R.I.R.O.
(1re année, nº 3, 1967, p. 63-70)

THEORIE DES SYSTEMES AUTOCODEURS

par Jacques F. VALLÉE (1)

SOMMAIRE

L'article propose une représentation théorique des systèmes de traitement de l'information qui donnent des réponses directes à des requêtes adressées à leurs catalogues de base. On démontre que dans le cas des interrogations séquentielles le problème de traduire un univers de données en un catalogue optimum qui servira de base à l'automate chercheur admet une solution. Cette solution s'obtient en composant la restriction des données d'origine à l'espace-requête et l'opérateur canonique d'une relation d'équivalence convenablement définie. La définition des systèmes autocodeurs est basée sur cette propriété, qui rend ces systèmes particulièrement applicables aux problèmes de télégestion.

THEORIE DES SYSTEMES AUTOCODEURS

L'avènement de la télégestion a ouvert la voie de la réalisation pratique pour une série de systèmes qui n'ont été étudiés jusqu'ici que d'une manière très expérimentale et sur une échelle limitée : les systèmes qui donnent accès direct à un large ensemble de données comprenant des centaines de milliers ou des millions d'enregistrements, et permettent l'interrogation de cet ensemble soit en langage naturel, soit dans un code utilisant des termes techniques naturels. Des systèmes du premier type, qui acceptent des questions directement en Anglais, ont été développés aux États-Unis au cours des dernières années. En 1961, des chercheurs du M.I.T. présentèrent un premier programme de ce type, BASEBALL [1] dont la structure était basée sur la notion de « liste de spécification ». L'auteur du présent article généralisa cette méthode et l'appliqua à un problème de recherche scientifique avec un programme appelé ALTAIR (Automatic Logical Translation And Information Retrieval) dont les résultats dans le domaine astronomique ont été décrits ailleurs [2] ainsi que l'organisation du programme lui-même [3]. ALTAIR a été écrit pour une machine

(1) Ph. D. Systems Analyst, Vogelback Computing Center, Northwestern University, Evanston, Illinois, U.S.A.

CDC 3 400 et traduit des questions en Anglais (à la vitesse de 250 questions/minute) dans un langage artificiel intermédiaire. Dans le présent article, nous avons pour objectif de donner une théorie plus précise des méthodes d'optimisation qui s'appliqueront à cette traduction lorsque l'usage de tels programmes se généralisera et que de sérieux problèmes de temps de calcul se poseront.

Aussi bien dans ALTAIR que dans BASEBALL, en effet, la dimension de l'ensemble de base est fixe, tous les chiffres significatifs des codes (« signatures ») de chaque item sont testés par le programme chercheur pour obtenir la réponse à chaque question : en d'autres termes, le langage artificiel de cet automate est constant. Dans cet article nous proposons l'idée de rendre ce langage artificiel variable pour obtenir une optimisation de la traduction des questions conduisant à une économie considérable au moment de l'exécution. Ces considérations nous conduisent aux définitions suivantes :

Définition 1

Étant donné un objet R et n ensembles d'attributs : A_1, $A_2 \cdots A_n$ dont les éléments peuvent être des nombres entiers ou réels, des symboles définis sur un alphabet quelconque, etc. avec la seule restriction qu'il existe dans A_h $(h = 1, \ldots n)$ une relation d'ordre $\prec$ telle que, étant donné deux éléments x_1 et x_2 appartenant à A_h, ou bien $x_1 \prec x_2$, ou $x_1 = x_2$, ou $x_2 \prec x_1$, nous noterons

$$C(R) = (r_1,\ r_2 \ldots,\ r_n)$$

le point représentatif de R dans $\prod_{h=1}^{n} A_h$.

Nous appellerons $C(R)$ *l'enregistrement de R par rapport* à la collection $\mathcal{F} = A_1 \times A_2 \times \ldots \times A_n$ et n est appelé la *dimension* de $\mathcal{F}$.

L'interprétation de cette définition est immédiate : Nous exigeons que tout objet figurant dans la collection soit identifié par ses attributs, que les ensembles sur lesquels ces attributs prennent leurs valeurs soient totalement ordonnés, et nous notons la dimension maximum de la collection. L'optimisation consiste à réduire cette dimensionalité en vue de résoudre une suite donnée de questions.

Définition 2

Étant donné n ensembles A_k totalement ordonnés et un ensemble de m triplets :

$$q_i = \{ (k^1, a_i^{k^1}, b_i^{k^1}), (k^2, a_i^{k^2}, b_i^{k^2}), \ldots (k^m, a_i^{k^m}, b_i^{k^m}) \}$$

Nous dirons que q_t est une *question de profondeur m par rapport a la collection* $\mathcal{F}$ si q_t vérifie les quatre conditions suivantes :

1. $k^\alpha \neq k^\beta$ pour tout $\alpha \neq \beta$
2. $m \leqslant n$
3. Pour tout k_j, a_t^{kj} et b_t^{kj} sont des éléments de l'ensemble A_{kj}.
4. $a_t^{kj} \leqslant b_t^{kj}$

Dans ce qui suit on notera Y l'ensemble des questions.

