Modèle mathématique de circulation des documents : distribution d'Usage, d'Utilité et de Contenu du Document: l'exemple de la circulation des articles scientifiques des périodiques.

Thierry Lafouge

Thierry Lafouge
Laboratoire RECODOC
Université Lyon1
43 Boulevard du 11 novembre 1918 Villeurbanne cedex.
Lafouge@enssib.fr

Modèle mathématique de circulation des documents : distribution d'Usage, d'Utilité et de Contenu du Document: l'exemple de la circulation des articles scientifiques des périodiques.

Résumé

Cet article fait une synthèse sur nos travaux publiés en matière de circulation d'articles chez un fournisseur de documents. Cet article explicite le modèle mathématique et rappelle les principaux résultats obtenus. Des nouveaux complètent les précédents et ouvrent une perspective plus large de notre modèle tout en l'ancrant dans des résultats anciens concernant les travaux bibliométriques sur des sujets concernant les modèles mathématiques de circulation.

1) Introduction

L'analyse des flux informationnels documentaires s'intéresse à quantifier, prévoir, la production et l'usage de l'information qui sont deux processus de même nature. Nous étudions ici les flux d'articles produits et utilisés. Par production nous entendons ici les articles publiés dans les revues scientifiques. L'usage quant à lui est mesuré par les demandes de ces derniers (commandes chez des fournisseurs, prêt inter entre bibliothèques, télé déchargement d'un article sur un serveur….).
Nos études antérieures ont privilégié un axe concernant la modélisation de ces phénomènes (LAFOUGE 1998). Les techniques utilisées ici sont celles de la bibliométrie distributionnelle. Cette dernière s'intéresse entre autre à quantifier certains processus liés à l'usage et à la production en Science de l'Information. On observe alors des fréquences d'évènements appelés généralement distributions bibliométriques qu'on représente avec des lois probabilistes.
Nous définissons et construisons dans cet article deux types de distribution, la distribution d'Usage et la distribution de Contenu pour analyser ces flux. Nous rappelons et complétons ensuite notre modèle distributionnelle qui va pouvoir rendre compte des liens entre distribution de Contenu et d'Usage.

2) Distribution d'Usage du Document

Historiquement les distributions relatives aux usages des documents ont été observées dans les bibliothèques (MORSE 1968). Etant donné une collection d'ouvrages, on s'intéresse durant une période de temps fixée (un an, un mois..) au nombre de prêts de chaque document de ce corpus. Parmi les techniques quantitatives utilisées, on peut citer par exemple la théorie des files d'attentes et les processus markoviens (EGGHE 1990), issues des techniques utilisées en recherche opérationnelle. Ces méthodes peuvent être utilisées pour aider à résoudre des problèmes de gestion de stock en bibliothèque. On peut citer, les problèmes liés au " fonds mort " des bibliothèques gérant des documents anciens ou au contraire aux manuels très demandés que les bibliothèques universitaires doivent acquérir en plusieurs exemplaires.

Dans les applications décrites précédemment, les calculs en général de nature probabiliste sont faits sur des objets physiques. Ors l'hypothèse avancée par tout le monde aujourd'hui face au document numérique, c'est que l'usager recherchera à travers les réseaux l'information dont il éprouvera le besoin. Il est raisonnable de penser que la circulation physique du document va diminuer et être remplacée par la circulation ou la communication de l'information. Le document "ouvrage" est alors un objet composé avec des objets porteurs . Nous ne définirons pas plus loin ici la notion de document qui est délicate et pose de nombreuses questions qui sont hors de propos dans cet article (BUCKLAND 1998).

Pour notre part nous nous sommes intéressés aux commandes d'articles chez un fournisseur de documents qui est l'Inist[1] (SALAUN 1999). La question qui se pose alors est la suivante: si l'on veut appliquer les modèles bibliothéconomiques à la prévision de la demande, quelle est l'entité documentaire pertinente? Le titre du périodique, le volume concerné ou l'article lui même. Ce type d'étude est importante pour la gestion d'un fournisseur de documents qui doit organiser et structurer ses données pour répondre à la demande sur un réseau électronique.. Dans le futur la notion de volume, n'étant plus lié au support physique de l'édition traditionnelle, va se modifier avec l'apparition des revues électroniques.
Formalisons cet usage Etant donné un corpus de périodiques scientifiques nous observons durant une période fixée (un mois, un an….) la demande des articles publiés dans ce corpus. Plus précisément nous observons combien de demandes -ces demandes dans nos études précédentes(LAFOUGE 1995, SALAUN 1999) sont matérialisées par des commandes de photocopies d'articles- ont fait l'objet chaque article du corpus. Nous représentons ce flux par une distribution statistique des fréquences mise sus la forme classique $\{i \ f(i)\} i = 1...i\max$, qui signifie qu'un pourcentage de $f(i)$ documents (périodiques, volumes, articles…) ont été demandés *i* fois, sachant qu'un document peut être demandé au maximum imax fois. Dans le cas de fournitures d'articles, la fameuse règle des 80/20 observée en économie (80% des commandes ne concernent que 20% des titres) est constatée régulièrement dans ce genre de distribution.
Pour collecter les données relatives à ces distributions il est nécessaire de bien préciser ce que l'on observe. Un périodique est composé de différents volumes, eux mêmes composés d'articles, aussi la circulation peut être observée à trois niveaux différents:

