Applications des lois infométriques en science de l'Information: dualité, champ infométrique d'usage et de production.

Thierry Lafouge, Boucif Boukacem

Applications des lois infométriques en science de l'Information: dualité, champ infométrique d'usage et de production.

Lafouge Thierry
Boukacem Boucif

Laboratoire URSIDOC
Université Claude Bernard Lyon1
Batiment OMEGA
43, Boulevard du 11 novembre 1918
69622 Villeurbanne Cedex
04 72 43 13 91
Thierry.Lafouge@univ-lyon1.fr

Résumé

L'objet de cet article est de montrer que l'étude des distributions statistiques en science de l'information fournit des outils pour appréhender et définir certains concepts relatifs à l'usage et à la production d'information. Après avoir rapidement rappelé ce que l'on désigne par infométrie et lois infométriques, on définira ce que l'on entend par dualité et champ infométrique. Quelques repères mathématiques simples sont donnés. On s'appuie principalement sur des expérimentations faites, dans une bibliothèque, chez un fournisseur d'articles, sur la fréquentation d'un site Web. Ces dernières sont prétexte à formaliser et modéliser, à l'aide des lois de probabilité, ces phénomènes d'usage et de production d'information, et définir ce que l'on appelle un champ infométrique de production et d'usage, à l'aide des mathématiques.

Mots clef : infométrie/ usage / loi de l'information /bibliothèque / fournisseur de document

Abstract

This article aims at demonstrating that the study of statistical distributions in information science provides the tools to apprehend and define concepts related with information use and production. One will briefly outline what should informetrics laws be, then one will define what is in our sense duality and statistic information field. Simple mathematical milestones will be given. The study relies on data gathered in libraries, in a document supplier, on a Web site visits. These data allow to formalize and modelize processes of information use and production . Thus, these data lead to define the concept of informetric field of use and production, with help of mathematics.

Keywords : informetric/ document supplier/ library /use/ bibliometric distribution/

1. Scientométrie, bibliométrie, infométrie

L'objet principal de la bibliométrie consiste à analyser, à l'aide de méthodes statistiques et mathématiques, un corpus documentaire, afin d'en extraire des relations significatives entre ses divers éléments. Elle a aussi pour objet d'étudier les livres ou revues scientifiques quant à leur usage et leur production. La scientométrie (Leydesdorff, 2001) a pour objectif d'étudier, toujours à l'aide de méthodes quantitatives, les processus de création, de diffusion et d'utilisation de la science. L'infométrie qu'on désigne par *informetrics* dans la langue anglo-saxonne vise à tirer profit de l'informatique documentaire. On va passer de la quantification des éléments bibliographiques du document au contenu de l'information qu'il contient. L'amalgame de ces trois termes pour désigner ces différentes disciplines est fréquent. Nous utiliserons par la suite uniquement le terme infométrie pour désigner l'ensemble des activités métriques relatives à l'information, couvrant aussi bien la bibliométrie que la scientométrie. Il est clair que cette dernière définition nécessiterait qu'on définisse avec précision ce que l'on entend par information. Ce n'est pas l'objet de cet article, cependant il nous semble utile de dire dans cette introduction que nous soutenons l'hypothèse que l'information peut être l'objet d'étude d'une science exacte[1] ce qui n'est pas la seule posture épistémologique en science de l'information (Fondin, 2001). Ceci explique peut-être, pourquoi nos recherches ont une proximité plus forte avec les sciences exactes qu'avec les sciences humaines et sociales.

2. Les distributions statistiques en science de l'information

Nous envisageons deux approches différentes pour traiter cette question, une liée à ce que l'on appelle classiquement les lois de l'information, une deuxième liée à la théorie de la circulation et des processus.

a. Les lois infométriques

L'infométrie s'intéresse entre autres à quantifier certains phénomènes rencontrés en science de l'information. Le point crucial d'une multitude d'études est alors l'observation de fréquences d'événements appelées généralement distributions. Rappelons les trois plus célèbres :

- On constate qu'il existe une relation inverse entre le nombre de publications dans un domaine scientifique et le nombre de ses membres. Cette régularité est représentée par une fonction hyperbolique établie par A.J. Lotka en 1926 (Lotka, 1926). Ce phénomène est connu sous le nom de loi de Lotka.

- S.C. Bradford en 1930 s'est intéressé à la répartition des articles scientifiques, pour un domaine précis, dans des revues; il a montré dans un article célèbre (Bradford, 1934) que les articles scientifiques sont distribués avec une régularité remarquable dans les revues. Ce phénomène est connu sous le nom de loi de Bradford. C'est la loi la plus intrigante de notre domaine.

[1] Nous nous reconnaissons dans le courant de pensée anglo-saxon désigné souvent par « Information Science ».

- G.K. Zipf en 1935 (Zipf, 1935) constate en étudiant des corpus de données textuelles des régularités sur la fréquence d'apparition des mots. Très grossièrement, nous pouvons dire que si nous ordonnons les mots suivant leur fréquence décroissante, nous nous apercevons qu'il existe une relation entre le rang et la fréquence: le produit rang× fréquence est à peu près constant. Ce phénomène est connu sous le nom de loi de Zipf[2]. Cette loi est particulière et n'est pas à notre avis de même « nature » que les deux précédentes.

Tous ces phénomènes peuvent être représentés par un schéma fonctionnel simple (voir figure 1) que Leo Egghe (Egghe, 1990) appelle IPP « *Information Product Process* » que nous nous proposons d'appeler ici « *Champ infométrique de production* »(CIP). Un CIP est un triplet composé d'une source bibliographique (S), d'une fonction[3] de production (P) et de l'ensemble des éléments (items) produits (I).

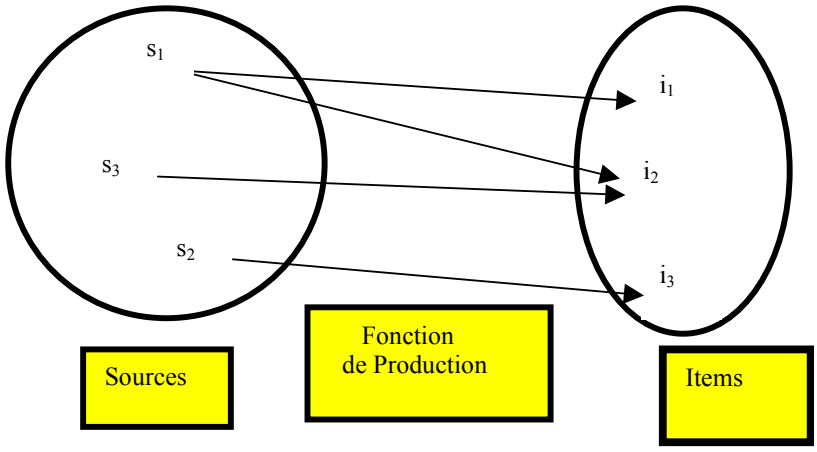


Figure -1- Champ infométrique de production

Avec ce formalisme très simple on écrit pour les trois phénomènes précédents:

- Loi de Lotka: les auteurs (sources) publient (produisent) des articles (items).
- Loi de Bradford: les revues (sources) éditent (produisent) des articles (items) sur un thème scientifique donné.
- Loi de Zipf: les mots (sources) produisent des occurrences (items).

