MATHEMATICA — REVUE D'ANALYSE NUMÉRIQUE
ET DE THÉORIE DE L'APPROXIMATION

**L'ANALYSE NUMÉRIQUE ET LA THÉORIE DE L'APPROXIMATION**
**Tome 6, N° 2, 1977, pp. 177—183**

# SUR LA SYNTAXE D'UN LANGAGE DE SÉLECTION

par

DANIEL PHAM
(Caen)

## 1. Préliminaires

Dans un très grand nombre de travaux effectués sur les fichiers d'une certaine importance on trouve deux grandes catégeories de problèmes : confection des listes et établissement des statistiques.

L'extraction d'une liste à partir des articles d'un fichier donné consiste à sélectionner les articles possédant certaines caractèristiques. Le sous-fichier ainsi formé peut être réorganisé selon ses indicatifs particuliers. Il peut être éclaté, condensé, regroupé avec d'autres sous-fichiers. L'expression „liste" se réfère à l'impression directe ou différée du sousfichier réduit à certains de ses éléments. Par exemple le fichier „CLIENT" d'une banque peut donner lieu, périodiquement, à l'impression des listes de clients ayant effectué des mouvements de fonds. Le fichier „PAIE" d'une entreprise peut fournir, à la fin de chaque mois, en dehors des bulletins de paie, des listes récapitulatives diverses. Le fichier „ETUDIANTS" d'une université peut fournir à la demande, des listes d'étudiants groupés par discipline ou année d'études.

Dans les travaux statistiques consistant la plupart du temps en un rapprochement de certains fichiers, la sélection des articles donne lieu en général à de simples comptages avec sortie des résultats.

La résolution de ces problèmes ne présente aucune difficulté théorique mais l'emploi des langages de gestion comme le COBOL nécessite une description souvent fastidieuse des articles des fichiers. D'autre part, sauf dans le cas d'une gestion intégrée bien conçue, chaque appel d'un sous-programme de liste ou statistique nécessite la lecture complète d'un gros fichier, ce qui contribue à augmenter le temps de traitement.

Nous pensons qu'il est possible de simplifier la résolution de ces problèmes par l'incorporation dans l'OPERATING SYSTEM d'un module

appelé MODULE DE SELECTION placé au même niveau que le MODULE DE GESTION DES FICHIERS.

Le rôle du MODULE DE SELECTION est double :
— il sélectionne les articles d'un fichier d'après les critères introduits par un fichier auxiliaire et procède à des comptages divers au besoin ;
— il facilite la formation de sous-fichiers en vue de la confection des listes diverses.

L'utilisation du MODULE DE SELECTION nécessite, en plus du langage général de traitement de fichiers, langage dont la richesse dépend de la complexité de l' OPERATING-SYSTEM, un mini-langage de SELECTION dont la syntaxe générale fait l'object de la présente communication.

La syntaxe décrite est, en principe, indépendante de la machine emplyoée. Cependant il suffit, dans la pratique, de s'en tenir aux informations stockées dans les fichiers sous l'une des formes suivantes :
— alphabétique et alphanumérique,
— chaîne décimale (décimal externe par module de 6 à 8 bits),
— entier binaire (mot, double mot, demi-mot dans les machines à octets).
— chaîne particulière en décimal-codé-binaire (caractère par module de 4 bits),
— éventuellement chaîne numérique en virgule flottante.

Les chaînes numériques en virgule flottante ne se recontrent pratiquement jamais dans les fichiers de gestion à cause de la diversité de la représentation interne des nombres en virgule flottante. On peut, sans restreindre la généralité, écarter ces chaînes du langage de sélection.

Pour faciliter le langage, nous supposons dans toute la suite, que les informations des fichiers traités sont stockées sous forme de chaînes d'octets (alphanumériques ou numériques). Une simple adaptation permet de passer aux chaînes de caractères de 6 ou 7 bits.

## 2. Critères de sélection et syntaxe

La sélection fait intervenir la représentation des items ainsi que leurs attributs. Les items peuvent être des items isolés (ou simples) ou bien des matrices qu'on traite comme des tableaux à une dimension. Les attributs considérés concernent la classe (numérique ou alphabétique), la nature (décimal, condensé, binaire), la situation et la taille (en octets) etc. ...