- du périodique,
- du volume,
- de l'article.

Lorsqu'un périodique ou un volume est demandé 10 fois il est possible que ce soit le même article qui soit concerné. Aussi le problème du gestionnaire d'un centre documentaire qui doit constituer son fonds est délicat: un périodique beaucoup demandé peut l'être pour deux raisons différentes, très peu d'articles par volume sont demandés, mais ils le sont très souvent où inversement, presque tous les articles d'un volume sont demandés mais chacun une seule fois. Dans une étude ancienne (LAFOUGE 1989) nous avions montré que les revues scientifiques ont des comportements différents: en ce qui concerne les revues appliquées ou technologiques, il semble que certains articles traitant des problèmes plus aigus ou plus pratiques que d'autres, focalisent l'intérêt, d'où une demande plus concentrée.
Inversement, pour des périodiques théoriques ou fondamentaux, les articles sont tous importants et l'intérêt des demandeurs se répartit. Aussi pour décrire ce phénomène il est nécessaire de prendre en compte le nombre d'articles par volume.

[1] Institut national de l'information scientifique et technique

3) Distribution de Contenu du Document

Nous rappelons ici ce que nous avons introduit dans les articles précédents (Lafouge 1999) la distribution de Contenu ou de structure qui quantifie le nombre d'articles par volume. Comme précédemment nous représentons cette distribution statistique sous la forme: $\{i\ g(i)\}i=1....p\max$ avec les mêmes conventions que précédemment, pmax étant le nombre maximum d'articles par volume. Si l'on regarde les indicateurs de bibliométrie concernant les citations, le calcul du fameux "Impact-factor" des revues (produit par l' ISI) prend en compte non seulement le nombre de citations pendant une année qu'a reçu un périodique (un usage ici est une citation) mais aussi le nombre d'articles publiés pendant les deux années précédentes. Comme pour l'usage d'un périodique où c'est toujours le même article qui est demandé, ce dernier peut avoir un " impact factor " très fort (GARDFIELD 1997) qui est dû au même article qui est toujours cité.

4) Modèle distributionnelle

Le modèle probabiliste que nous avons construit et qui va être présenté ici va relier ces deux distributions, Usage et Contenu. Soit un corpus de périodiques scientifiques, nous modélisons deux situations bien distinctes: dans un premier temps nous nous intéressons au nombre d'articles demandés dans chaque volume (nous notons fu la distribution correspondante).

$fu(2)$ = 50% signifie que la moitié des volumes du corpus de périodiques ont eu deux articles demandés au moins une fois (dans ce cas imax est le nombre maximum d'articles que peut avoir un volume : imax = pmax).

Dans un deuxième temps nous nous intéressons au nombre de commandes d'articles par volume (nous notons fc la distribution correspondante).

$fc(2)$ = 50% signifie que la moitié des volumes du corpus de périodiques ont eu deux commandes. (dans ce cas imax est le nombre maximum de commandes d'articles que peut avoir un volume : imax n'est pas borné et connu d'avance).

Ces deux distributions sont distinctes et n'ont pas la même signification. Seule $fu(0)\ et\ fc(0)$ les proportions de volumes jamais utilisées ou demandés sont les mêmes. D'un point de vue pratique cette donnée est souvent inconnue ou du moins difficile à connaître. Nous allons définir maintenant un modèle distributionnelle qui prenne en compte ces deux situations, représentées par ces deux distributions.

