REVUE FRANÇAISE D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE. SÉRIE ROUGE

GUY LOUIS-GAVET

Brève communication. Étude mathématique pour la concentration de fichiers occupant un volume important

Revue française d'informatique et de recherche opérationnelle. Série rouge, tome 5, n° R3 (1971), p. 101-111

<http://www.numdam.org/item?id=M2AN_1971__5_3_101_0>

R.I.R.O.
(5e année, N° R-3, 1971, p. 101-111)

ETUDE MATHEMATIQUE POUR LA CONCENTRATION DE FICHIERS OCCUPANT UN VOLUME IMPORTANT

par Guy LOUIS-GAVET (1)

Sommaire. — *Un aspect important des travaux faits sur les Bibliothèques, outre les problèmes spécifiques à la gestion des fichiers (notamment les chaînages), a été de réduire des fichiers trop importants, ces derniers après réduction devant rester l'image exacte des fichiers d'entrée.*

En fait le problème était de respecter, quel que soit le nombre de rubriques en général, une biunivocité *entre les rubriques d'entrée et celles de sortie. Ce fut un des buts de notre recherche.*

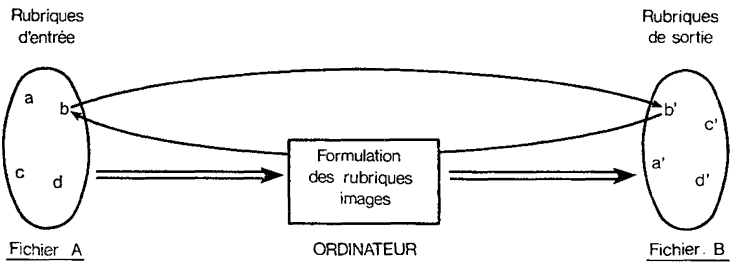

Figure 1

Il va sans dire que la résolution de tels problèmes dépassent largement le cadre de gestion d'une Bibliothèque et nous servira notamment dans les recherches sur la documentation automatique.

Nous allons développer sommairement les problèmes rencontrés à la Bibliothèque qui nous ont fait déboucher sur cette recherche.

(1) I.U.T., Informatique Lyon, boulevard du 11 Novembre, 69-Villeurbanne.

VOLUME DU FICHIER DE BASE DANS LES BIBLIOTHEQUES

Dès qu'une Bibliothèque a un fonds d'ouvrage important, elle se trouve obligée, pour avoir une bonne gestion de ce fonds (investigation, réception, prêt des livres), de faire appel à l'informatique. Aussi devra-t-elle disposer d'un fichier de référence où se trouveront tous les livres que la Bibliothèque possède avec leurs caractéristiques. C'est en quelque sorte le fichier matière des bibliothécaires, avec, en plus bien d'autres renseignements tels que : date d'arrivée, prix...

Il va sans dire que ce Fichier de BASE aura un volume très important de l'ordre de plusieurs centaines de millions de caractères (et même du milliard) pour une Bibliothèque assez importante (on compte entre 800 et 1 000 caractères par ouvrage).

Ainsi il est pratiquement impossible de se servir de tels fichiers lorsque l'on veut une réponse rapide pour un renseignement déterminé, à moins de les mettre sur des mémoires de masse, solution bien trop onéreuse.

Aussi a-t-on eu l'idée de faire un fichier beaucoup plus petit, de l'ordre du vingtième, qui soit l'image du fichier de base dans ses caractéristiques majeures et qu'ainsi, on puisse l'interroger facilement. A savoir pour un livre : son titre, son auteur, son édition, avec d'autres informations importantes. En fait on a eu à prendre l'empreinte de ces caractéristiques d'où le nom de ce fichier : Fichier des EMPREINTES. Il nous servira pour toute la gestion des livres en Bibliothèque, mais surtout lors de l'acquisition des livres; le fichier de BASE étant alors sur bandes.

APPROCHE GENERALE DU PROBLEME

Pour partir sur des bases solides, nous avons dû analyser dans le détail, statistiquement parlant, comment étaient formés les Titres-Auteurs-Édition d'un livre. Partant de ces statistiques d'où nous avons obtenu certaines courbes (fréquence d'un caractère, d'un couples de caractères...) nous avons dénoté certaines répartitions à éviter dans la structure de l'empreinte, pour qu'en fait nous aboutissions à un maximum de possibilités d'empreintes. Obtenant ainsi une probabilité la plus faible possible d'avoir une homonymie entre deux empreintes provenant de deux livres différents, en fait obtenir une biunivocité parfaite entre tous les Titres-Auteur-Édition d'un livre et leur empreinte.