Ces notations prennent leur sens quand on les applique a un exemple donné. Supposons que nous disposions d'un catalogue astronomique donnant la magnitude, le spectre et la distance de chaque étoile, et que l'on veuille obtenir le nombre d'objets dont la magnitude est entre 5 et 6 et qui sont situés entre 20 et 100 années-lumière du Soleil. Cette question serait représentée par l'ensemble :

$$q_1 = \{(1, 5, 6), (3, 20, 100)\}$$

puisque A_1 est l'ensemble des magnitudes et A_3 l'ensemble des valeurs de la distance.

Dans une seconde question, il se peut que l'on demande le nombre des étoiles dont le spectre est entre les codes 23 et 35 et dont la distance est entre 50 et 200 années-lumière. On aurait alors :

$$q_2 = \{(2, 23, 35), (3, 50, 200)\}$$

Il est intéressant de définir maintenant de telles suites de questions :

Définition 3

Étant donné n ensembles totalement ordonnés et une suite de questions par rapport à la collection qu'ils définissent, nous appelons une telle suite une *interrogation* de la collection et nous la notons :

$$Q = (q_1, q_2, \ldots q_L)$$

On observe maintenant que toutes les questions d'une interrogation donnée ne s'adressent pas en général à l'ensemble de la collection — d'où une réduction considérable de la dimension du problème si l'on tire parti des répétitions observables :

Définition 4

Étant donné une question q, soit f la fonction :

$$f : \mathcal{F} \rightarrow [\{0, 1\}]^n$$

telle que, étant donné un objet quelconque R, on ait :

$$f(R) = (\delta^1, \delta^2, \ldots \delta^n)$$

avec $\begin{cases} \delta^h = 1 \text{ si l'ensemble } A_h \text{ apparaît dans la question } q, \text{ c'est-à-dire} \\ \text{s'il existe un entier } k^j \text{ dans } q \text{ tel que } A_{kj} = A_h \\ \delta^h = 0 \text{ dans le cas contraire.} \end{cases}$

Nous appelons f la *fonction-requête* de la question q.

La fonction que nous venons de définir associe un index avec chaque ensemble d'attributs : cet index est 1 si cet attribut est utilisé dans la question et 0 dans le cas contraire. Pour la question q_1 prise pour exemple plus haut, et pour une étoile donnée, on aurait ainsi :

f_1 (étoile) = (1, 0, 1) puisque A_2 (spectre) n'est pas utilisé.

De même, pour la seconde question :

$$f_2 \text{ (étoile)} = (0, 1, 1)$$

L'utilisation judicieuse des fonctions-requêtes permet une première réduction de la dimension du problème pour une interrogation donnée. Il est donc naturel d'introduire la définition suivante :

Définition 5

Étant donné une interrogation $Q = (q_1, q_2, \ldots q_L)$,
soit $\Phi(Q)$ l'expression :

$$\Phi(Q) = f_1 \vee f_2 \vee \ldots \vee f_L = (\varphi^1, \varphi^2, \ldots \varphi^n)$$

f_i étant la fonction-requête de la question q_i, et soit $\mathfrak{U}$ le produit de tous les ensembles A_h tels que $\varphi^h = 1$.

Soit p le nombre de tels ensembles : nous appelons p *l'étendue de l'interrogation* Q par rapport à la collection $\mathcal{F}$ et nous appelons $\mathfrak{U}$ l'*Espace-Requête* défini par Q sur la collection $\mathcal{F}$. On a évidemment :

$$\dim \mathfrak{U} = p$$

A chaque point R de $\mathcal{F}$ correspond un point représentatif $P(R)$ dans l'Espace-Requête.

A l'aide de ces notations, nous pouvons formuler mathématiquement de manière simple le problème de la représentation optimale de l'information dans une interrogation donnée, c'est-à-dire dans le cas des questions séquentielles :

Problème

Soit N l'ensemble des entiers positifs. Étant donné une interrogation Q d'étendue p par rapport à la collection $\mathcal{F}$, trouver une fonction $\sigma : \mathcal{F} \rightarrow N^p$ telle que, étant donné deux éléments R et R' de $\mathcal{F}$, leurs images $\sigma(R)$ et $\sigma(R')$ ne soient distinctes que si les deux conditions suivantes sont vérifiées :

1) $\Phi(Q) \times C(R) \neq \Phi(Q) \times C(R')$ (produit scalaire)

2) Il existe un entier h tel que $r_h < a_i^h < r'_h$ ou $r'_h < a_i^h < r_h$ pour une valeur de i au moins.