Nous appellerons dans la suite, ces deux dernières, respectivement **distribution d'Utilité** fu **et distribution d'Usage** fc Nous avons montré en étudiant les commandes d'articles d'un corpus de revues en chimie que ces distributions ont des caractéristiques similaires: décroissantes avec une variance très supérieure à la moyenne, ce qui signifie qu'on a une dispersion très forte et en général une longue queue. Si ce type de propriété était prévisible pour l'usage, il en était pas de même pour l'utilité. Nous allons donner (LAFOUGE 1995) à titre d'exemple les commandes durant l'année 1985 à l'Inist de 22 titres en chimie classés dans le domaine " Matériaux de Construction " en donnant les distributions d'Usage et d'Utilité au niveau du volume et de l'article (cf. Tableau 1)

Taleau 1 Commandes d'articles en 1985 pour la collection " Matériaux de Construction " de l'Inist "

Commandes	*Articles :* $A(i)$ %	Volumes $fc(i)$ USAGE %	Volume $fu(i)$ UTILITE %
1	887 95.7	498 75.8	513 78.3
2	38 4.1	97 14.8	91 13.9
3	2 0.2	31 4.7	29 4.4
4		12 1.8	7 1.1
5		8 1.2	6 0.9
6		4 0.6	5 0.8
7		2 0.3	1 0.2
8		2 0.3	2 0.3
9		2 0.3	1 0.2
10		0 0	0 0
>=11		1 0.2	1 0.2
	$\sum iA(i)$ = 969	$\sum ifc(i) =$ 969	$\sum ifu(i)$ = 926
	$\sum A(i) =$ = 927	$\sum fc(i)$ = 657	$\sum fu(i)$ = 656

La lecture de la ligne 3 du tableau nous permet de dire qu'environ 4% des articles ont fait l'objet de deux commandes, 15% des volumes ont été utilisées deux fois, alors que 14% des volumes ont eu deux articles demandés au moins deux fois.
Les deux dernières lignes du tableau expriment sous forme fréquentielle l'usage et l'utilité.
Soit de façon globale 969 commandes ont eu lieu; ces 969 commandes ont concerné 926 articles répartis dans 656 volumes. Nous ignorons ici la proportion d'articles jamais demandée et le nombre d'articles produits par les volumes de ces 22 revues. (Les différences 656/657, 926/927, sont dues aux troncatures des distributions).

Définitions et hypothèses du modèle

a) Paramètres du modèle

Soit un corpus de périodiques, composés de volumes, contenant des articles, nous notons:

. *pu* la probabilité qu'un article appartenant à un volume quelconque d'un périodique soit commandé au moins une fois.

. $po = 1 - pu$: appelé "No use" . Cette quantité est souvent très importante et voisine de 1. Le poids des "No Use" est très important et va conditionner de nombreux modèles mathématiques de circulation . En général cette valeur est inconnue, elle nécessite un corpus bien délimité. Si on collecte les commandes d'articles chez un fournisseur on pourra par exemple s'intéresser uniquement à la proportion d'articles (respectivement volumes) jamais demandée appartenant à des volumes (respectivement périodiques) demandés au moins une fois.

Supposons que le nombre d'articles d'un volume soit fixé:
on désigne par $Bc(j\ pu)(i)$ la probabilité qu'un article appartenant à un volume ayant *j* articles soit commandé *i* fois: *i* = 0, 1…….imax (nombre maximum de commandes).

On suppose que cette probabilité dépend uniquement du nombre *j* d'articles, et de la probabilité donné à priori qu'un article soit commandé.
De même soit $Bu(j\,pu)(i)$ la probabilité que *i* articles appartenant à un volume ayant *j* articles soit demandé au moins une fois : $i = 0,1 \ldots j$. Cette probabilité dépend des mêmes paramètres que précédemment. Elle mesure la proportion d'articles utilisés par volume de périodique du corpus.
D'après la remarque précédente on doit avoir pour tout *j* : $Bu(j\,pu)(0) = Bc(j\,pu)(0)$.
On suppose que l'on a quantifié le nombre d'articles dans des périodiques par la distribution précédente de Contenu que l'on va formaliser également par une distribution de probabilité .
On désignera par $G(j)$ la probabilité qu'un volume de périodique du corpus ait *j* articles.

b) Distribution d'Usage et d'Utilité des volumes en fonction de la distribution de Contenu

On va maintenant définir la distribution de probabilité qu'un volume soit demandé *i* fois qu'on notera $Pc(i)$ en fonction des distributions G et Bc et la distribution de probabilité qu'un volume ait *i* articles commandés au moins une fois qu'on notera Pu (i), en fonction de G et Bu .
Pour ce faire on utilise l'axiome d'additivité des probabilités :

$$Pc(0) = \sum_{j=1}^{\infty} Bc(j, pu)(0).G(j) \;\; ; \; Pc(i) = \sum_{j=1}^{\infty} Bc(j, pu)(i).G(j) \;\; i = 1,2.... \text{ [A]}$$

$$Pu(0) = \sum_{j=1}^{\infty} Bu(j, pu)(0).G(j) \;\; ; \; Pu(i) = \sum_{j=i}^{\infty} Bu(j, pu)(i).G(j) \;\; i = 1,2.... \text{[B]}$$

Pc et Pu sont respectivement les distributions d'Usage d'Utilité définies précédemment.