C'est au niveau de l'algorithme, qui formulait l'empreinte, que nous avons eu à éviter dans celle-ci, les défauts de répartition des caractères dans le titre-auteur-édition. Ce qui sous-entend que nous avons eu à développer des statistiques différentes au niveau de l'empreinte (complémentaires de celles faites au niveau des Titre-Auteur-Édition du livre, ces dernières ayant surtout un

but informationnel). Notamment au niveau global de l'empreinte, car en fait nous pouvons considérer que l'ensemble Titre-Auteur-Édition d'un livre est une approximation d'un système ergodique, mais en aucun cas nous devons avoir une telle approximation au niveau de l'empreinte.

Prenons un exemple simple. Nous avons noté que la lettre E apparaissait dans une proportion de 22 % en deuxième position dans le titre. Nous avons fait en sorte que ce « pic » s'amenuise en améliorant peu à peu l'algorithme au vu des théories statistiques appliquées sur le Titre-Auteur-Édition. Dans un 1er temps cela nous a suffi. Mais très rapidement nous nous sommes aperçus que nous ne maîtrisions plus d'autres phénomènes qui se développaient par ailleurs au niveau de l'empreinte (catégories de caractères revenant trop souvent, mauvaises corrélations entre différents groupes de caractères...). Aussi comme l'empreinte est de longueur fixe, nous avons été amené à penser à une théorie statistique qui lie la place du caractère au caractère lui-même, nous permettant ainsi de résorber ces différents phénomènes.

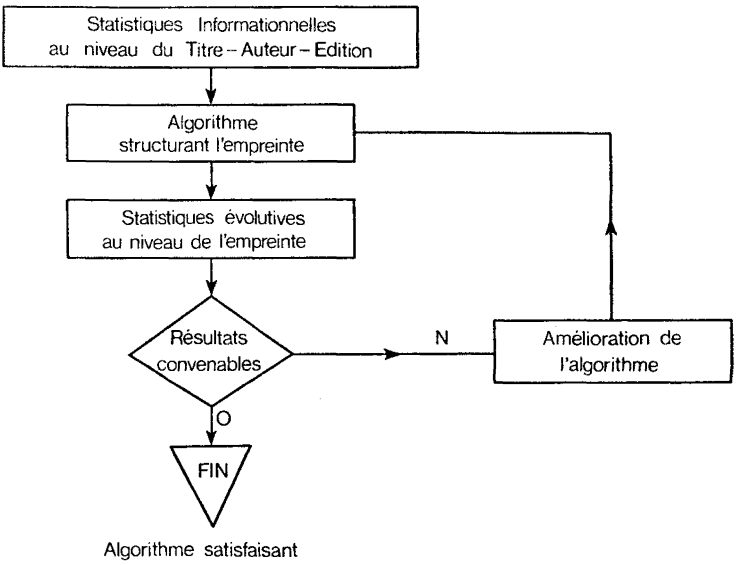

Figure 2

REMARQUE : nous ne développerons pas ici le mode de fonctionnement de l'algorithme, ce dernier devant être relativement simple, afin que l'ordinateur, comme nous sommes en mode conversationnel, ne passe pas un temps trop long pour construire l'empreinte, (car nous avons à ajouter aussi le temps de recherche dans le fichier). Ceci est une restriction importante.

ETUDE MATHEMATIQUE

Introduction

Dans un but d'uniformisation, un ensemble de définitions a été constitué suivant ce tableau en notation de Backus.

⟨ Monogramme ⟩ ::= ⟨ caractère ⟩

⟨ Caractère ⟩ ::= ⟨ lettre ⟩|⟨ espace ⟩|⟨ délémiteur ⟩|⟨ caractère rare ⟩

⟨ lettre ⟩ ::= A | B | C | ..... | Z

⟨ espace ⟩ ::= |—|

⟨ délémiteur ⟩ ::= . Point|entre Titre et Auteur

/ Barre oblique| entre les différentes parties de l'édition

* Astérisque| entre Titre-Auteur et Édition

⟨ Caractère rare ⟩ ::= ⟨ Chiffre ⟩|⟨ Signe orthographique ⟩|

⟨ Signe de ponctuation ⟩|⟨ Caractère spécial ⟩

⟨ Chiffre ⟩ ::= 0 | 1 | 2 | ..... | 9

⟨ Signe orthographique ⟩ ::= ' Apostrophe|

\- Trait d'union|

⟨ Signe de ponctuation ⟩ ::= , | ; | : | ! | ? | « | » | (|)

⟨ Caractère spécial ⟩ ::= + | % | =

⟨ Bigramme ⟩ ::= ⟨ monogramme ⟩ ⟨ monogramme ⟩

⟨ Trigramme ⟩ ::= ⟨ monogramme ⟩ ⟨ monogramme ⟩

⟨ monogramme ⟩

Ce tableau sera complété par ces définitions :

⟨ Identificateur ⟩ ::= Titre-Auteur-Édition d'un livre

⟨ Cardinal ⟩ ::= nombre d'éléments d'un ensemble

⟨ Redondance ⟩ ::= Identité de monogramme, bigramme, ...