On distingue les conditions élémentaires (simples ou multiples) des conditions booléenns constituées par des polynômes booléens dont les variables sont les conditions élémentaires.

### Syntaxe de la representation des items du fichier

a) Tout item du fichier — à l'epception des tableaux ou chaînes complexes-est représenté par le triplet

$$(n1,\ n2\mathscr{X})$$



où la paire de parenthèses caractérise un item de tableau ; $n1$ désigne le numéro de l'octet de gauche de l'item dans l'article du fichier (coopté à partir de 1) ; $n2$ désigne la longeur de l'item comptée en octets ; $\mathscr{X}$ est vide si l'item est alphanumérique ; dans les autres cas $\mathscr{X}$ doit être l'un des caractères suivants :

D si l'item est en décimal externe,
C si l'item est en décimal condensé,
B si l'item est en entier binaire.

Ainsi

(14, 23) désigne un item alphanumérique commencant au 14e octet et de longeur 23 octets.

(15, 2B) désigne un entier binaire occupant 2 octets et commencant au 15e etc. ...

b) Tout tableau ramené à une dimension est représenté par le quintuplet

$$(n1,\ n2\mathscr{X}/n3,\ n4)$$

où le triplet initial désigne le premier élément du tableau ; $n3$ désigne l'incrément (en octets) pour passer d'un élément au suivant, $(n3 \geqslant n2)$ ; $n4$ désigne le nombre d'éléments du tableau.

Le fait que $n3$ ne dépend pas de $n2$ permet de représenter les tableaux ,,à trous" ce qui est commode lorsqu'on ne considère qu'une fraction de chanque élément d'un tableau compact ou lorsqu'on traite une matrice pluridimensionnelle.

La barre oblique ,,/" caractérise un tableau.

### Représentation des constantes

— Une constante alphanumérique est enserrée entre 2 quotes, tout caractère indéterminé est représenté par $\mathscr{C}$. Ainsi

'A B C $\mathscr{C}\mathscr{C}$ X'

représente une chaîne alphanumérique de 6 caractères, les trois premiers ainsi que le dernier sont déterminés, les deux autres peuvent prendre n'importe quelle valeur parmi les codes utilisés.

— Une constante numérique entière est représentée sous la forme habituelle précédée le cas échéant du signe ,,—" s'ill y a lieu (les zéros à gauche ne sont pas significatifs).

— Une liste de constantes comporte des virgules comme séparateurs.

Exemples: '1', '2', '3' représente une liste de 3 constantes alphanumériques, 1, 2, 3 représente une liste de 3 constantes numériques

### Tests et cumuls

On considère les test de classes : information alphabétique ou numérique appliquées respectivement à un item déclaré alphanumérique et décimal externe, et les tests arithmétiques et logiques par l'intermédiaire des opérateurs

$= \neq > <$ : (compris entre)

lorsqu'un test est appliqué à tous les éléments d'un tableau, le nombre d'éléments du tableau satisfaisant au test peut faire l'objet lui-même d'un test de grandeur. Ce dernier test s'appelle une condition de cumul ; le vocable cumul désigne le nombre d'éléments du tableau satisfaisant au premier test, ce nombre est représenté par la lettre S.

## Syntaxe des conditions

Nous désignons par (IT) un item simple, par (IT) un tableau, par (IT/T) une liste de constantes toutes alphanumériques ou toutes numériques pouvant se réduire à un seul élément.

— Tests de classe :

(IT) = A ou (IT/T) = A

(IT) = N ou (IT/T) = N

La condition considérée est satisfaite si l'item (ou *tous* les items quand il s'agit d'un tableau) est (ou sont) de la classe spécifiée par le second membre (A pour alphabétique, N pour numérique).

— Conditions arithmétiques et logiques

(IT) Ω LST ou (IT/T) Ω LST

(Ω désigne l'un des opérateurs $= \neq > <$ :).

La liste LST comporte un ou plusieurs éléments sauf lorsque l'opérateur est „ : " pour lequel LST doit comporter obligatoirement deux éléments, le premier étant *inférieur* au second.