Cette représentation montre la spécificité de la loi de Zipf où source et item sont de même nature: un mot produit une occurrence de mot. On verra par la suite que cette loi présente d'autres particularités. En bibliométrie on utilise très souvent les CIP suivants pour caractériser la distribution des mots[4] dans les articles scientifiques:

- les mots (sources) produisent des articles (items).

[2] Cette caractéristique est étudiée en linguistique quantitative (lexicométrie).
[3] Le terme fonction n'est pas à prendre au sens mathématique. Une fonction de production peut faire correspondre plusieurs items à une source. En toute rigueur on devrait parler de relation.
[4] Dans ce cas on désignera par mot une forme linguistique minimale porteuse de sens, ce qui n'est pas le cas de la loi de Zipf originelle où un mot est une chaîne de caractères délimitée par des séparateurs.

Par la suite, lorsque nous parlerons de distribution zipfienne pour la régularité des mots dans un corpus de textes, c'est à ce type de distribution que nous ferons allusion. Dans ce cas on s'intéresse uniquement à la présence ou absence du mot dans le texte[5]. Lorsque l'on analyse des références bibliographiques contenant des descripteurs, on a le CIP suivant qui est de même nature que le précédent:

- les descripteurs[6] (sources) indexent les articles référencés (items).

Les distributions décrites précédemment présentent des régularités semblables à la loi de Zipf, elles sont bien connues lorsque l'on fait des études quantitatives sur des corpus de références bibliographiques, et sont interprétées par les bibliomètres (Rostaing, 1996)[7].

Pour exploiter ces différentes régularités, la représentation la plus classique consiste à écrire ces distributions sous la forme fréquentielle suivante utilisée généralement en statistique:

$F_i \; i = 1..pmax$, désignent le nombre de sources qui ont produit i items; *pmax* étant le nombre maximum d'items produits par une source.

L'observation de ces fréquences révèle pour les phénomènes précédents des régularités du type:

- Un petit nombre de chercheurs publient beaucoup et par contre ils sont nombreux à ne publier que quelques articles.
- Un grand nombre d'articles fondamentaux (« la littérature cœur ») dans un domaine est produite par un petit nombre de revues. Un grand nombre de revues publient quelques articles dans un domaine (on parle de dispersion de la littérature scientifique).
- Un très petit nombre de descripteurs sont très utilisés alors qu'une grande partie des descripteurs ne sont utilisés qu' une seule fois.

Un des traits communs de ces distributions dans le domaine de l'information est leur grande dispersion. Leur étude confirme des régularités et des rapports mesurables, qui vont amener certains chercheurs du domaine à parler de lois de l'information; le terme lois infométriques nous semble mieux adapté.
Très grossièrement on peut dire que les régularités de ces lois sont mathématisées avec la relation hyperbolique ci-dessous[8]:

$F_i = \frac{k}{i^{a+1}} \quad i = 1,2.....$ où k et a sont des constantes positives, F_i désignant le nombre de sources théoriques qui ont produits i items.

[5] On ne compte pas le nombre d'apparitions du mot dans le texte, mais le nombre de textes dans lequel le mot est présent.
[6] Les descripteurs ne sont pas forcément des mots issus du texte.
[7] Les distributions de descripteurs sont découpées en trois zones qui répartissent empiriquement le vocabulaire et qu'on dénomment habituellement par: bruit, information intéressante, information triviale.
[8] Il existe d'autres représentations mathématiques de ces lois utilisant des techniques de rang (Lada 2000).

Ces distributions sont connues sous le nom de zipfiennes et les propriétés mathématiques de ces dernières (Haitun, 1982) ont été largement étudiées. Elles sont de forme hyperbolique (en « j renversé») décroissantes et possèdent une longue queue avec un écart type supérieur à la moyenne. On les rencontre fréquemment en sciences humaines et sociales et en science de l'information. En général on les oppose aux distributions gaussiennes rencontrées fréquemment en statistique lorsqu'on étudie des distributions physiques de populations humaines (répartition des tailles et des poids des individus d'une population…..).

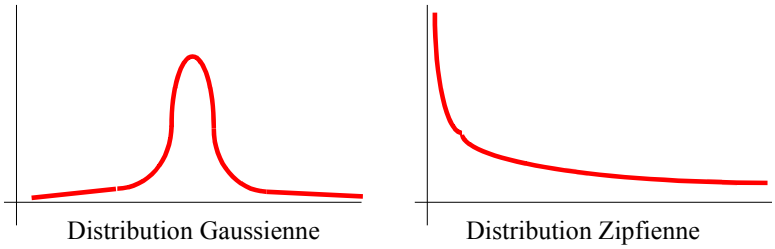


Figure -2- Distributions statistique

La théorie des probabilités permet d'expliquer la forme de ces courbes. On sait qu'une distribution gaussienne ne fait que refléter la distribution au hasard de très petits effets additifs indépendants les uns des autres. Une distribution Zipfienne résulte d'un processus dans lequel un événement élémentaire va être beaucoup plus important que les autres dans la construction du phénomène étudié.

- Un chercheur publie facilement un article parce qu'il a déjà publié plusieurs articles, un article est cité dans un article parce qu'il a déjà été cité.
  ……

- Pour la loi de Zipf originelle il est plus difficile d'expliquer ce phénomène du langage. L'usage d'un mot représentant un effort pour un locuteur, ce dernier essaye de minimiser cet effort, d'où cette distribution. Le lien entre la relation rang-fréquence et la loi du moindre effort conserve un intérêt historique et n'est plus une réalité scientifique pour l'étude de la langue. Cependant loi d'effort et distributions infométriques sont fortement liées (Lafouge et Michel, 2001) par la théorie de l'information statistique et méritent de notre part une plus grande attention.