Pour notre problème nous avons été amené à penser à une structure statistique (Ω, a, p).

Ω Ensemble des caractères identifiant 3 600 livres

a Ensemble des parties de Ω

p Ensemble des lois de probabilités

Choix de la loi de probabilité

A priori nous ne savions quelle loi faire coïncider avec notre problème. Cependant nous sommes partis du fait que si tous les caractères d'un identificateur étaient indépendants nous pouvions penser à une loi binomiale, or le principal inconvénient est que l'identificateur d'un livre est une approximation d'un système ergodique.

Dans un tel système les « messages » qui se suivent ont une probabilité d'apparaître suivant un certain ordre; mais s'il y a une corrélation entre les messages elle ne s'étend pas au-delà d'une certaine limite.

Compte tenu de ce fait nous avons fait plusieurs approximations successives en considérant 1 caractère, 2 caractères ... x caractères comme groupe d'éléments complètement indépendant du caractère suivant.

Les résultats le prouvent, le test d'adéquation avec la loi binomiale s'applique d'une manière d'autant plus satisfaisante que l'on considère des groupes d'éléments de plus en plus importants.

Au départ nous considérons tous les monogrammes comme indépendants entre eux, comme nous venons de le voir cette approximation est obligée d'être très grossière. Cependant pour cette première approximation certains caractères suivent une distribution binomiale (ceux qui reviennent très rarement dans un identificateur).

Ensuite nous considérons que les bigrammes sont indépendants puis nous appliquons de nouveau le test d'adéquation sur ceux-ci. De proche en proche nous recommençons ces études sur des groupes de monogramme de plus en plus importants. Ce qui nous permet de cerner, avec une certaine approximation, « l'ergocité » de l'ensemble des identificateurs.

Ainsi d'une façon générale nous pourrons savoir à partir de quelle valeur de x, x-uples de monogrammes suivent la loi Binomiale. Bien entendu cette hypothèse ne sera valable qu'avec un seuil de signification α, de la loi d'ajustement du Chi-2, que l'on aura choisi *a priori.*

Calcul des fréquences empiriques et théoriques, Ajustement avec la loi Binomiale

Considérons X identificateurs de taille n, dans lesquels nous enregistrons le nombre de redondances de chaque monogramme. Après les avoir recherchés dans ces X identificateurs nous pouvons établir une distribution en classe.

Avec a_j^{mono} le nombre de redondances du monogramme mono dans la classe j, nous en tirons la fréquence empirique pour un monogramme donné :

$$f_{\text{mono}} = \frac{\sum_{j=0}^{n} a_j^{\text{mono}} * j}{x * n}$$

Ensuite nous calculons la fréquence théorique par la loi Binomiale, puis nous faisons le test d'ajustement du CHI-2, pour un seuil de signification α donné.

$$X^2_{\text{mono}} = \sum_{j=0}^{n} \frac{(a_j^{\text{mono}} - Np_j)}{Np_j} \qquad \text{(voir tableau p. 110)}$$

Estimateur de p dans la loi Binomiale

Si nous appliquons la loi Binomiale encore faut-il être sûr que la fréquence empirique d'un monogramme, d'un bigramme... est un bon estimateur de la probabilité obtenue avec la loi binomiale, aussi allons-nous rechercher son estimation ponctuelle d'une part et son estimation ensembliste d'autre part.

A) Estimation ponctuelle en considérant la présence ou l'abscence d'un caractère donné. Envisageons la structure statistique $(\Omega, a, p)^n$ avec n nombre d'observations indépendantes.

Ω pouvant prendre les valeurs $\{ 0 \}$ ou $\{ 1 \}$ suivant la présence ou l'absence du caractère.