Nous pouvons déjà faire deux remarques qui vont conditionner les choix qui vont suivre.

Dans les deux cas la proportion des volumes jamais demandée étant la même, le choix des lois doit respecter la condition : $Pc(0) = Pu(0)$

Si tous les volumes ont un article (volume et article sont la même entité) et un seul la distribution des usages des volumes doit être la même que celle des articles, ce qui induit :
$Pu(0) = po$; proportion d'articles ou de volumes jamais demandés.
$Pu(1) = pu$; proportion d'articles ou de volumes demandés au moins une fois.
Dans ce cas la distribution des commandes des volumes est égale à celle des articles et on a:
$Pc(i) =$ probabilité qu'un article soit demandé i fois i = 0,1,2….

b) Choix des lois
Nous allons choisir deux lois de probabilité usuelles pour modéliser l'utilisation (*Bu*) et l'usage (*Bc*) dans le cas où le nombre d'articles est fixé Pour l'utilité on suppose qu'elle se fait suivant la " loi du hasard ", d'autres hypothèses seraient difficilement envisageables ici puisque cette distribution de probabilité est finie et dépend d'un seul paramètre. Aussi si un volume a *j* articles, nous avons choisi la distribution binomiale classique de paramètre *j* et *pu.*

$$Bu(j\ pu)(i) = \frac{(j-i+1).....(j-1).j}{i!}\ pu^i.po^{j-i}\ \ i = 0,1....j$$

Pour les usages, on s'appuie sur les expérimentations faites depuis longtemps en bibliothèque qui utilisent souvent une distribution géométrique pour modéliser la circulation des documents (BURREL 1986). Nous faisons également cette hypothèse. Nous allons de plus supposer que si plusieurs articles sont dans un même volume, ceci n'a aucune influence sur la distribution de la circulation de chaque article . En fait nous supposons qu'il y a indépendance, ce que les probabilités traduisent pour un volume de *j* articles, par la convolution de *j* distributions géométriques. Un résultat classique de probabilité (CALOT 1988) nous montre alors que $Bc(j\ pu)$ est une distribution binomiale négative de paramètre j et pu :

$$Bc(j\ pu)(o) = po^j\ \ Bc(j\ pu)(i) = \frac{j(j+1)...(j+i-1)}{i!}\ po^j.pu^i\ ;\ i = 1,2...\ ....$$

Cette loi est souvent observée dans le cas de la circulation des ouvrages dans une bibliothèque (LEEMAN 1992).

5) Résultats: relations entre Contenu, Usage et Utilité

A l'aide des modèles ([A] et [B]) et des choix précédents nous allons essayer de montrer les relations qui existent entre la distribution de Contenu et la distribution d'Usage ou d'Utilité.

La problématique est alors la suivante:

on observe très souvent des distributions d'Usage qui s'ajustent suivant des lois simples de type Poisson, Géométrique, Binomiale négative. La question que nous posons est alors la suivante : les propriétés que l'on observe sur la forme de la distribution d'Usage ou d'Utilité (Poisson, Géométrique, Binomiale négative), ne sont elles pas une conséquence des propriétés de la distribution de Contenu?

Tableau 2 : Principaux résultats du modèle

Distribution : *Bu, Bc*	*G:* distribution de contenu	*Pu: Pc*: Distibution des volumes	N°
Binomiale : UTILITE	Géométrique	Géométrique d'espérance M	[1]
	Binomiale négative	Binomiale négative d'espérance M	[2]
	Poisson	Poisson d'espérance M	[3]
Neg. Binomiale: USAGE	Géometrique	Géomérique	[4]
	Binomiale négative	Binomiale négative d'espérance M	[5]
	Poisson	Poisson d'espéranceM	[6]

En d'autres termes si on fait l'hypothèse que la distribution de contenu *G* est de type Géométrique, Binomiale négative ou Poisson, est ce que la distribution d'Utilité ou d'Usage définies par les équations [A] ou [B] est du même type? Le tableau 1 ci dessus résume les résultats (numérotés de [1] à [6]) que l'on a obtenu (LAFOUGE 1999). Nous avons testé trois hypothèses relatives au Contenu pour l'Utilité et l'Usage. A part le résultat [4] qui est direct et s'obtient en faisant un simple calcul formel en utilisant les définitions précédentes (On montre

dans ce cas que Pu est une distribution géométrique d'espérance $E(G).(1-po).\frac{1}{po}$). Les autres résultats s'obtiennent en passant à la limite. Rappelons les trois conditions qui conditionnent le passage à la limite et qui définisse M, espérance de la distribution d'Usage et d'Utilité.

