REVUE FRANÇAISE D'AUTOMATIQUE, INFORMATIQUE, RECHERCHE OPÉRATIONNELLE. MATHÉMATIQUE

GUY LOUIS-GAVET

Brève communication. Étude mathématique pour la concentration de fichiers occupant un volume important (2e partie)

Revue française d'automatique, informatique, recherche opérationnelle. Mathématique, tome 6, n° 1 (1972), p. 71-80.

<http://www.numdam.org/item?id=M2AN_1972__6_1_71_0>

R.A.I.R.O.
(6e année, R-1, 1972, p. 71-80)

ETUDE MATHEMATIQUE POUR LA CONCENTRATION DE FICHIERS OCCUPANT UN VOLUME IMPORTANT

(2e partie)

par Guy LOUIS-GAVET (1)

Résumé. — *Ceci est la suite de l'article paru dans la revue R3-1971 de l'AFCET Rouge, qui traitait de la 1re partie d'une théorie mathématique étudiant la concentration de fichier occupant un volume important. Notre recherche fut de trouver une théorie mathématique qui puisse : d'une part, nous faire découvrir un algorithme capable de réduire des rubriques d'un fichier (entraînant une biunivocité parfaite entre les rubriques d'entrées et celles de sortie) ; d'autre part, maîtriser des phénomènes de redondances au niveau des rubriques de sortie.*

Afin de trouver quelle forme prendraient ces algorithmes, il nous a fallu développer des statistiques au niveau des identificateurs : statistiques appelées informationnelles. Elles firent l'objet du premier article.

Introduction

Il est évident que si ces statistiques nous ont permis de structurer notre algorithme, notamment en cernant l'ergodicité d'un identificateur, nous avons dû par ailleurs développer d'autres statistiques au niveau de l'empreinte, nous permettant ainsi d'améliorer notre algorithme sur les identificateurs. Pour notre application, ce sera l'ensemble des titres-Auteur.

Ceci est résumé dans le schéma page suivante.

Structuration de l'empreinte

Nous avons montré que le test d'adéquation de la loi binomiale était d'autant meilleure que nous nous adressions à des groupes d'éléments plus importants. Qu'est-ce que cela voulait dire? Que si nous cernions l'ergodicité d'un identificateur, nous pourrions savoir à partir de quelle valeur de x,

(1) I.U.T. Informatique Lyon, Villeurbanne.