Par définition on a $a = p(\Omega)$. Soit X la variable aléatoire qui indique la présence du caractère, nous aurons donc :

$$P(X = 1) = P \qquad \text{et} \qquad P(X = 0) = 1 - P = q$$

Si on dénote k apparitions du caractère donné, on aura comme fonction de vraisemblance

$$L(X, P) = \sum_{j=1}^{n} P(X = x_j) = P^k q^{n-k} \qquad \text{avec} \qquad k = \sum_{j=1}^{n} x_j$$

ainsi :

$R = \sum_{j=1}^{n} X_j$ est une statistique exhaustive de X, par conséquent cela nous amène à la structure statistique suivante :

$$\Omega = \{ 0, 1, 2, ; ....., n \}$$
$$a = p(\Omega)$$

Nous pouvons dire que R suit la loi binomiale $B(n, p, q)$. Prenons la fréquence $f = \frac{k}{n}$ comme estimateur de p. Démontrons que cet estimateur convient.

Calculons $E(F)$ avec $F = \frac{K}{n}$

$E(F) = \frac{1}{n} E(K) = \frac{1}{n} np = p$ (car $E(K) = np$) Comme il n'y a pas de facteur correctif cet estimateur est sans biais.

Calculons sa *variance minimale :*

$$\sigma(F) = \sigma\left(\frac{K}{n}\right) = \frac{1}{n^2}\,\sigma(K) \qquad \text{or} \qquad \sigma(K) = npq$$

$$\sigma(F) = \sigma\,\frac{p(1-P)}{n} \qquad n \to +\infty \quad \text{Cet estimateur est convergent.}$$

Calculons le *maximum de vraisemblance.*

Dérivons la fonction de vraisemblance par rapport à p.

$$\frac{L}{P} = kp^{k-1}(1-p)^{n-k} - p^k(n-k)(1-p)^{n-k-1}$$

$$= p^k(1-p)^{n-k}\left(\frac{k-np}{p(1-p)}\right)$$

ainsi le maximum a lieu pour $p = \dfrac{k}{n}$ valeur de la fréquence.

Cet ensemble de faits prouve que *la fréquence empirique est un excellent estimateur de p.*

B) Estimation ensembliste.

En ayant comme hypothèses, que K suit une loi binomiale et $f = \dfrac{k}{n}$, nous allons essayer de trouver le plus grand intervalle possible, telle que la probabilité de contenir p soit supérieur à un nombre donné $1 - \alpha$. Lorsque n est assez grand, suivant les hypothèses de départ, X suivra la loi Normale centrée réduite, pourvu que $npq > 20$.

$$X = \frac{(f-p)\sqrt{n}}{\sqrt{pq}}$$

On sait que si $N(X) = \alpha$ on a $X = N^{-1}(\alpha)$. En prenant un espace symétrique on a cette double inégalité (la médiane jouant le rôle du second quartile $\alpha = 1/2$).

$$(N^{-1}(\alpha/2)\sqrt{pq})/\sqrt{n} \leqslant f - p \leqslant -(N^{-1}(\alpha/2)\sqrt{pq})/\sqrt{n}$$

en prenant $\beta = N^{-1}(\alpha/2)/\sqrt{n}$ on obtient :

$$\beta\sqrt{pq} \leqslant f - p \leqslant -\beta\sqrt{pq}$$

en remplaçant q par $1 - p$ on arrive à l'inéquation suivante :

$$p(1+\beta^2) - p(2f+\beta^2) + f^2 \leqslant 0$$

p devra donc être compris entre les deux racines de cette inéquation. Développant au premier ordre : l'expression (I).

$$\frac{-(2f+\beta^2)+\sqrt{(2f+\beta^2)^2-4f^2(1+\beta^2)}}{2(1+f^2)} \equiv \frac{-(2f+\beta^2)+\sqrt{4(f-f^2)+\beta^2}}{2(1+f^2)} = (I)$$

$$I \simeq (2-2\beta^2+0(\beta))+(\beta^2+2f+\beta\sqrt{4(f-f^2)}+0(\beta^2))$$

On obtient finalement :

$$f-\sqrt{(f-f^2)} \leqslant p \leqslant f+\sqrt{(f-f^2)}$$

Ainsi

$$\boxed{p = f \pm N^{-1}(\alpha/2) * (f-f^2) * 1/\sqrt{n}}$$

En prenant $N^{-1}(\alpha/2) = 1{,}88$, la conclusion importante qu'il nous faut tirer de cette formule est que la précision sur p sera d'autant meilleure que la taille de l'échantillon sera grande.

De plus, nous pouvons conclure que f sera un excellent estimateur de p, puisque pour notre problème la précision sera de l'ordre de 15^{-4};

Intervalle entre deux redondances d'un même monogramme

Une redondance d'un monogramme étant apparue, qu'elle est le nombre de monogrammes qui va le séparer de sa prochaine redondance? La réponse à cette question est primordiale pour la formulation de notre algorithme, car ainsi, après nos statistiques informationnelles, nous pouvons mettre un « poids » à chaque monogramme, de telle façon que dans l'empreinte, il y ait une meilleure répartition de ceux-ci. De la même manière nous pouvons le faire pour x-uples de monogrammes.