De nombreuses autres régularités ont été observées et seront également qualifiées de loi; on parle de la loi de Brookes, Mandelbrot, Leimkmuler. Le même phénomène de non linéarité est observé. Les distributions de citations dans le domaine scientifique présentent les mêmes caractéristiques.
Des travaux mathématiques ont montré que certaines lois sont équivalentes (Egghe, 1985) et qu'on peut les classer par groupes. Deux lois sont dites équivalentes si l'une peut être déduite de l'autre et vice versa par des arguments purement logiques ou mathématiques.[9]Très souvent on aura une équivalence des lois dans des conditions extrêmes c'est-à-dire en passant à la limite: on parle alors de lois asymptotiquement équivalentes.
La formulation déterministe précédente trouve un cadre et une interprétation probabiliste. La

[9] Il est très facile de montrer mathématiquement par exemple que la formulation de la loi de Lotka et la formulation de la loi de Zipf, dans le cas idéal sont équivalentes.

plus connue est le principe des avantages cumulés (Price, 1976). Price a retenu la règle qui consiste à augmenter la probabilité que le succès engendre le succès sans avoir à tenir compte de l'influence des échecs. Pour ce faire il utilise le modèle de l'urne de Polya qui permet à l'aide de la combinatoire de générer les lois de probabilité discrètes courantes (binomiale, hypergéométrique, binomiale négative……) (Reyni, 1966, chapitre 3) et explique pourquoi on obtient des distributions hyperboliques.

b. Les lois de circulation

Il n'existe pas à notre connaissance de loi empirique (comme la loi de Lotka, de Bradford,….) relative aux usages des documents. Historiquement ce sont les distributions relatives aux usages des ouvrages dans les bibliothèques qui ont été observées les premières. Etant donné une collection d'ouvrages, on s'intéresse durant une période de temps fixé (un an, un mois…) au nombre d'emprunts de chaque document de ce corpus.
Nous avons pour notre part étudié ce type de phénomène en bibliothèque (Lafouge, 1989) en utilisant les travaux de Morse en recherche opérationnelle pour modéliser les distributions de circulation d'ouvrages. Ces derniers s'inscrivent plutôt dans ce que les anglo-saxons (Sengupta, 1992) désignent sous le terme de *librametry,* c'est-à-dire dans l'utilisation des méthodes quantitatives dans le domaine de la gestion bibliothéconomique. Plus généralement toutes ces méthodes quantitatives sont utilisées dans le contexte d'évaluation des systèmes d'information: bibliothèques, centres documentaires, fournisseurs de documents, musées, services Web…..
La communauté des chercheurs qui a étudié les phénomènes de circulation a utilisé la théorie classique des processus stochastiques. Par exemple dans le modèle des emprunts d'ouvrage développé principalement par Burrel (Burrel, 1987) on suppose que pendant un intervalle de temps fixé les emprunts d'ouvrages se comportent comme un processus poissonien avec une moyenne qui varie d'un ouvrage à l'autre suivant une loi de probabilité connue. On peut écrire mathématiquement cette relation:

$P_i = \int_0^{\infty} \pi(h(t)x)(i).f(x).dx \quad i = 0,1...$ où $\pi(h(t)x)$ est une loi de Poisson de moyenne $h(t)x$, h une fonction du temps, f une fonction de densité d'une loi de probabilité continue. P_i désigne la probabilité qu'un ouvrage soit emprunté i fois pendant l'intervalle de temps [o T]

Sous certaines conditions, ce modèle donne une distribution binomiale négative. Cette loi de probabilité discrète a été vérifiée très souvent pour la circulation d'ouvrages dans une bibliothèque. Dans beaucoup de cas l'approximation par une loi géométrique simple est suffisante.

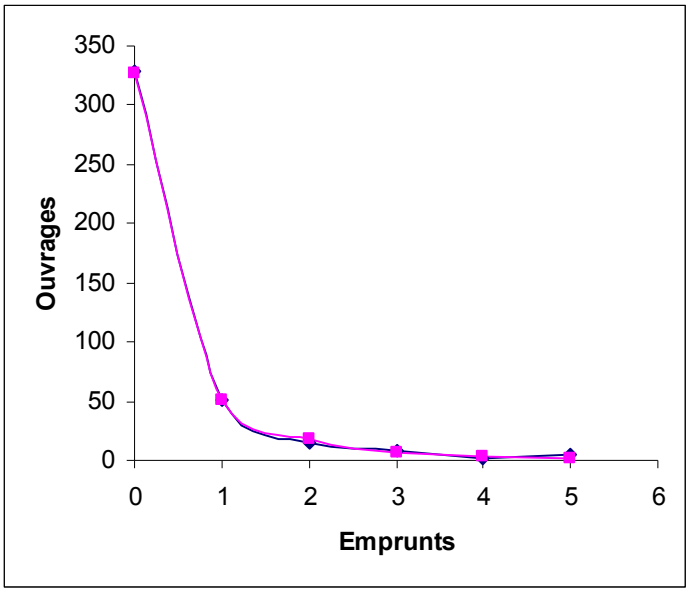
Exemple-1

On a relevé le nombre d'emprunts de la collection Payot-Science durant l'année 1984 à la bibliothèque municipale de Bordeaux. Une grande partie de ce fonds, 80%, n'a pas circulé durant cette année. On fait l'hypothèse que les effectifs sont distribués suivant une loi binomiale

négative; on calcule[10] alors ces derniers puis on les compare aux effectifs observés.

Emprunts	Ouvrages Observés	Ouvrages Attendus
0	328	326,73
1	51	51,36
2	15	17,42
3	8	6,97
4	1	3,00
5	5	1,35
Total	408	

Tableau - 1- Emprunts des ouvrages de la collection Payot-Science en 1984 à la bibliothèque municipale de Bordeaux.



Graphe -1 - Distribution d'Usage de la collection Payot-Science à la bibliothèque municipale de Bordeaux en 1984.

Nous constatons que le nombre d'ouvrages observés dans chaque classe d'emprunts est très près du nombre d'ouvrages prévus par le modèle.[11]

Les distributions relatives à la circulation de l'information sont caractérisées entre autres par deux

[10] On utilise la méthode des moments pour calculer les paramètres de la loi.
[11] Le test statistique du Khi² permet de valider l'hypothèse: « la distribution est binomiale négative ».

paramètres: le temps, les « no use »[12].

.

♦ Le facteur temps

Il est implicite dans beaucoup de distributions au moment de la constitution du corpus. Lorsque l'on cherche à vérifier la loi de Lotka par exemple, la question est: pendant combien d'années observe-t-on la production d'articles d'une communauté de chercheurs? Dans la formulation des lois précédentes (paragraphe 2. a) le paramètre temps n'est pas modélisé, on parle dès lors de distributions stationnaires. La formulation de Burrel de la page 6, utilisée ici pour des processus de circulation d'ouvrages est générale: elle permet de formuler tous les processus infométriques où des sources produisent des items pendant une période de temps donnée.

♦ Les « no use »

Ce deuxième facteur est un point crucial dans ce type de distribution : en effet ces fréquences ne sont pas directement observables et nécessitent qu'on précise bien les conditions d'expérimentation. Si on observe par exemple dans une bibliothèque les emprunts d'ouvrages, on n'observe pas le même phénomène à la banque de prêt ou dans la réserve de la bibliothèque.
Si on peut parler du nombre de chercheurs qui n'ont publié aucun article, du nombre d'ouvrages qui n'ont jamais été empruntés, cela n'a pas grand sens de parler du nombre de descripteurs qui n'apparaissent jamais dans l'indexation d'un corpus d'articles sauf peut être dans le cas d'un vocabulaire contrôlé. De plus on remarquera que le modèle hyperbolique ne nous permet pas de prendre en compte le cas des « no use », ce qui n'est pas le cas des lois de probabilité discrètes comme on vient de le voir dans l'exemple précédent. Les deux approches loi de l'information et loi de circulation sont complémentaires; elles suscitent souvent des polémiques (Burrel, 2001) d'ordre mathématique que nous n'aborderons pas ici.