Sauf lorsque l'opérateur vaut „ ≠ ", la condition imposée est satisfaite lorsque *l'un* des items du tableau est lié, par la condition Ω, à *l'un* des éléments de LST.

Lorsque l'opérateur Ω vaut „ ≠ " la condition imposée est satisfaite si et seulement si aucun item du tableau n'est égal à un élément de la liste LST.

Ainsi „ ≠ " est bien la négation de „ = ".

Exemples :

a) (14, 23) = A
b) (1, 13D) = N
c) (388, 1) = ,1', '2', '3'
d) (237, 2B) < 0
e) (237, 2B) : 0,500
f) (239, 5C) < −256
g) (115, 2B) = (387, 5C)
h) (350, 2D/18, 10) = (∗)



i) (350, 2D/18, 10) = (17, 6D/32, 8)
j) (350, 3C/18, 40) = 0, 1S < 4
k) (350, 6/26, 15) = 'AB @@@ 1', 'AB @@@ 2'S > 3

*Remarques* — Dans la condition h) on trouve au second membre la notation (∗) qui représente, *conventionnellement*, le même tableau qu'au premier membre. Cette convention est destinée à éviter une répétition. La condition h) exprime que le tableau figurant au premier membre comporte au moins deux éléments égaux.

— La condition k) exprime que le tableau figurant au premier membre admet plus de 3 éléments alphanumériques (de 6 caractères) commençant par ‚AB' et se terminant par '1' *ou* '2'.

## Libellé des conditions — Conditions simples et multiples

Les conditions décrites dans les paragraphes précédents seront appelées *conditions simples*. Elles seront numérotées continûment de 1 à $p$ pour une sélection effective. Le numéro d'ordre d'une condition simple en constitue son libellé[1].

En vue d'éviter des longueurs on peut introduire des *conditions multiples* qui régissent les items *successifs* d'un tableau.

Une condition multiple résume $k$ conditions simples (sans cumul possible) et obéit à la syntaxe

(IT/T) Ω LST (Ω représente '=' ou '>' ou '<')

ou LST est une liste de k éléments exactement, le tableau devant comporter au moins k éléments ; *la j-ieme condition simple s'obtient en liant le j-ieme élément du tableau au j-ieme élément de la liste*. Quand la liste est numérique on peut employer la notation :

élément initial/incrément, nombre d'éléments.

Ainsi 1/2, 10 représente la liste des 10 premiers nombres impairs.

Pour distinguer une condition simple d'une condition multiple on emploie, pour désigner le libellé de cette dernière la notation $q$A$r$ (par exemple 6A15) où $q$ et $r$ sont deux entiers ($q < r$) ; la condition multiple équivaut donc aux conditions simples de libellé $q$, $q + 1$, ..., $r$.

## Conditions Booléennes

Toute combinaison des conditions décrites ci-dessus à l'aide des opérateurs 'ET' 'OU', 'NON' est appelée condition booléene. Le libellé d'une condition booléenne est indiqué par la lettre B suivie de son numéro

[1] Chaque condition se présente sous forme d'une carte ou d'une image carte (avec possibilité de carte 'suite') ; l'emplacement du 'libellé' évite toute confusion avec les items du fichier ou les constantes.

d'ordre. On emploie les signes. + — pour les opérateurs 'ET' 'OU' 'NON'. Par exemple dans

B5 1. (—2). (3 + 4)

on a quatre conditions simples de libellés 1, 2, 3, 4 et la condition booléenne est la cinquième imposée à la sélection.

On considère aussi une condition booléenne multiple résumant plusieurs conditions booléennes ordinaires ; le libellé d'une telle condition s'écrit B*q*A*r* et équivant à B*q*, B*q* + 1, ..., B*r*. Dans une telle condition on doit trouver une liste numérique de $r - q + 1$ nombres entiers déisgnant $r - q + 1$ conditions simples par exemple

B12 A21 1.(2/1, 10)

désigne 10 conditions booléennes de libellés B 12 à B 21 représentant 1·2, 1·3, ..., 1·11

*Remarques.* La sélection d'un article de fichier se fait, dans la pratique de la gestion, à l'aide d'un petit nombre de conditions en général ne faisant intervenir que les informations alphanumériques et des nombres entiers sous diverses représentations.