Pour résoudre ce problème nous devons introduire deux nouvelles définitions qui sont basées sur l'idée d'utiliser l'*information déclarative* contenue dans les questions pour déterminer une partition de la collection :

Définition 6

Considérons l'espace-requête $\mathfrak{U}$ défini par l'interrogation Q sur la collection $\mathcal{F}$. Soit $P(R) = (s_1, \ldots s_h \ldots s_p)$ l'image d'un point R de $\mathcal{F}$. Considérons d'autre part le plus grand et le plus petit éléments de chaque ensemble A_h, et notons-les a_0^h et b_0^h, respectivement. Appelons D_h l'ensemble :

$$D_h = \bigcup_{i=0,\ldots L} \{ a_i^h,\ b_i^h \} \subset A_h$$

Les éléments de cet ensemble sont tous les points extrêmes des intervalles définis par Q, auxquels s'ajoutent le plus petit et le plus grand élément de chaque ensemble ; en vertu des hypothèses faites plus haut, nous pouvons toujours changer le nom des points de D_h pour mettre en évidence le fait que D_h définit une véritable partition :

$$D_h = (c_h^1, c_h^2, \ldots c_h^K)$$

avec

$$c_h^1 = a_0^h < c_h^2 < \ldots < c_h^K = b_0^h$$

Remarque. — Puisque Q est une suite de L questions, D_h contient au plus $(2L + 2)$ points et divise A_h en $(2L + 1)$ intervalles disjoints.

Définition 7

Nous définissons maintenant une fonction α de la façon suivante : si le h-ième attribut de l'objet R est dans l'intervalle $[c_h^\tau, c_h^{\tau+1}[\subset A_h$, alors $\alpha_h = \alpha(s_h) = \tau$

En d'autres termes, nous associons à la valeur du h-ième attribut de R *le nombre entier* τ si cette valeur se trouve dans le τ-ième intervalle de la partition de A_h définie par l'interrogation.

D'autre part nous définissons une fonction $T : \mathfrak{U} \rightarrow N^p$ par la relation : $T(P(R)) = (\alpha(s_1), \alpha(s_2), \ldots \alpha(s_p)) = (\alpha_1, \alpha_2, \ldots \alpha_p)$

Dans ce qui suit, $T(P(R))$ est appelé la *signature* de R.

Nous avons ainsi défini un ensemble de codes (signatures) en utilisant d'abord la fonction P qui projette la collection sur l'espace-requête (définition 6) puis la fonction T qui « numérote » les points de l'espace-requête d'après les intervalles où leurs attributs prennent leurs valeurs.

Proposition 1

La fonction composée $\sigma = T \cdot P$ constitue une solution au problème que nous nous sommes proposé.

La démonstration de cette proposition ne présente pas de grande difficulté. Elle a été donné ailleurs intégralement[4] et permet d'établir que le système de codage ainsi obtenu constitue un système optimal pour l'interrogation donnée.

En résumé, étant donné une liste de questions soumises par un utilisateur, nous partons de ces questions elles-mêmes pour créer un nouveau catalogue dont le système de codage est optimal. Ce nouveau catalogue est considérablement plus petit que la collection de base — d'où une économie considérable de temps-machine. En particulier, il est possible qu'un nombre important d'opérations se ramènent ainsi à des traitements *internes* de l'information, éliminant la lecture de bandes magnétiques ou de disques.

Ces considérations permettent de formuler précisément les propriétés des systèmes qui utilisent comme base les signatures (obtenues par ce procédé) des objets originaux plutôt que la collection elle-même :

Définition 8

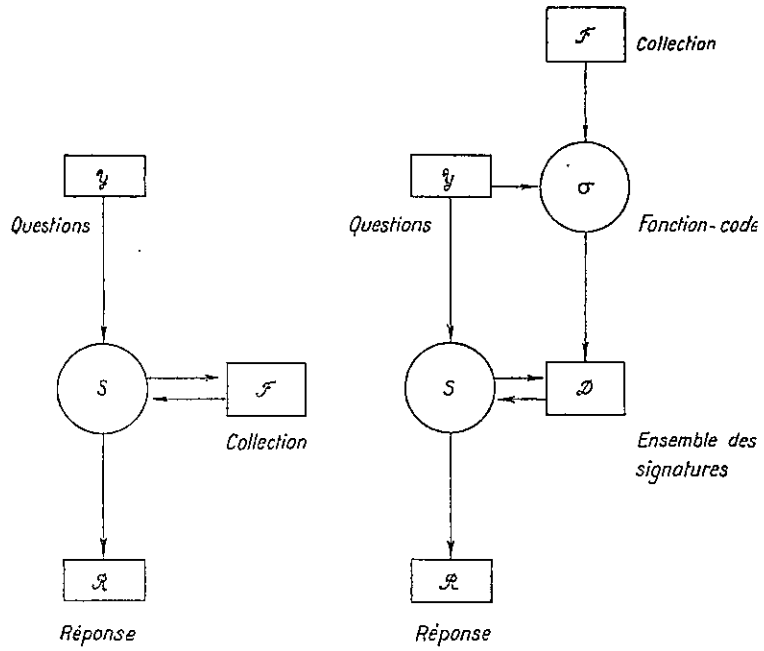
Soit $\mathfrak{R}$ l'ensemble des nombres réels. Une fonction $S : Y \rightarrow \mathfrak{R}$ est appelée un *système direct* sur la collection $\mathfrak{F}$ si, étant donné une question q,