3. Les distribution d'usage et de production: dualité en infométrie.

De nombreux autres phénomènes liés à l'usage de l'information, c'est-à-dire à de multiples processus informationnels sont de même nature et produisent des régularités statistiques semblables. Ces distributions relatives à ces processus sont dites distributions d'usage et peuvent se formuler comme précédemment en terme de production, on peut citer:

- ♦ des ouvrages (sources) suscitent (produisent) des emprunts (items),
- ♦ des articles (sources) suscitent (produisent) des commandes (items),
- ♦ des sites Web (sources) génèrent (produisent) des visites (items).

Ces trois formulations s'expriment de façon duale[13] en faisant intervenir une autre source, une autre fonction de production et le même ensemble d'items.

- ♦ des lecteurs (sources) font (font usage) des emprunts (items),

[12] On désignera par « no use » le cas où la fréquence d'événements est nulle: revue scientifique qui ne produit aucun article sur un sujet donné, ouvrage qui n'est jamais emprunté……..

[13] La notion de dualité ici est différente de celle de L. Egghe (Egghe 1990) qui est mathématique.

- des usagers (sources) font (font usage) des commandes (items),
- des internautes (sources) font (font usage) des visites (items).

La définition de la distribution duale nécessite cinq éléments: deux sources bibliographiques, deux fonctions, une dite de production, l'autre dite d'usage, et l'ensemble des items produits. On suppose que chaque item est produit par au moins une source, et utilisé par au moins une source.

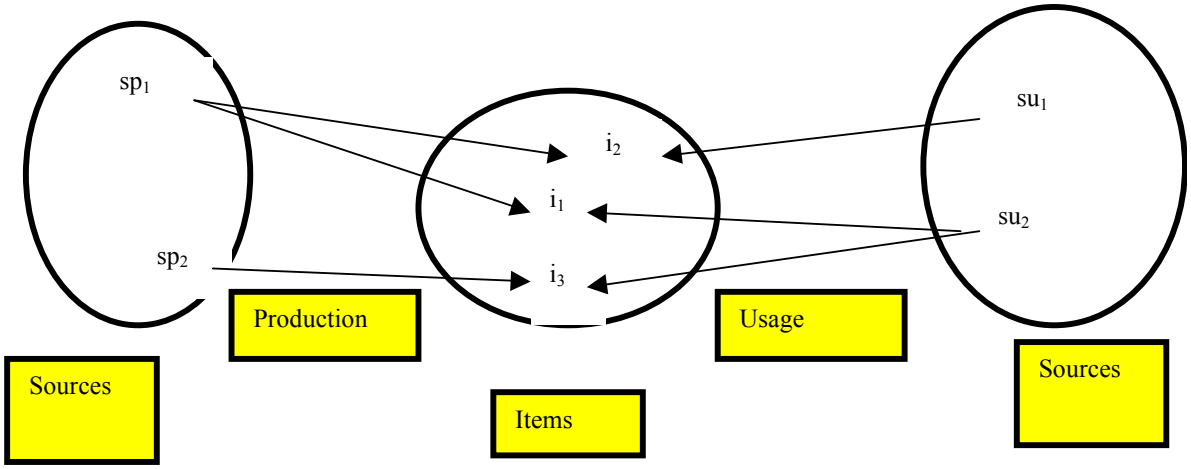


Figure 3 – Distributions duales en bibliométrie

Dans les exemples précédents les deux distributions que l'on dira duales sont indépendantes: on ne peut déduire l'une de l'autre. On sait simplement que le nombre d'items produits est égal au nombre d'items utilisés, ce qui se traduit à l'aide des fréquences par l'égalité mathématique triviale suivante:

$\sum_{i=1}^{p\max} FP_i.i = \sum_{j=1}^{u\max} FU_j.j$ FP_i désignent le nombre de sources (sources productrices) qui ont produit *i* items, (*pmax* désignant le maximum), FU_j désigne le nombre de sources (sources utilisatrices) qui ont utilisé *j* items (*umax* désignant le maximum).

Très souvent les deux distributions $\{(FP_i)\ (FU_i)\}$ sont de nature hyperbolique. C'est ce qui nous permet de dire qu'en infométrie la production et l'usage de l'information sont deux processus de même nature qu'on ne peut distinguer.

Exemple-2

Notre corpus est celui des commandes d'articles à l'Inist[14] au mois de janvier 1997, soit 50000 commandes. On trouvera dans (Salaün, Lafouge, et Boukacem 2000) une étude bibliométrique complète à partir de ces données. Nous travaillons ici uniquement sur les données de la première

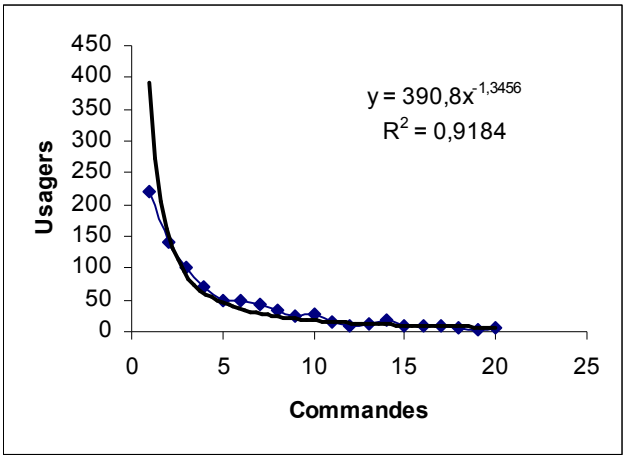
[14] Institut national d'information scientifique et technique :http:\\www.inist.fr.

semaine, soit 14000 commandes. Chaque commande est caractérisée par deux codes, un identifiant la revue, l'autre le client. On peut alors construire les deux distributions duales de production (les revues produisent des commandes) et d'usage (les clients font usage de commandes).