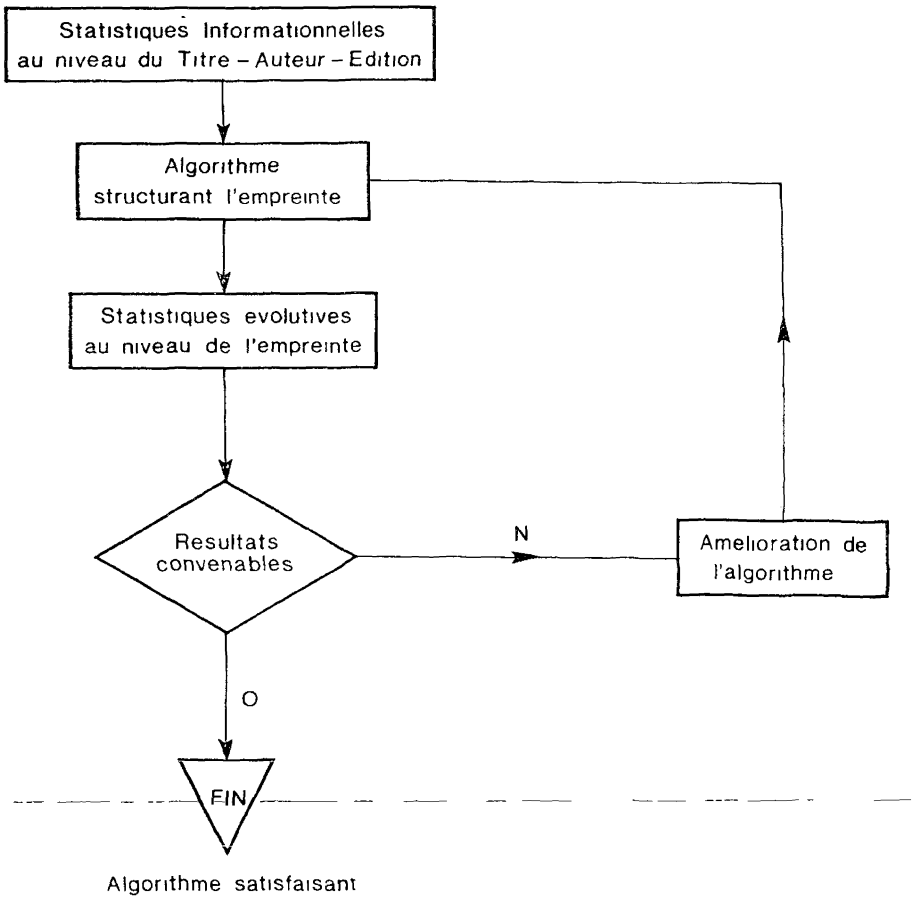

Figure 1

x-uple éléments avait une totale indépendance entre ce dernier et l'élément suivant. Pour ce faire nous avons calculé les probabilités conditionnelles d'un groupe d'éléments qui suit ou qui précède un élément donné. Nous les avons obtenu très simplement par l'intermédiaire des probabilités des bigrammes, des trigrammes...

Or à de très rares exceptions, les résultats le montrent, l'ergodicité d'un identificateur s'arrête à $x = 3$. Ainsi, pour des groupes de 3 éléments les probabilités conditionnelles sont négligeables vis-à-vis de l'élément qui le suit ou qui le précède.

En partant de ces statistiques conditionnelles, une structuration très simple consiste à prendre un caractère à intervalle régulier, le pas de l'intervalle devant

être au minimum de 3. Nous n'insisterons pas sur l'élaboration de notre algorithme, nous dirons simplement qu'il faut faire attention de :

— ne pas prendre un intervalle trop petit vis-à-vis de la longueur de l'identificateur, car nous risquons de saturer l'empreinte avec un jeu de caractères ne correspondant pas à toutes les caractéristiques de l'identificateur;

— ne pas prendre un intervalle trop grand, car nous n'aurions alors que des espaces à la fin de l'empreinte, augmentant la probabilité d'homonymie.

Aussi pour éviter tous ces écueils, avons-nous modulé nos intervalles suivant la longueur de l'identificateur, de telle façon que nous soyons sûr de parcourir entièrement toutes ces caractéristiques.

Calcul de la probabilité d'homonymie de deux empreintes

Cette technique d'empreinte très simple répond aux normes que nous nous étions fixées. Mais il conviendrait de montrer que cet algorithme engendre une probabilité d'homonymie très faible.

Développement théorique

Soient :

— une chaîne d'entrée (appelée Titre-Auteur) σ :

$$\sigma = \{ \sigma_1 \sigma_2 \ldots \sigma_i \ldots \sigma_p \}$$

avec $\sigma_i \in \{$ lettres, chiffres, espaces, caractères spéciaux $\}$

— une chaîne de sortie appelée empreinte ν.

$$\nu = \{ \nu_1 \ldots \nu_j \ldots \nu_q \} \text{ avec } \nu_j \in \{ \alpha_n \} \text{ et } q < p$$

Les σ_i se distribuent suivant les distributions des caractères dans le Titre-Auteur. L'algorithme peut, dans une première approche, consister à prélever q caractères (nombre fini) parmi les p caractères donnés. Considérons dans une première approche que tous les Titres-Auteurs ont la même longueur (ceci ne change rien à l'étude du problème, sinon que cela diminue d'autant l'ensemble de définition que nous pouvons subdiviser en différentes classes de longueur égale).

Donc le $i^{ème}$ élément de σ va se trouver en $j^{ème}$ position dans ν, ou bien être abandonné. Soient I l'ensemble des q valeurs de i retenues et J l'ensemble des $j^{èmes}$ positions correspondantes. Cette correspondance correspond à l'algorithme.

Ainsi la probabilité d'avoir deux Titres-Auteur ayant une empreinte identique est donc d'avoir q éléments identiques distribués suivant les rangs de i.

Soit P_1 cette probabilité :

$$P_1 = \begin{pmatrix}\text{Probabilité d'avoir} \\ \sigma_{i1} = \sigma'_{i1}\end{pmatrix} * \begin{pmatrix}\text{Probabilité d'avoir} \\ \sigma_{i2} = \sigma'_{i2}\end{pmatrix} * * \begin{pmatrix}\text{Probabilité d'avoir} \\ \sigma_{iq} = \sigma'_{iq}\end{pmatrix}$$

Nous voyons que nous avons intérêt à ne prendre que des caractères indépendants, c'est-à-dire avec un intervalle de 4 caractères, comme nous venons de le dire dans les pages précédentes. Sinon la probabilité conditionnelle à l'apparition d'un caractère est à prendre en considération et elle diminue la probabilité P_1.