On suppose:

(a) $po \to 1$

(b) $E(G)\,(E(G)\,désigne\,l'espérance\ de\,la\,distributin\ G) \to \infty$

(c) $(1-po).E(G) \to M$ (L'espérance de *Pu* et *Pc* est donc définie par une limite)

La condition (a) traduit le fait que la proportion des " No Use" tends vers sa valeur maximum c'est à dire que la probabilité qu'un article soit utilisé devient infime. La condition (b) nous dit que le nombre d'article par volume est de plus en plus grand. Cette condition nécessite une définition du volume qui n'est pas habituel. Avant d'interpréter (c) nous allons expliciter un résultat:

Par exemple le résultat [1] signifie que sous les conditions limites (a) (b) (c) on a :

$$Limite\ Pu(i) = Lim \sum_{j=1}^{\infty} Bu(j, pu)(i)q(1-q)^{j-1} ... = m(1-m)^i \quad i = 0,1.2..$$

$Bu(j\ pu)$:*distribution binomiale de paramètre pu et j.*

Où les paramètres des lois géométriques q et m des distributions d'Usage et de Contenu sont liés aux espérances par les relations:

$$q = \frac{1}{E(G)} \ ; \ m = \frac{1}{M}$$

Après passage à la limite, *Pu* est une distribution géométrique de moyenne *M*.

D'autre part Il est aisé de montrer que les distributions Pc et Pu ont pour espérance les valeurs suivantes (LAFOUGE 1995) :

$$E(Pu) = (1-po)E(G)$$

$$E(Pc) = (1-po)E(G).\frac{1}{po}$$

Ce résultat nécessite que la distribution *G* ait au moins un moment d'ordre un, ce qui est vérifié ici pour tous les types de distribution de Contenu que nous avons choisi. On voit que lorsqu'on suppose les conditions limites (a) et (b) la condition (c) signifie que l'espérance de la distribution d'Usage et d'Utilité est la même, et tends vers une limite finie. Les résultats du tableau traduisent deux phénomènes qui nous semblent importants:

- c'est la nature de la distribution de Contenu qui est transmise à la distribution d'Usage.[2]
- la distribution d'Usage et d'Utilité sont identiques ([2] et [5] [3] et [6]).

On remarquera que pour les mêmes conditions limites les résultats [1] et [4] sont identiques.

[2] Nous voulons dire par la(cf [6]), que bien que Bc (distribution d'usage pour un nombre d'articles fixés) soit binomial négatif, si G la distribution de Contenu est Poisson, alors la distribution d'Usage Pc est Poisson.

6) Interprétation des conditions limites et de la distribution de contenus

Nous allons expliciter ces conditions par un exemple afin de montrer qu'elles sont réalistes : supposons qu'un fournisseur de documents à l'instant t possède une collection complète de périodiques sur les dix dernières années écoulées; dans notre modèle ce que nous appelons volume dans ce cas est l'ensemble des volumes (numéros ou fascicules) d'un périodique parus ces dix dernières années. Il est clair que le nombre moyen d'articles par volume augmente au cours du temps puisque le nombre de périodique reste constant (ou du moins varie peu). Nous allons formaliser cet exemple.

Ici le volume est le périodique :
soit $X(t)$ le nombre d'articles produits par cette collection de N périodiques sur la période $(0,t)$; supposons que $C(dt)$ articles ont été commandés au moins une fois sur la période $(t, t+dt)$.

Lorsque t devient de plus en grand avec dt restant constant les conditions (a) (b) et (c) s'écrivent :

(*condition (a)*): $\frac{X(t)-C(dt)}{X(t)}$: la proportion des articles du fonds jamais utilisés devient de plus en grande au cours du temps.

(*condition (b)*) : $\frac{X(t)}{N}$:le nombre moyen d'article publiés par un périodique est de plus en plus grand, ceci traduit simplement le fait trivial que la taille de la collection augmente avec le temps.

(condition (c)) : $\frac{C(dt)}{N}$: le nombre moyen d'articles utilisés durant la période dt par périodique devient(ou reste) stable. On est dans une situation stationnaire de type classique dans un processus.
En résumé ces conditions sont vérifiées si on suppose :
La limite de X(t) lorsque t tends vers l'infini est infini alors que celle de C(dt) est finie.
(Cela peut se traduire par le fait que le « volume du fonds augmente plus vite que son utilisation »)

<u>Analogie de notre modèle avec un processus stochastique</u>

La modélisation de la circulation dans notre modèle utilise des distributions stationnaires (cf. les équations [A] et [B]). Nous voulons dire par là que le facteur temps qui est primordial dans ce type d'étude sur la circulation n'est pas un paramètre explicite des équations. On peut le faire intervenir de façon implicite lors du passage à la limite comme nous venons de voir.