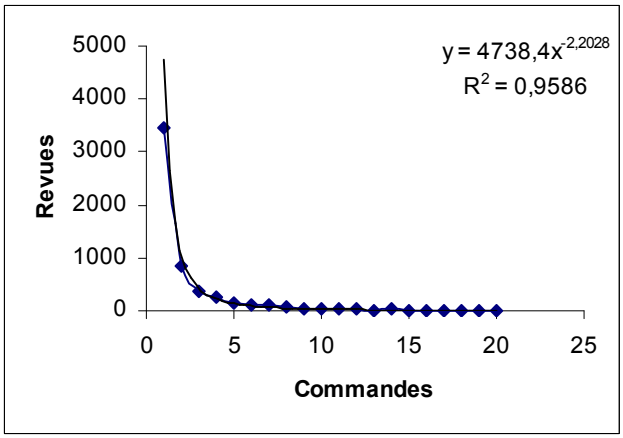
Commandes	Usagers	Revues
1	220	344
2	142	859
3	100	375
4	70	248
5	50	141
6	48	105
7	42	92
8	34	56
9	25	46
10	27	37
11	16	35
12	8	24
13	11	12
14	17	20
15	9	15
16	8	17
17	10	8
18	6	7
19	4	5
20	6	2
Plus de 20	147	47
Total	1000	5595

Tableau - 2 – Commandes des revues à l'Inist durant la première semaine de janvier 1997

La troncature du nombre de commandes à 20 dans le tableau, ne nous permet pas de vérifier que le nombre total de commandes calculé, soit à partir des revues, soit des usagers, donne le même résultat. En réalité il existe une revue qui a produit 117 commandes et un usager qui a passé 547 commandes.



Graphe -2- Distribution d'usage des usagers à l'Inist durant la première semaine de janvier 1997



Graphe -2 bis - Distribution de production des revues à l'Inist durant la première semaine de janvier 1997

Nous avons fait pour chaque distribution un ajustement de type hyperbolique (voir page 4 de l'article): nous faisons une régression linéaire après avoir transformé les coordonnées sur une échelle logarithmique . R^2 est le carré du coefficient de détermination de la régression linéaire. On remarquera que l'ajustement de la distribution de production des périodiques est meilleure que celui de la distribution d'usage des usagers. Pour cette dernière un modèle exponentiel donnerait de meilleurs résultats. On se trouve la devant les deux grands types de distributions rencontrées fréquemment en science de l'information: les distributions en fonction puissance et les distributions exponentielles (Barbut 1990).

Les deux distributions duales sont de nature zipfienne[15] :
Un grand nombre de périodiques (62%) n'est utilisé une seule fois, tandis qu'un petit nombre de périodiques sont utilisés très souvent. D'autre part un grand nombre d'usagers, en moins grande proportion (22%), ne fait qu'une seule commande tandis qu'un petit nombre d'usagers fait beaucoup de commandes. La dispersion pour la distribution des usagers est beaucoup moins forte

[15] Elles sont toutes les deux décroissantes et ont un écart type (3,96 pour les revues, 35,24 pour les usagers) supérieur à la moyenne (2,5 pour les revues, 14 pour les usagers).

que pour les revues[16]. Le nombre de commandes très importantes sont dues à des organismes qui groupent leur achat (un seul code client pour un grand institut de recherche).

A propos de la dualité des lois infométriques

Quelle est la distribution duale de Lotka, Bradford et Zipf ?
Pour Lotka la dualité nous amène à considérer le CIP suivant: des revues (sources) produisent des articles (items); ce sont les revues où sont publiés les articles des chercheurs; si les chercheurs travaillent sur une thématique commune, on observera une concentration d'articles publiés dans quelques revues. La distribution duale de Lotka est celle de Bradford. Pour montrer ce résultat il est nécessaire de mettre en place des expérimentations.
Pour Bradford on aura: des auteurs (sources) publient des articles (items). La distribution duale de Bradford est alors celle de Lotka, ce qui n'est pas surprenant d'après la définition de la dualité !
Peut-on observer simultanément les deux lois, c'est à dire les deux types de régularités mathématiques?[17] Tout dépendra de la limitation du corpus. En général on délimite un ensemble de sources, les items correspondants puis l'autre ensemble source.
Dans l'exemple précédent, on a choisi d'abord les sources, les revues de l'Inist qui ont fait l'objet d'au moins une commande, puis les items produits que sont les commandes et enfin le deuxième ensemble de sources, qui sont les usagers qui ont fait ces commandes. La différence entre les deux ensembles de sources est claire. Toutes les revues sont attachées à l'Inist. Elles ont un dénominateur commun fort, ce qui n'est pas le cas des usagers. L'article de la revue commandé est un « document situé » (Lafouge, 1998 - chapitre 2) (c'est à dire localisé dans l'espace et muni de deux propriétés qui sont son type de support et sa localisation) qui est extrait d'une collection
La loi originelle de Zipf n'a pas de distribution duale[18]. Cette remarque montre bien la spécificité de cette dernière en infométrie qui n'est pas une distribution d'usage de même nature: cependant cette dernière n'est pas étrangère à notre discipline : elle est reliée à la théorie statistique de l'information et connaît des applications dans le domaine de l'indexation (Losee, 2001).
Une autre approche pour la loi de Zipf nous semble nécessaire. Il faut introduire une deuxième source qui serait le « lexique » ; la fonction de production serait : les entrées du lexique produisent des formes graphiques.

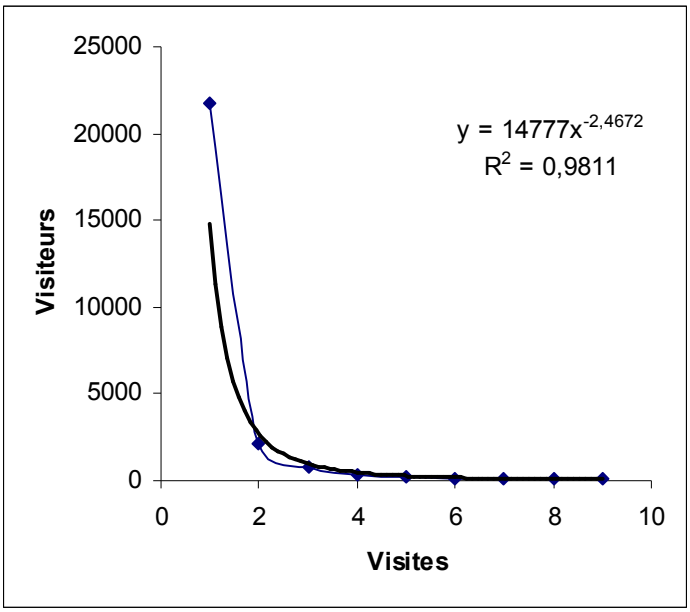
Il serait paradoxal aujourd'hui de ne pas parler d'Intenet. Les lois infométriques se vérifient aussi dans cet environnement (Egghe, 2000) où l'on observe les mêmes phénomènes.

Exemple –3
Soit un site web quelconque où on comptabilise les visites des internautes. Le CIP correspondant est :
Sources = ensemble des internautes ayant visité le site durant une période de temps,
Items = ensemble des visites sur le site durant la même période.
Ce type d'étude donne au webmestre des informations sur la fréquentation de son site. Il lui permet de répondre à la question: dans quelle proportion les internautes ont-ils tendance à visiter plusieurs fois le site ?