Soit X la variable aléatoire indiquant le rang n de sortie d'un monogramme. Cette dernière suit la loi géométrique.

Nous avons donc : $P(X = n) = pq^{n-1}$

dont le moment d'ordre 1 est $E(X) = \sum_{n=1}^{\infty} npq^{n-1}$

$$E(X) = p\sum_{n=1}^{\infty} nq^{n-1} = \frac{p}{(1-q)^2} = \boxed{\frac{1}{p} = E(X)}$$ (voir tableau p. 110)

Ainsi l'intervalle entre deux redondances d'un même monogramme est *l'inverse de sa probabilité.* Cela ne peut indiquer qu'une moyenne, ne serait-ce le fait qu'il existe des bigrammes de même monogramme.

Conclusion

Nous n'avons développer que les statistiques que nous avons appelé « informationnelles », ces dernières donnant des renseignements primordiaux sur les indentificateurs afin de structurer les empreintes d'une manière satisfaisante. Il nous restera à développer une théorie statistique qui lie, au niveau de l'empreinte, un monogramme donné à sa place dans celle-ci. Nous permettant ainsi de maîtriser certains défauts dans la structure générale de l'empreinte, que les statistiques informationnelles appliquées à l'empreinte ne nous permettaient pas de contrôler totalement.

De plus, nous n'avons pas cité d'autres statistiques importantes faites sur l'ensemble des identificateurs (longueur des mots, place d'un monogramme dans un mot, probabilités conditionnelles...) ces dernières ne demandant pas un développement théorique (voir ci-après un des tableaux de ces statistiques).

NOMBRE TOTAL DE MOTS 08008

LONGUEUR DES MOTS	NOMBRE	POURCENTAGE
01	00096	01,19
02	01545	19,29
03	00714	08,91
04	00697	08,70
05	00924	11,53
06	01151	14,37
07	01068	13,33
08	00703	08,77
09	00414	05,16
10	00242	03,02
11	00178	02,22
12	00073	00,91
13	00035	00,43
14	00015	00,18
15	00014	00,17
16	00018	00,22
17	00001	00,01
18	00000	00,00

Figure 3

Répartition des mots dans un identificateur

FREQUENCE ET INTERVALLE D'APPARITION D'UN MONOGRAMME DONNE

MONOGRAMMES	REDONDANCES	FREQUENCE CORRIGEE	INTERVALLE D'APPARITIONS
A	04065	08,89	000011
B	00703	01,53	000065
C	01002	02,19	000045
D	01582	03,46	000028
E	06573	14,38	000006
F	00413	00,90	000111
G	00729	01,59	000062
H	00871	01,90	000052
I	02814	06,16	000016
J	00166	00,36	000277
K	00194	00,42	000238
L	03179	06,95	000014
M	01377	03,01	000033
N	02596	05,68	000017
O	02465	05,39	000018
P	00785	01,71	000058
Q	00128	00,28	000357
R	03269	07,15	000013
S	02990	06,54	000015
T	02325	05,08	000019
U	02098	04,59	000021
V	00653	01,42	000070
W	00109	00,23	000434
X	00183	00,40	000250
Y	00339	00,74	000135
Z	00215	00,47	000212
	04165	09,11	000010
)	00064	00,14	000714
'	00299	00,65	000153
.	01620	03,54	000028
[	00002	00,00	000000
,	00331	00,72	000138
;	00001	00,00	000000
(	00088	00,19	000526
&	00005	00,01	000000
-	00119	00,26	000384
%	00003	00,00	000000
=	00002	00,00	000000
:	00006	00,01	000000
/	00001	00,00	000000
0	00602	01,31	000076
1	00061	00,13	000769
2	00036	00,07	001428
3	00016	00,03	003333
4	00020	00,04	002500
5	00018	00,03	003333
6	00006	00,01	000000
7	00013	00,02	005000
8	00040	00,08	001250
9	00014	00,03	003333

Figure 4

Fréquence et intervalle d'apparition d'un monogramme donné

BIBLIOGRAPHIE

G. CALOT, *Statistique descriptive*, Dunod, 1964.
J. R. BARRA et A. BAILLE, *Problèmes de statistique mathématique*, Dunod, 1969.
R. BOREL, *Statistique et analyse linguistique*, PUF, 1968.