Les résultats obtenus avec notre modèle (cf. Tableau 2) sont de même nature avec des hypothèses différentes: passer à la limite ici est un artifice pour tenir compte du facteur temporel. Rappelons le modèle de circulation défini entre autre par Burrel (Burrel 1987)

X désigne le nombre de documents demandés pendant l'intervalle de temps $[0\ ,\ t]$.
Si on suppose que le processus est connu pour un paramètre *k* (nombre positif) supposé varier pour chaque document suivant une distribution définie par une densité f_k (appelée "loi de désirabilité"), et si on note $P(X=i)$ la probabilité qu'un document soit demandé i fois , la distribution d'Usage s'écrit :

$[\mathrm{x}]$: $P(X=i)=\int_0^{\infty} P(X=i/k=x).f_k(x)dx$ (ici x varie de façon continue de o à ∞)

pour i = 0,1,2…….
On montre alors que si on suppose que le processus conditionnel est de type Poisson :

$$P(X=i/k=x)=\frac{e^{-h(t).x}}{i!}.(h(t).x)^i$$

(dans le cas le plus simple où *h* est indépendant de la date d'observation et où t =1 on a alors *h* qui est constant et vaux 1)
et si on suppose que la distribution de désirabilité est une fonction γ de paramètre ν alors P est une distribution binomiale négative d'index ν et de paramètre $\frac{\beta^{-1}}{H+\beta^{-1}}$.
Si ν est égale à 1, la fonction de désirabilité est la fonction exponentielle et *P* est une distribution géométrique.

Avec les mêmes notations notre modèle s'écrit :

$[\mathrm{y}]$: $P(X=i)=\sum_{x=1}^{\infty} P(X=i/Y=x).P(Y=x)$ (ici x varie de façon discrète de 1 à ∞)

pour i = 0,1,2…….
Y est la variable aléatoire relative à la distribution de Contenu.
Nous avons supposé que le processus conditionnel est binomial (distribution d'Utilité) ou binomial négatif (distribution d'Usage).
L' analogie des formules $[\mathrm{x}]$ et $[\mathrm{y}]$ apparaît ici clairement.

Interprétation de la distribution de Contenu

Le nombre d'articles produit chaque année par un périodique est très variable. Dans le tableau 3 ci dessus nous avons relevé le nombre d'articles publiés en 1996 et 1997 par 8 titres de périodique internationaux dans le domaine de la pharmacie; nous voyons que le nombre d'articles peut varier très de façon significative d'une année à l'autre.
Aussi il n'est pas ridicule de chercher à quantifier le nombre d'articles produits par une distribution probabiliste.

Tableau 3 : Nombre d'articles publiés en 1997 et 1996 (Source JCR)

Titre des revues Pharmaceutiques	Articles 1997	Articles 1996
INTERNATIONAL JOURNAL OF PHARMACEUTICS	344	489
BIOCHEMICAL PHARMACOLOGY	376	438
BRITISH JOURNAL OF PHARMACOLOGY	704	795
DRUGS	170	164
BRITISH JOURNAL OF CLINICAL PHARMACOLOGY	190	245
ARZNEIMITTEL-FORSCHUNG (DRUG RESEARCH)	234	224
JOURNAL OF PHARMACY AND PHARMACOLOGY	258	249
JOURNAL OF ANTIMICROBIAL CHEMOTHERAPY	295	289

Dans notre modèle nous introduisons une distribution de probabilité qui quantifie le nombre d'articles par volume. (La notion de volume peut avoir plusieurs interprétations comme nous l'avons vu précédemment). D'autre part, la nature de ces articles, pour un même domaine, peut être très variée: article fondamental ouvrant une nouvelle piste de recherche, article de synthèse, etc. Parler du nombre d'articles par volume n'est pas toujours une mesure pertinente lorsque nous nous intéressons à l'usage des articles par les lecteurs. Certes le fait de remplacer ce comptage par une probabilité ne résout pas le problème lié à la remarque précédente. Aussi il peut être etre préférable d'interpréter cette distribution de Contenu comme une mesure du nombre d'articles "leader" par volume. Nous disons qu'un article est "leader" pour un volume si il est important (article écrit par un chercheur prestigieux dans un domaine) et va être susceptible non seulement d'être éventuellement demandé, mais va faire en sorte que les autres articles du même volume soient également demandés. Le fait de choisir pour distribution de Contenu une distribution géométrique qui est nécessairement décroissante paraît alors moins arbitraire, avec l'interprétation précédente et n'est pas sans rappeler la fameuse loi de Bradford (BRADFORD 1934) qui quantifie la répartition des articles traitant d'un domaine précis dans des périodiques scientifiques.