[16] On peut remarquer que le coefficient calculé lors de l'ajustement hyperbolique est plus faible pour la distribution des usagers (1,34) que pour la distribution des revues (2,2).
[17] Nous ne connaissons pas d'études bibliométriques dont l'objet serait celui-ci.
[18] L'ensemble des sources de la distribution duale serait réduit dans ce cas à un seul élément : l'auteur du texte.

Visites	Visiteurs	% Visiteurs
1	21735	81,5
2	2178	8,2
3	774	2,9
4	374	1,4
5	258	1
6	159	0,6
7	116	0,4
8	115	0,4
9	86	0,3
Plus de 9	872	3,3
Total	26667	100

Tableau -3- Fréquentation du site Web de l'Enssib[19] décembre 2001



Graphe -3 - Distribution d'usage des visites du site web de l'Enssib

[19] Ecole nationale supérieure des sciences de l'information et des bibliothèques :http:\\www.enssib.fr

L'ensemble des sources de la distribution duale correspondante pourrait être l'ensemble constitué des sous ensembles de pages du site Web qui ont produit les visites. Cette distribution quantifierait le nombre de pages utiles dans un site.

Cette dualité nous amènera à parler indifféremment de champ infométrique de production et /ou d'usage.

4. Champ infométrique de production ou d'usage

Nous allons maintenant enrichir le modèle précédant, en utilisant le formalisme des CIP , à partir d'un exemple pris dans le domaine des usages, celui de la fourniture d'articles à l'Inist (Lafouge,1998, chapitre 4) en introduisant la distribution de contenu.

a. Usage des articles (Exemple-4)

On a collecté durant l'année 1985 à l'Inist les demandes de photocopies d'articles de 13 titres de revues scientifiques dans le domaine: « Parfums, cosmétiques, corps gras ».

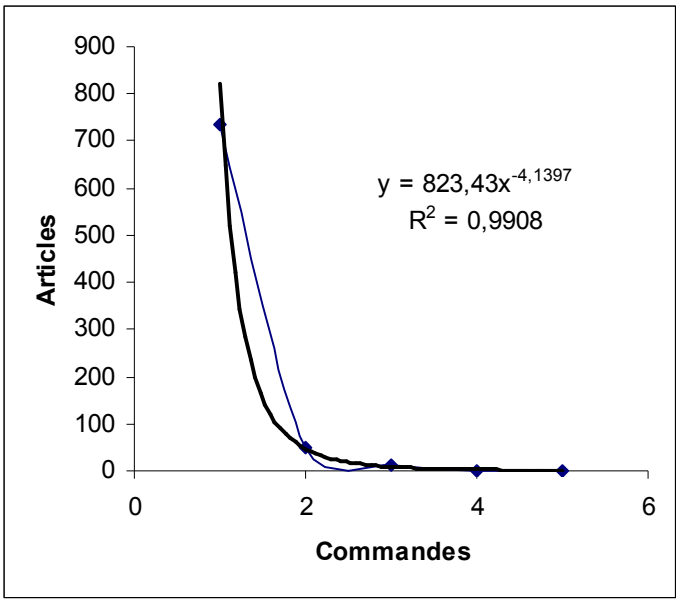
Commandes i	**Articles** FUP_i	**Volumes** FUS_i
1	734	382
2	49	70
3	13	37
4	2	21
5	1	15
6		6
7		2
8		1
9		1
10		0
Plus de 10		2
Total	799	537

Tableau – 4 -Commande d'articles à l'Inist pour l'année 1985 pour 13 titres dans le domaine « Parfums, cosmétiques, corps gras »

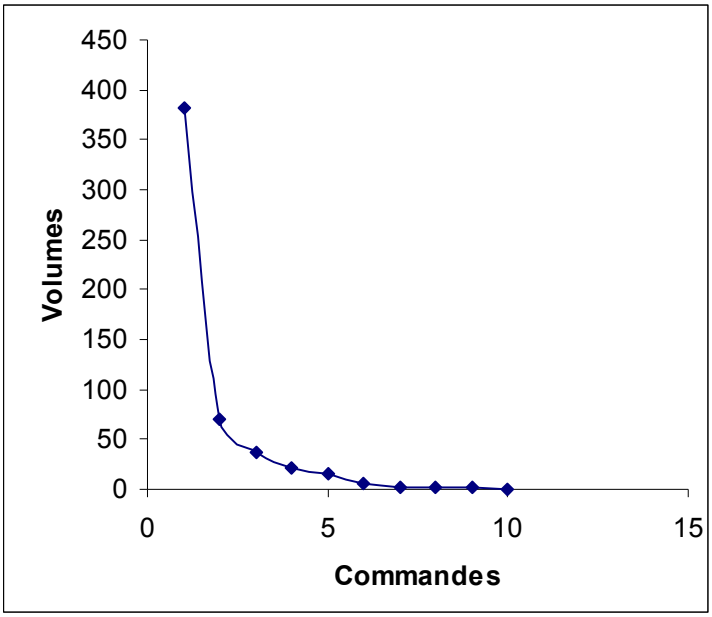
FUP_i est le nombre d'articles demandés i fois et représente la distribution de production ou d'usage vue précédemment. FUS_i est le nombre de volumes (chacun des 13 titres des revues est

composée de volumes, chacun contenant des articles) demandés *i* fois. Dans le premier cas on mesure l'usage car c'est l'article qui est demandé et commandé par l'usager. Dans le second cas, la mesure de l'usage est moins directe. Le volume n'est pas demandé, il a peut être servi de support pour cette commande, c'est en consultant ce volume que l'usager a pris connaissance de l'article. On n'oubliera pas le fait qu'un volume peut être demandé *i* fois, un seul article du volume étant commandé. D'autre part il existe des articles appartenant à des volumes demandés qui ne sont jamais commandés.

Le nombre d'usages est égale à $\sum_{i=1}^{5} FUP_i.i$, soit après calcul 884 demandes de photocopies d'articles. Ce dernier peut aussi être calculé par la formule $\sum_{i=1}^{11} FUS_i.i$, soit après calcul 881. La différence s'explique par la troncature.



Graphe - 4 – Distribution d'usage des articles de la collection « Parfums, cosmétiques, corps gras » en 1985à l'Inist.