$S(q) = \text{Card}\,(\{R \in \mathfrak{F} \mid a_k < r_k < b_k \text{ pour tout } k \text{ dans } q\})$

Définition 9

Un système direct, couplé à une fonction σ qui obéit les critères du problème défini plus haut, est appelé un *système autocodeur*.

La structure des systèmes directs et autocodeurs a été représentée schématiquement sur la figure 1.



(a) Système direct (b) Système autocodeur

Figure 1

Proposition 2

Étant donné un système autocodeur (S, σ) sur une collection $\mathcal{F}$ et étant donné une question q,

La réponse à cette question est le nombre réel :

$$S(q) = \sum_{i_1 \in I_1} \dots \sum_{i_m \in I_m} \text{Card} \left(\{ R \in \mathcal{F} \mid \sigma(R) = (i_1, i_2 \dots i_m) \} \right)$$

où m représente l'étendue de l'interrogation dont q est un élément, et $I_1, \dots I_m$ les intervalles qui correspondent à q.

Cette proposition, qui a été démontrée ailleurs [4] indique que si les cardinalités de tous les ensembles élémentaires (correspondant à toutes les signatures possibles sous une interrogation donnée) sont connues, la réponse à chaque question sera obtenue par une simple suite d'additions. Un corollaire important est le suivant :

Corollaire

La réponse à chaque question d'une interrogation Q (telle qu'elle a été définie plus haut) peut être obtenue à partir de $(2L + 1)^p$ nombres fondamentaux au plus.

En effet, nous avons observé plus haut (voir notre remarque à la suite de la définition 6) que si Q était une suite de L questions chaque ensemble A_h était divisé en $(2L + 1)$ intervalles au plus. Puisque p est la dimension de l'espace-requête il y a au plus $(2L + 1)^p$ signatures distinctes.

La proposition 2 a deux conséquences importantes :

1) Dans un système autocodeur, toute question admet une représentation pseudo-algébrique sous la forme d'une somme de nombres fondamentaux qui peuvent être calculés une fois pour toutes. Ces *formules d'acquisition* ont une grande importance pour le spécialiste des systèmes car elles offrent un *langage artificiel* optimal pour l'automate chercheur. Nous avons ainsi généralisé la notion de formule d'acquisition sur laquelle la programmation d'ALTAIR était centrée, dans le cas de systèmes dont la dimension est bien plus grande. D'autre part nous avons fait de ce langage artificiel *une variable* du système total *sous contrôle* du programme lui-même : le langage du programme chercheur est totalement manipulé par le moniteur.

2) La propriété énoncée dans le corollaire suggère que le remplacement de systèmes ordinaires, de type « direct », par des systèmes autocodeurs, sera hautement rémunérateur par l'efficacité des opérations ainsi atteinte. Ceci sera valable en particulier dans le cas de collections de très grande dimension mises à la disposition d'équipes d'utilisateurs travaillant dans un contexte très spécifique, et ayant un grand nombre de questions se rapportant à ce contexte — or cette situation est en voie de généralisation rapide dans les problèmes de gestion des entreprises comme dans les problèmes de recherche scientifique proprement dite.

REFERENCES

[1] GREEN, WOLF, CHOMSKY et LAUGHERTY, BASEBALL : An Automatic Question-Answerer, *Proceedings of the Western Joint Computer Conference*, May 1961, 219-224.

[2] VALLÉE, J. F. et HYNEK, J. A., An Automatic Question-Answering System for Stellar Astronomy, *Publications of the Astronomical Society of the Pacific*, vol. 78, n° 463, August 1966.

[3] VALLÉE, J. F., KRULEE, G. K. et GRAU, A. A., *Retrieval Formulae for Inquiry Systems*, inédit.

[4] VALLÉE, J. F., *Search Strategies and Retrieval Languages* (thèse de Doctorat), Northwestern University, 1967.