— La syntaxe décrite dans les paragraphes précédents n'est pas directement applicable au cas du rapprochement de deux ou plusieurs fichiers ; ainsi une condition telle que

(ITEM FICHIER 1) = (ITEM FICHIER 2)

exige une identification de chacun des fichiers.

Il n'est pas difficile de compléter la syntaxe précédente. Pour prendre en compte ce cas, il suffirait de faire suivre l'item de fichier (IT/T) par un numéro qui désigne le fichier considéré.

Dans les applications statistiques habituelles il est aisé de ramener le cas général au cas particulier de traiement d'un seul fichier à la fois.

### 3. Application à la construction d'un module de sélection

La partie centrale d'un module de sélection est composée d'un analyseur syntaxique du langage de sélection assisté d'une pile de stockage de conditions et d'une pile de compteurs statistiques.

La numérotation séquentielle continue des conditions facilite et le stockage et la recherche des résultats.

Deux autres sous-modules traitent respectivement, le premier de la réservation de mémoire pour le stockage des listes et de la pile de compteurs pour statistiques, le second de l'impression, directe ou différée, des résultats.

En vue de l'organisation de listes et statistiques de sortie (titres, disposition de lignes etc. ...) le dernier sous-module peut introduire, à l'instar du REPORT-WRITER du langage COBOL, unde syntaxe simplifiée des lignes d'impression, syntaxe sur laquelle il n'est pas nécessaire d'insister dans cet article.



En gros, le fonctionnement du module de sélection, placé sous le contrôle du MONITEUR, peut être envisagé de la facon suivante :

— A l'entrée, un fichier CARTE (ou IMAGE-CARTE)) contient les informations suivantes :

Nombre de listes, nombre de statistiques à confectionner ;

Leurs caractéristiques : titres, disposition etc. ...

Les critères de sélection, par liste et par statistique, critères regroupés et numérotés séquentiellement.

— Le MODULE procède alors à la réservation de mémoire pour les sous-fichiers listes et statistiques (sous forme d'un fichier disque unique organisé en RANDOM permettant le classement des listes au fur et à mesure de leur formation).

— Le fichier central à traiter est ensuite lu, article par article. Après la lecture de chaque article, le contrôle est renvoyé au MODULE qui analyse les conditions, délivre, le cas échéant, des messages d'erreur, stocke les résultats dans les piles, confectionne les articles des sous-fichiers de listes, les écrit aux places voulues, procède au comptage et garnit les piles statistiques.

— Dès que la lecture du fichier central est terminée, le sous-module d'impression entre en jeu et fait sortir, dans l'ordre, les listes une par une, puis les statistiques. De la sorte, avec une seule lecture du fihicer central, on peut traiter un nombre pratiquement illimité de listes et de statistiques. Seule la capacité des mémoires de stockage (mémoire centrale et disque) constitue une limite au volume des sous-fichiers résultats.

Nous avons appliqué les idées précédentes à l'écriture d'un MODULE DE SELECTION sur une IRIS 50 de la C.I.I. de 94 K fonctionnant en monoprogrammation.

Compte-tenu des dimensions modestes des mémoires auxiliaires utilisées (disques du type 2311 de 6M octets) le nombre de listes à confectionner, en une seule fois a été arbitrairement limité à 200 et le nombre de statistiques simultanées à élaborer, à 10.

Des tests concluants ont été faits sur un fichier d'étudiantes en vue de l'établissement simultané des listes d'inscription à une centaine d'UNITES DE VALEURS dans le cadre d'une faculté ainsi que des statistiques de ces mêmes étuidiants ventilés par sexe, par valeurs déjà obtenues etc.... D'autres tests sont en cours concernant l'utilisation d'un tel module à la gestion administrative d'une commune (listes électrorales et statistiques diverses) ainsi qu'à la gestion de la paye du personnel d'une entreprise de taille moyenne.

Reçu le 8 XI. 1973.