Graphe -4 bis– Distribution d'usage des volumes de la collection « Parfums, cosmétiques, corps gras » en 1985 à l'Inist

b. **Définition d'un champ infométrique de production ou d'usage avec distribution de contenu**

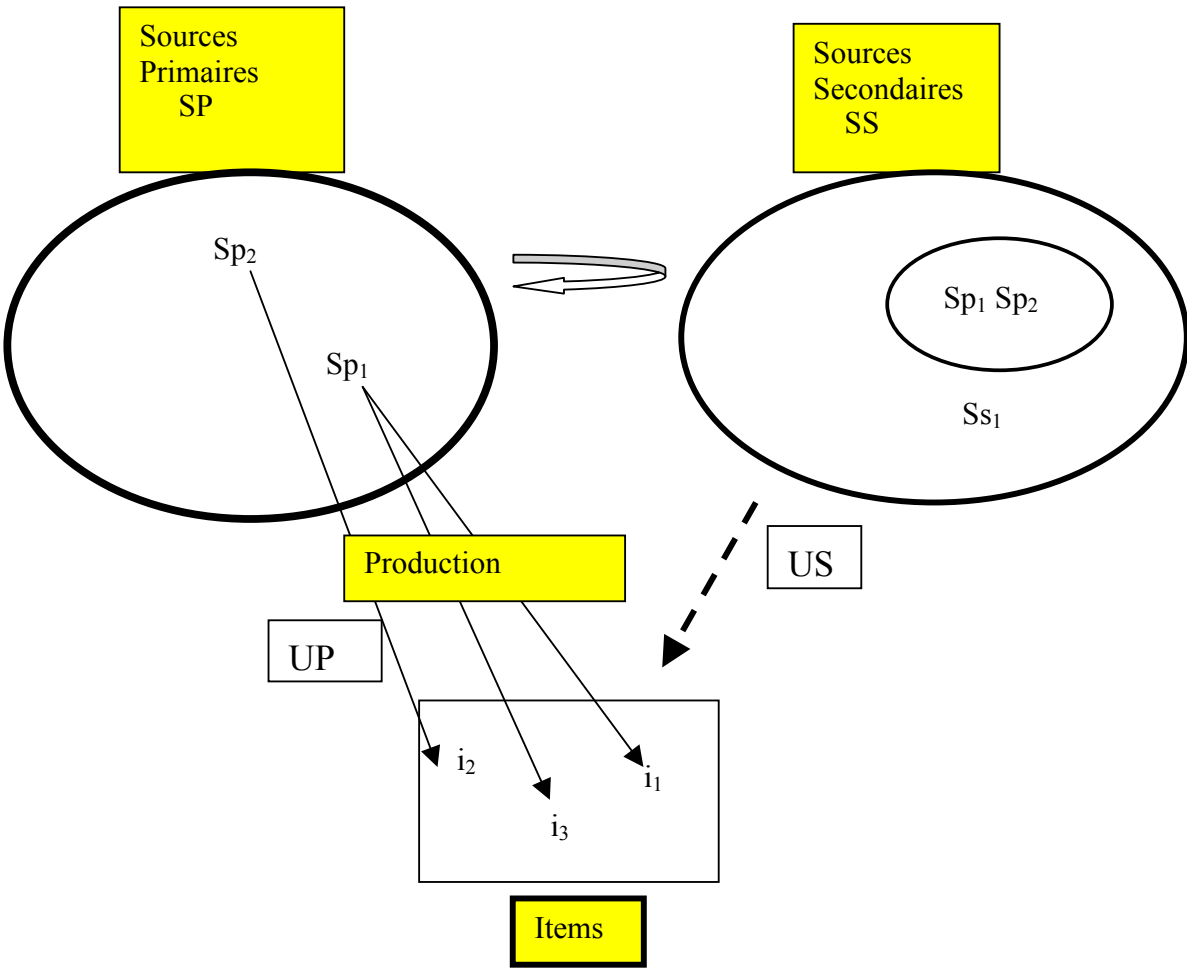


Figure –4- Champ infométrique d'usage avec distribution de contenu

Généralisons la situation précédente. Appelons SP et SS les deux ensembles de sources productrices, nommés respectivement ensemble des sources primaires, ensemble des sources secondaires. I est l'ensemble des items produits par ces sources. UP et US désignent les deux fonctions de production ou d'usage correspondantes. Les deux sources sont dépendantes l'une de l'autre. Toute source secondaire est réunion de sources primaires. Le nombre total de sources primaires est donc toujours supérieur ou égal au nombre de sources secondaires. D'autre part on fait l'hypothèse que toute source primaire appartient au plus à une source secondaire.

Les propriétés mathématiques d'un tel champ informationnel se traduisent à l'aide des fréquences par les relations mathématiques triviales ci-dessous :

$\sum_{i=1}^{p\,max} FUP_i.i = \sum_{j=1}^{s\,max} FUS_j.j$, FUP_i désigne le nombre de sources (sources primaires) qui ont produit i items (pmax désignant le maximum), FUS_j désigne le nombre de sources (sources secondaires) qui ont produit *j* items (smax désignant le maximum).

$\sum_{i=1}^{p\,max} FUP_i \leq \sum_{j=1}^{s\,max} FUS_j$, si on a l'égalité les deux distributions sont identiques, c'est à dire

$$FUP_i = FUS_i \quad i = 1, p\,max = s\,max$$

Dans la pratique on détermine en général l'ensemble des sources secondaires, puis on en déduit l'ensemble des sources primaires. Dans l'exemple précédent, on a choisi les 13 titres d'une collection dont on a étudié les commandes d'articles pendant l'année 1985.

En introduisant la distribution de contenu entre revues et articles on a le schéma ci dessous.

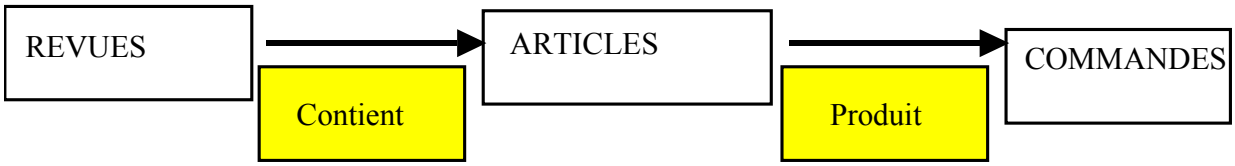


Figure -4-bis – Champ infométrique d'usage de fourniture d'articles

Donnons des exemples dans d'autres domaines que celui de la fourniture d'articles.

♦ Analyse des citations

Des articles (sources primaires) produisent des citations (items), des revues (source secondaires) produisent des citations. Une revue, qui est un ensemble d'articles, est dite citée lorsqu'un article de cette dernière est cité (Voir le calcul du facteur d'impact de l'ISI[20]).

♦ Production scientifique

Des chercheurs (sources primaires) produisent des documents au sein d'équipes de recherche (sources secondaires);
la production d'une équipe est la somme de tous ces documents

♦ Visites des sites Web

Des pages (sources primaires) produisent des visites. Les pages appartiennent à des sites (sources

[20] Institut for Scientific Information: http://www.isinet.com/isi/

secondaires) qui sont visités.

Dans les trois cas il peut y avoir des sources primaires (articles, chercheurs, pages) qui ne produisent aucun item (citations, articles, visites) et qui appartiennent à des sources secondaires qui produisent des items.