Nous voyons aussi que nous avons intérêt à avoir des Titres-Auteur le plus long possible.

Nombre théorique d'empreintes engendré par cet algorithme. Liaison d'un monogramme avec sa position dans l'empreinte

Ces statistiques que nous venons de développer nous disent a priori le risque d'homonymie de deux empreintes, mais ne nous éclaire pas sur les défauts que pourraient avoir l'ensemble des empreintes dans leur structure. Comme nous avons une longueur d'empreinte fixe, nous pourrions avoir des phénomènes au niveau de la répétition de mêmes caractères dans de mêmes positions de l'empreinte. Défauts qui proviendraient de la structure des Titres-Auteurs. Dans ces conditons, prendre une chaîne de 18 caractères pour l'empreinte, ne serait-ce pas superflu ou trop peu?

Ainsi c'est au niveau de l'algorithme, qui formule l'empreinte que nous avons eu à éviter dans celle-ci, les défauts de répartition des caractères dans le Titre-Auteur. 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 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 suffit. 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 (catégorie 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, un monogramme à sa place dans l'empreinte, nous permettant ainsi de résorber ces différents phénomènes, et de dévoiler certains résultats qui peuvent être trompeurs comme l'exemple ci-dessous.

EXEMPLE

Il est évident, que si l'algorithme est bien formulé le test d'adéquation de la loi binomiale doit pouvoir s'appliquer parfaitement. Nous avons montré que l'intervalle de redondance d'un même monogramme dans un identificateur est l'inverse de sa fréquence. Elle est variable et c'est normal. Imaginons que pour les empreintes cet intervalle soit devenu uniforme pour tous les monogrammes et même pour certains bigrammes. Cela pourrait paraître très satisfaisant a priori, mais cela pourrait dire aussi que l'ensemble des empreintes aurait cette forme suivante :

$$A_1 \quad A_2 \quad A_3 \ldots A_{12} \quad A_{13} \ldots A_{17} \quad A_{18}$$
$$A_1 \quad A_2 \quad A_5 \ldots A_{12} \quad A_{16} \ldots A_{14} \quad A_{18}$$

Ce qui serait désastreux, cela voudrait dire, avec une probabilité élevée que certaines positions de l'empreinte auraient toujours le même monogramme en l'occurence dans notre exemple la 1re, 2e, 12e et 18e positions.

Développement théorique

Considérons les ensembles E et F.

— E représentant les positions dans l'empreinte e { 1, 2, 3, ..., 18 }

— F représentant les différents monogrammes f { A, B, ..., Z, *, ;, ... }.

Si nous avons un bon algorithme, la position d'un monogramme dans la chaîne des caractères d'une empreinte est le résultat, par cet algorithme, d'un choix indifférent entre *tous* les caractères. Ainsi pour un élément donné de E nous avons le choix entre f_i possibilités.

De la même façon, si nous considérons un élément quelconque f_i, sa position e_j dans l'empreinte devra être indifférente.

Nous pouvons transposer ceci sous une forme matricielle, la matrice P ayant tous ses éléments $p(i, j)$ égaux.

(i représentant un monogramme donné et j sa position dans l'empreinte.)

Matrice P

Nous pouvons dire aussi que nous avons défini une application E dans F, or ces deux *ensembles sont finis.*

Dans ces conditions nous aboutissons à un arbre des exponentielles ou arbre des applications où les points terminaux sont bien de même génération

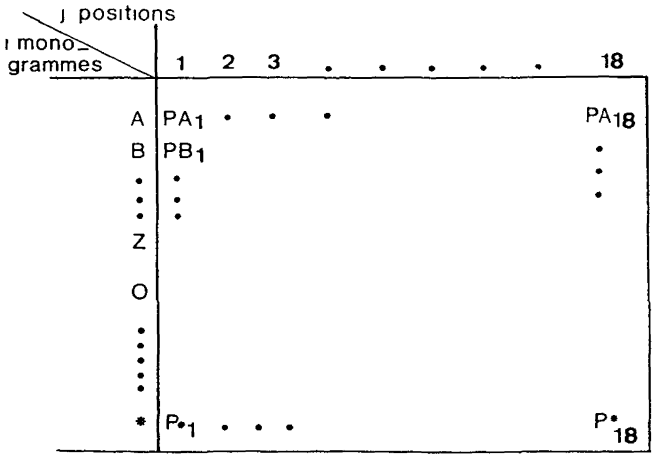

Figure 2

(ici faisant partie d'un même ensemble) et qu'en outre de chaque point peut partir le même nombre de ramification (c'est-à-dire le nombre total de caractères).

