_0$$

$$p_{k,k-1} \geqslant \xi_0$$

If we mark by X a set with informational pair elements, and by Y a set with converse pair members, then

$$X \subset Y$$

The conversion relation is not an equivalence relation, because if

$$p_{k-1,k} > \xi_0$$

$$p_{k,k-1} > \xi_0$$

$$p_{k,k+1} > \xi_0$$

$$p_{k+1,k} > \xi_0$$

that does not mean that

$$p_{k-1,\,k+1} > \xi_0$$

$$p_{k+1,\,k-1} > \xi_0$$

Definition VII

A chain through x is any sequence containing x, so that

$$x_k \quad \text{and} \quad x_{k+1}$$

are converse.

THEOREM II

The set of all chains traversing x forms an equivalence class.
The dependence relation to a chain is an equivalence relation, since

$$x_k \sim x_k$$

$$\text{implies } x_{k-1} \sim x_k$$

$$x_k \sim x_{k-1}$$

If there is a chain from x_k to x_{k+1} and a chain from x_{k+1} to x_{k+2}, then there is a chain from x_k to x_{k+2}.
We state that two elements belonging to a chain intercommunicate.

Strategy III

Since the conversion implies communication, set S may be covered by a set of disjoint classes $C_1, C_2, \ldots, C_n$ with the property that if $x \in C_1$, all the members of S converse with x and, consequently, informational with x, are also in C_1.
Due to the nature of an equivalence relation, each class may be represented by an arbitrary element.
A set $G_1 = (C_{11}, C_{12}, \ldots, C_{1n})$ is thus formed.
Instead of questioning set S, the system will question set G_1, with three possibilities.

1. there is no element C of set G_1, so that $F_q(C) > \xi_0$,
2. there is an element C of set G_1, so that $F_q(C) > \xi_0$,
3. there are several elements C of set G_1, so that $F_q(C) > \xi_0$.

If the first instance occurs a second set is chosen, namely

$$G_2 = (C_{21}, C_{22}, \ldots, C_{2n})$$

and so on, until the second or third instance occurs.
If the second instance occurs, there is a class C_i with the biggest density of elements relevant to q. The system will then determine the set $A_i < C_i$, that is the elements of class C_i, so that $F_q(x) > \xi_0$.
The third instance implies the existence of several classes with equal densities. The system must decide on the introduction in the class containing

the element of G with the highest probability, then to proceed as in the second instance.

CONCLUSIONS

Strategy I is conducive to a reply in which the elements are introduced on the basis of their absolute probability, each element being considered as independent towards the other elements of the set.

There are circumstances when the elements of the reply are not independent, their relevance varying as to the position of the element within the sequence.

Strategy II reveals an optimum ordering of the reply in closed disjoint subsets, not affected by the order of the elements' survey. Yet, the system is poorly efficient, considering that for a question the whole archives must be surveyed.

Strategy III eliminates this disadvantage too, for it enables the direct obtaining of a subset containing all the informational elements with a given element.

For a person wishing to have a „literature" in connection with a subject-matter, strategy I is sufficient. For classification, most useful is strategy II. Most cases of retrieval imply a strategy of the third type, which enables a rapid access to the storage.

REFERENCES

1. BECKER, J., HAYES, R.M. *Introduction to Information Storage and Retrieval—Tools, Elements, Theories*, Wiley, New York, 1963.
2. GHIKA, A. *Analiza Funcțională*, Editura Academiei R.S.R., București, 1967.
3. GOFFMAN, W. *On Relevance as a Measure*, Information Storage and Retrieval, vol. 2, 1964, p. 201.
4. BELZER, J., GOFFMAN, W. *Theoretical Considerations in Information*, Retrieval Systems, Communications of the ACM, vol. 7, No. 7, July 1964, p. 439.
5. GOFFMAN, W. *Probabilistic Models in Information Retrieval*, Information Retrieval in Action, Western Reserve University Press, Cleveland, 1963.
6. GOFFMAN, W.A *Searching Procedure for Information Retrieval*, Information Storage and Retrieval, vol. 12, 1964, p. 73.

SUR LA STRATEGIE DE RECHERCHE DE TYPE PROBABILISTIQUE POUR LES SYSTEMES DE REUTILISATION DES INFORMATIONS

L'ouvrage porte sur les systèmes informationnels qui fournissent à l'issue des listes de documents contenant des informations pertinentes au sujet d'une demande d'information introduite à l'entrée.

Après avoir défini le processus de réutilisation de l'information comme la pose d'une question, représentée comme un ensemble fini de descripteurs avec une collection de documents enmagasinés dans la mémoire du système par des identifieurs, et la relèvance comme une mesure

de l'information portée par un document au sujet d'une certaine question posée au système, après avoir donc défini ces deux notions, l'auteur examine trois stratégies de recherche, à savoir: la première, qui conduit à une réponse dont les élements sont indiqués sur la base de leurs probabilités absolues; la seconde, permettant la meilleure ordination de la réponse en ensembles fermés disjoints; la troisième permettant le meilleur groupement des éléments de telle manière que le système ne doit plus parcourir toute l'archive pour une demande d'information.

О СТРАТЕГИИ ПОИСКА ВЕРОЯТНОСТНОГО ТИПА ДЛЯ СИСТЕМ СБОРА ИНФОРМАЦИИ

Статья касается информационных систем, обеспечивающих выход списков документов, которые содержат точную информацию на требование информации, введенное у входа.

После определения процесса сбора как согласования какого-либо вопроса, представленного как множество, установленное описанием, с коллекцией документов, хранящихся в памяти системы при помощи отождествителей, а также и выявления как меры информации, содержащейся в одном документе в связи с определенным вопросом, заданным системе, экзаменуются три стратегии поиска, а именно: стратегия I ведет к ответу, в котором элементы специфицируются на основе их абсолютных вероятностей; стратегия 2 позволяет оптимально упорядочить ответ в выключенные закрытые множества; стратегия 3 разрешает оптимальную группировку элементов таким образом, что система не должна больше проходить весь архив для одного требования информации.