7) Cas où la distribution de contenu est binomiale

Notre modèle ne serait pas complet si on n'envisageait pas le cas où la répartition du nombre des articles dans les volumes était du au hasard *(ce qui semble « naturel » si on interprète la distribution de Contenu dans son sens courant)* c'est à dire si la distribution de Contenu était de nature binomiale (tronquée) de paramètre n et p où n désigne le nombre maximum d'articles par volume et p un nombre compris entre 0 et 1. La question posée est alors la

même que précédemment: comment se comporte la distribution d'Usage et d'Utilité (définie par les équations [A] et [B] si on suppose que la distribution de Contenue est du type binomiale avec les mêmes conditions limites? La condition (b) ici traduit le fait que le nombre maximum d'articles devient infini. On montre que l'on une distribution de Poisson d'espérance *M* (où M est définie comme précédemment).

Démonstration

On va donner ici un simple aperçu de la démonstration . Les techniques utilisées sont semblabes à celles des articles précédants(Lafouge 1999).

La distribution de Contenu *G* qui est une loi binomiale tronquée de paramètre *n* et *p* s'écrit :

$$G(j)=C_n^j.p^j.q^{n-j}\ \frac{1}{1-q^n} \qquad j=1....n$$

où l'on a : $0\le p\le 1;\ \ q=1-p;\ \ C_n^j=\frac{(n-j+1)....(n-1).n}{j!}$

D'après [A] et[B] on peut écrire:

$$Pu(0)=Pc(0)=\frac{1}{1-q^n}.\sum_{j=1}^{\infty} po^j.C_n^j.p^j.q^{n-j}$$

On notera : $P(0)=\sum_{j=1}^{\infty} po^j Pn(j)\ \ où\ \ Pn(j)=C_n^j.p^j q^{n-j}$

Le calcul des moments d'ordre k de la loi binomiale de paramètre *n* et *p* nous donne (CALOT 1988) :

$$\sum_{j=1}^{k} j^k Pn(j)=p^k.\frac{n!}{(n-k)!}$$

Pour le passage à la limite les conditions précédentes (a) (b) et (c) s'écrivent :

(a) $po\to 1 \Leftrightarrow pu\to 0$
(b) $E(G)=n.p\to\infty \Rightarrow n\to\infty$
(c) $(1-po).E(G)\to M\ \Rightarrow (1-po).n\to K$.

Le calcul de $P(0)$ nous donne :

$$P(0)=\sum_{j=1}^{n} po^j Pn(j)=\sum_{j=1}^{n}\Big(1+\sum_{k=1}^{j}(-1)^k.\frac{pu^k}{k!}.\frac{j!}{(j-k)!}\Big)Pn(j)\ )$$

$$P(0)=1+\sum_{k=1}^{j}(-1)^k\ \frac{pu^k}{k!}\Big(\sum_{j=1}^{n}\frac{j!}{(j-k)!}Pn(j)\Big)\ ;$$

$$\sum_{j=1}^{n}\frac{j!}{(j-k)!}Pn(j)=\sum_{x=1}^{k}(\sum_{j=1}^{n}Pn(j).j^{x})a(x);où\ a(k)=1$$

Soit en utilisant le résultat précédant sur les moments d'une loi binomiale :

$$=\sum_{x=1}^{k}a(x).p^{x}.n.(n-1)...(n-x+1)$$

d' où on a : $P(o)=1+\sum_{k=1}^{n}(-1)^{k}.\frac{pu^{k}}{k!}(\sum_{x=1}^{k}p^{x}.a(x).(n-x+1).....n)$

On passe à la limite sous les trois conditions (a) (b) et (c) précédentes:

$$\text{Limite } P(o)=1+\sum_{k=1}^{k=\infty}(-1)^{k}.\frac{(K.p)^{k}}{k!}=e^{-Kp}$$

D'où on déduit que l'on a : Limite $Pu(o)=\text{Limite } Pc(o)=e^{-Kp}$

Avec les mêmes techniques et les mêmes conditions on montre :

Limite $Pu(i)=LimitePc(i)=e^{-Kp}.(Kp)^{i}.\frac{1}{i!}\quad i=1,2.....$

D'où on déduit le résultat qu'à la limite Pc et Pu est une distribution de Poisson de d'espérance *M* où *M=Kp.*

CQFD

Ce résultat n'est pas sans rappeler le théorème classique de probabilité (CALOT 1988) qui montre que la distribution de Poisson d'espérance *M* est obtenue comme limite d'une distribution binomiale de paramètre *n* et *p* lorsque : $n \to \infty ; p \to 0 ; n.p \to M$.