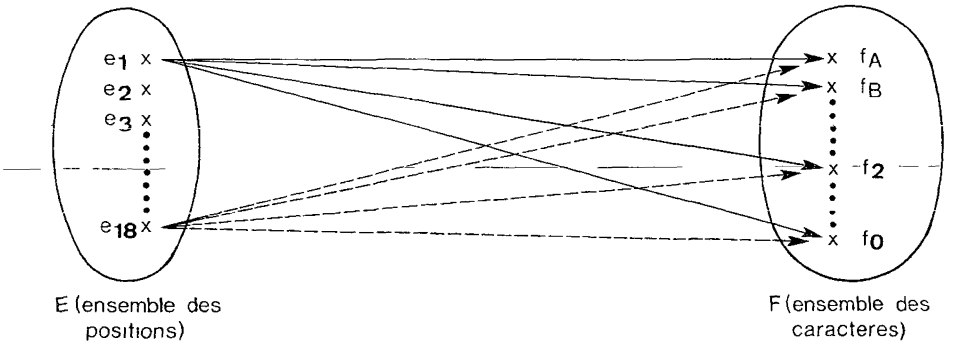

Figure 3

Ainsi cet arbre représentant un cas particulier des arbres exponentielles, il peut être associé aux applications de E dans F.

Si nous appelons card $(E) = m$ et card $(F) = n$, nous connaissons la propriété suivante : si K est l'ensemble des applications de E dans F.

$$\text{Card } (K) = n^m = \text{nombre de combinaisons possibles.}$$

Nous pouvons dire alors que si à partir d'un ensemble d'identificateurs composé d'un nombre c de catégories de monogrammes différents, nous vou-

lons réduire tous les identificateurs sur p éléments, c^p est la borne supérieure du nombre d'empreintes possibles. Chiffre idéal qu'aucun algorithme ne peut atteindre.

Exemples simples

Imaginons que l'on construise une empreinte de 3 positions avec un ensemble d'identificateurs construits à l'aide de seulement deux monogrammes différents. Nous aurons :

$$2^3 = 8 \text{ possibilités d'empreintes}$$

AAA, AAB, ABA, ABB, BBB, BAB, BBA, BAA.

Si on la construit à partir de deux positions avec un ensemble d'identificateurs comportant 3 monogrammes différents, nous aurons :

$$3^2 = 9 \text{ possibilités d'empreintes.}$$

AA, AB, AC, BA, BB, BC, CA, CB, CC.

Ces deux exemples auront pour matrices finales :

M_1

	1	2	3
A	4	4	4
B	4	4	4

M_2

	1	2
A	3	3
B	3	3
C	3	3

En généralisant : la valeur de P_{ij} sera égale à nombre de combinaisons/ nombre de monogrammes

$$P_{ij} = \frac{C^p}{C}$$

Ce qui vérifie la théorie développée précédemment. Ainsi en cette matrice nous avons obtenu un outil de contrôle très puissant, qui nous permettra de savoir à chaque instant suivant le nombre de rubriques d'entrées, comment évolue tous les caractères dans chaque position de l'empreinte. Et d'une façon

générale quelle est l'incidence de notre algorithme sur un ensemble d'empreintes. Étudier ainsi comment évoluent certaines fréquences de monogrammes, noter les amplifications, faire des statistiques sur leur nombre d'apparitions. Et par là-même, essayer d'améliorer l'algorithme afin que nous arrivions à une matrice idéale. Mais il faut bien penser que ceci est une vérification a posteriori du calcul des probabilités que nous avons effectué théoriquement en nous basant sur la forme générale de notre algorithme. Par conséquent sur ce dernier rien ne sera fondamentalement changé, sinon la longueur de certains pas ou l'inclusion et d'autres caractéristiques du Titre-Auteur; comme nous le verrons par la suite.

Variation de la matrice P

Les propos précités se résument à ceci : comment notre matrice P va évoluer. En fait tous les P_{ij} seront le reflet exact de la répartition de nos monogrammes au niveau global de la matrice entière. En fait nous avons fait un transfert d'une application sur un outil-mathématique. Nous avons à étudier celui-ci pour connaître si l'application est bonne, et si elle est, essayer de l'améliorer.

Il est évident que tous les éléments P_{ij} de la matrice ne seront pas égaux lorsque nous la consulterons pour un échantillon donné d'empreintes. Il faudrait pour cela que notre algorithme fut parfait, nous pourrions alors construire environ 30^{18} empreintes ($3^{18} * 10^9$ milliards!) chiffre astronomique. Mais notre but était de savoir comment, pour un monogramme donné, quelle loi suivait sa fréquence d'apparition dans chaque position de l'empreinte. Avons-nous une loi qui va se stabiliser, s'infirmer, pour un nombre d'empreintes allant croissant?

Loi équiprobable

Ici nous devons théoriquement aboutir à une loi équiprobable, c'est-à-dire que toutes les fréquences d'un monogramme donné doivent se situer sur une droite de la forme générale $y = Ax + B$ où A serait très proche de 0 et B de la fréquence moyenne.

Aussi sommes-nous ramené à un ajustement d'une courbe à des données. Ces droites de régression sont calculées par la méthode des moindres carrés, cette dernière nous a semblé être la plus précise pour notre problème.

L'application de la méthode des moindres carrés conduit pour la détermination des constances a et b aux formules suivantes :

$$\text{(I)} \qquad a = \frac{n\Sigma x_i y_i - \Sigma x_i * y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} \qquad b = \frac{\Sigma x_i^2 * \Sigma y_i - \Sigma x_i * \Sigma x_i y_i}{n\Sigma x_i^2 - (\Sigma x_i^2)}$$

Or dans notre cas, nous pouvons aboutir à des formules beaucoup plus

simples, car ici les x_i sont régulièrement espacés : aussi peut-on faire la transposition suivante :

$$X_i = x_i - \bar{x} \qquad \text{et} \qquad Y_i = y_i - \bar{y}$$

d'où nous obtenons en partant de (I)

$$A = \frac{X_i Y_i}{X_i^2} \qquad \text{et} \qquad B = \frac{Y^i}{n} = \bar{y}$$

En définitive, en revenant aux axes (o_x, o_y) la droite ajustée aux données de notre problème aura la forme générale

$$\boxed{y = ax + (\bar{y} - a\bar{x})} \qquad \text{avec} \qquad \underline{A = a}$$

Comme les abscisses étaient représentées par des points fixes et arbitrairement mis. (Ils représentent les positions dans l'empreinte.) Nous ne pouvons pas nous fier à une seule droite de régression, en fait à une seule valeur du *coefficient angulaire a.* Aussi avons-nous étudié la *variation de a* en prenant plusieurs échantillons de l'ensemble des empreintes.

Si nous avons à faire à une loi équiprobable les valeurs de *a* doivent tendre vers 0, si les échantillons deviennent de plus en plus importants.

Ceci corroboré par le fait que si nous prenons l'ensemble des fréquences d'un monogramme donné, nous devons obtenir des courbes dont les amplitudes s'effacent peu à peu.

Nous ne parlerons pas des améliorations apportées à notre algorithme, notamment la notion de poids que nous avons apporté à certains monogrammes, soient qu'ils fussent trop rares ou trop fréquents. Cependant quelles que soient les améliorations nous n'arriverons jamais à trouver la même fréquence pour tous les monogrammes. Nous arrivons à des classes de monogrammes qui découpent dans notre matrice P, des sous-matrices P_1, P_2 ... où les éléments sont égaux.

Conclusion

Nous avons trouvé là un outil mathématique qui nous permette sans aucune inquiétude de réduire des fichiers. Nous voyons tout l'intérêt que cela peut avoir. Ainsi pour les Bibliothècaires qui doivent avoir une réponse rapide pour un renseignement déterminé (nombre de livres possédés, lieu où ils se trouvent, leur cote...) cela est primordial. Autrement le coût pour stocker leur fichier serait trop important, ce qui empêcherait l'expansion de la lecture publique à Lyon.

Cependant nous avons essayé d'aller plus loin. Nous nous sommes posés

la question de savoir si notre fichier de pure gestion, ne pourrait pas nous servir pour faire de la recherche documentaire. Il nous fallait pour cela que notre relation biunivoque entre un ouvrage et son empreinte devienne bijective. Sans donner de détails, disons simplement qu'il ne faut pas oublier que nous avons dans nos rubriques d'entrées un système ergodique. Qu'est-ce que cela sous-entendait? Si nous avions *n*-uples éléments indépendants, ces derniers ne peuvent être construits que suivant un type de structure logique qui doit au-delà d'un certain nombre, se retrouver identique. Nos statistiques ont vérifié ceci, ce qui nous permet de créer un fichier d'un volume très faible qui représente le complément de l'ensemble des empreintes.

BIBLIOGRAPHIE

A. CULLMAN, *Éléments de calcul informationnel*, 1960.

BONDY, *Éléments de phonétique*, Cahier Baillere, 1968.

J. L. SAVAGE, *A note on the Evaluation of methods for systematically abbreviating english words*, 1970.