La distribution de Poisson étant considérée comme la loi du hasard pour les évènements rares, ce résultat est conforme à notre modèle où c'est la nature de la distribution de Contenu qui est transmise à la distribution d'Usage ou d'Utilité.

8) Conclusion

Notre conclusion se situe à deux niveaux. Tout d'abord nous comparons les résultats obtenus avec notre modèle et certains modèles connus en bibliométrie, puis nous donnons une interprétation de la distribution de Contenu.

La loi empirique des 80/20 observée dans les bibliothèques qui stipule qu'environ 80% des demandes sont assurés par 20% du stock a été observée et discutée depuis longtemps (TRUESWELL 1966). Il a été également démontré (BURREL 85) le lien simple de cette loi avec la loi géométrique. D'autre part en construisant un modèle stochastique de circulation la loi binomiale négative apparaît bien comme une première approximation du phénomène. Les résultats obtenus avec notre modèle (cf Tableau 1) sont de même nature avec des hypothèses différentes: passer à la limite ici est un artifice pour tenir compte du facteur temporel.

D' autre part nous pensons que l'introduction de la distribution de Contenu est une traduction distributionnelle possible de ce qu'on appelle communément la granularité de l'information. Cette granularité est matérialisée par les valeurs discrètes de la distribution de Contenu. qui ne prends pas la valeur 0. La distribution d'Usage et d'Utilité du document qui
est liée à la distribution de Contenu, devrait nous permettre de mesurer alors « l'information utile » du document.

REFERENCES BIBLIOGRAPHIQUES

(BRADFORD 1934) S. C. BRADFORD
Sources of information on specific subjects.
26 janvier 1934, Enginering, p85-86.

(BURREL 1985) Q.L BURREL
The 80/20 Rule :library lore or statistical law
Journal of Documentation, Vol 41, N°1, mars 1985, p24-39.

(BURREL 1987) Q.L. BURREL
Predictive aspects of some bibliometrics Process.
Informetrics 87/88: Select proceedings of the first international conference on bibliometrics and theorical aspects of information retrieval
Elsevier Amsterdam 1988.

(BUCKLAND 1998) M. BUCKLAND
What is a " digital document "
Document numérique Vol N°2, Hermès, 1998.

(CALOT 1988) G. CALOT
Probabilités, théorie et applications
Chapitre 12 :Lois de probabilité discrètes d'usage courant.
Eyrolles, 290 pages, 1988.

(EGGHE 1990) L. EGGHE L., R. ROUSSEAU
Introduction to Informetrics.
Quantitative Methods in Library Documentation and Information Science.
Elsevier, 450 pages, Amsterdam, 1988.

(GARDFIELD 1997) GARFIELD E.
Dispelling A Few Commons Myths About Journal Citation Impacts.
The Scientists, Vol 11, #3, p11 February 3.

(LAFOUGE 1989) T. LAFOUGE, A. DELARBRE
Des Statistiques à la Bibliométrie.
Revue Française de Bibliométrie N°3, p38-49.

(LAFOUGE 1995) T. LAFOUGE
Mesures relatives de l'information utile dans des périodiques scientifiques.
Revue Française de Bibliométrie N°14, p135-146.

(LAFOUGE 1995) T. LAFOUGE
Information Stochastic Field
JISSI : The International Journal of Scientometrics and Informetrics.
Vol N°1 (2), juin 1995, p 57-64.

(LAFOUGE 1998) T. LAFOUGE
Mathématiques du document et de l'information : Bibliométrie distributionnelle
Thèse d'Habilitation à Diriger des Recherches
Soutenue le 15 Octobre 1998 à l'Université Jean-Moulin Lyon3
Chapitre 4 : Modélisation des processus de circulation des documents.
http://193.51.109.173/memoires/ThierryLafouge_ext.pdf

(LAFOUGE 1999) T. LAFOUGE et E. GUINET
Relations between distributions of use and distributions of contents in the case of library circulation data.
Proceedings of Seventh Conference of The International Society for Scientometrics and Infometrics (ISSI). Colima, México, 5-8 Juillet 1999. p.267-277.

(LEEMANS 1992) MJ. LEEMANS, M. MAES, R. ROUSSEAU R.,C. RUTS
The negative binomial distribution for circulation data in Flemish public libraries,
Scientometriccs, 25(1), p47-57.

(MORSE 1968) PH. MORSE
Library effectivness
The M.I.T Press, Cambridge.

(SALAUN 2000) JM. SALAUN, T. LAFOUGE ,C. BOUKACEM
Trading in ideas, articles and journals : a document supplier cade study.
(à paraître N° Spécial de Scientométrics