On distinguera deux types de champ infométrique :

fermé : toute source secondaire produit au moins un item,
ouvert : il existe des sources secondaires qui ne produisent aucun item.
Le champ étudié précédemment était fermé: chaque revue sélectionnée a donné lieu à au moins une commande d'un article.

c. Distribution de contenu

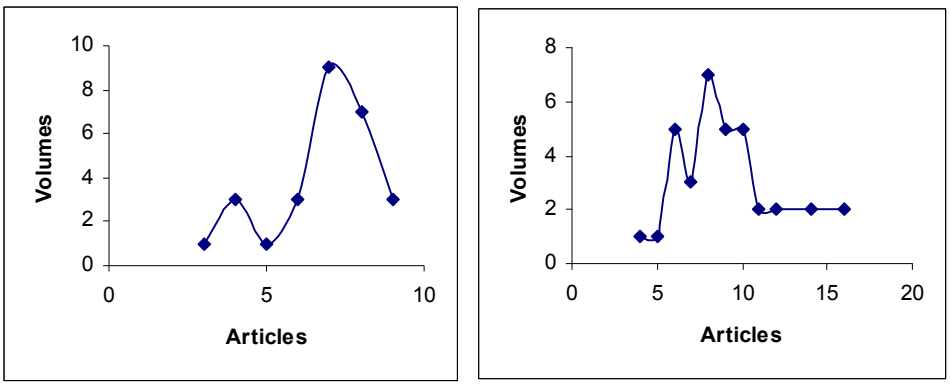
La définition de champ infométrique passe par celle de distribution de contenu, qui exprime le fait que toute source secondaire est réunion de sources primaires. Nous allons donner des exemples dans le domaine des articles de revues afin d'illustrer cette notion de distribution de contenu en nous situant à deux niveaux : celui du volume de la revue, puis celui de la revue elle-même.

1. Nombre d'articles par volume de revue

Dans ce cas les sources secondaires sont les volumes des revues.
Pour ce faire on a comptabilisé le nombre d'articles de 35 volumes de la revue *Scientometrics,* de 1997 à 2001, et de 27 volumes de la revue *Journal of Information Science* de 1994 à 1998.

Articles	Volumes Scientometrics	Volumes JIS
3		1
4	1	3
5	1	1
6	5	3
7	3	9
8	7	7
9	5	3
10	5	
11	2	
12	2	
14	2	
16	2	
Somme	35	27

Tableau –5- Nombre d'articles par volume dans les revues *Scientometrics et JIS*



Graphe –5 Distributions de contenu des volumes des revues *Scientometricst JIS*

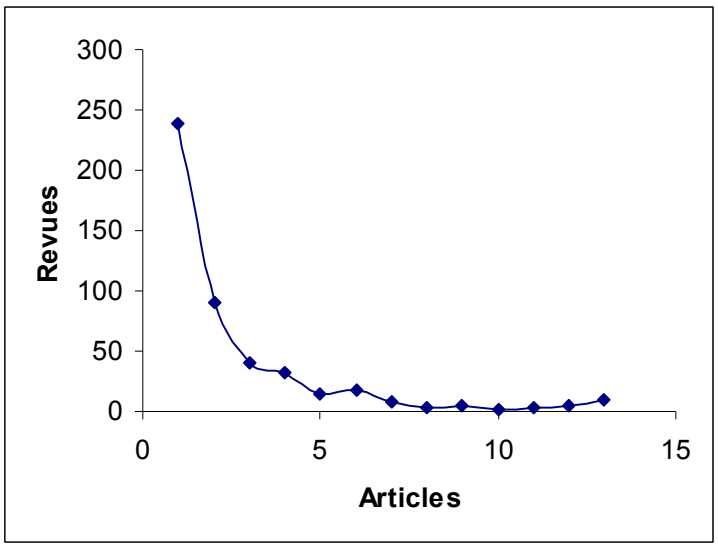
Ces deux distributions n'ont pas de forme régulière. Pour certaines revues le nombre d'articles par volume est constant, c'est le cas par exemple de la revue *Journal of Documentation* qui a entre 4 et 5 articles par numéro.

2. Nombre d'articles par revue

Dans ce cas les sources secondaires sont les revues elles-mêmes.
Pour ce faire on a comptabilisé le nombre d'articles de 471 revues dans le domaine des sciences exactes extraites du *JCR* de l'ISI de 1999 . On présente les résultats (voir tableau et graphique 6) après voir fait des classes d'amplitude de 50.

Articles	Revues
[0 50[	239
[50 100[	91
[100 150[	41
[150 200[	33
[200 250[	15
[250 300]	17
[300 350[	8
[350 400[	3
[400 450[	5
[450 500[	2
[500 550[	3
[550 600[	5
>600	9
Somme	471

Tableau –6- Nombre d'articles produits par 471 revues *JCR 1999*



Graphe –6- Distribution de contenu[21] des articles des revues-*JCR - 1999*

d. Modèle mathématique

Avant de passer à la modélisation voici quelques remarques qui délimitent le modèle mathématique qui caractérise l'usage des articles des revue scientifiques:
La commande du texte de l'article est une pratique de recherche documentaire courante dans le domaine des sciences exactes. Bien sûr, le fait que celui-ci soit publié dans une revue « renommée » peut être un critère, en plus ou complémentaire, qui déclenche la commande. En sciences sociales et humaines, les pratiques de recherche ne sont pas exactement semblables, l'article ne joue pas le même rôle informatif. On va plutôt avoir tendance à consulter un numéro thématique d'une revue. D'autre part beaucoup d'ouvrages spécialisés sont en fait composés d'une suite de points de vue de plusieurs auteurs et sont plus près des revues thématiques que des ouvrages classiques.

Le modèle mathématique déjà utilisé est indifféremment pour l'usage des revues ou l'usage des volumes de revues. La distribution de contenu concerne donc aussi bien le nombre d'articles par revue ou le nombre d'articles par volume de revues. On explicitera les notations du modèle au niveau de la revue et non pas du volume.

Nous avons construit un modèle mathématique simple pour un champ ouvert qui rend compte des liens entre:
1) La distribution d'usage des revues,
2) L''usage des articles,
3) La distribution de contenu des articles dans les revues.

[21] L'axe horizontal est gradué en classes : 5 = nombre de revues ayant entre 200 et 250 articles.

Notre modèle est de nature probabiliste, nos variables sont les suivantes:

$V(i)\ \ i = 0,1.....$: probabilité qu'une revue soit demandée i fois,

p_o : probabilité qu'un article ne soit jamais commandé,

$G(j)\ \ j = 1,2.....$: probabilité qu'une revue contienne j articles.

On introduit la suite de lois de probabilités suivantes : $Pj(p_o) : j = 1,2.....$ fonction de p_o,

$P_j(p_o)(i) : i = 0,1....$: probabilité qu'une revue ayant j articles soit demandée i fois.

L'axiome d'additivité des probabilité nous permet d'écrire : $V(i) = \sum_{i=1}^{\infty} P_j(p_o)(i).G(j)$ [1]

La distribution d'usage (V) est une distribution semblable à celle que l'on observe dans les bibliothèques . Elle est obtenue après comptage: c'est la commande du texte de l'article dans une revue qui fait que cette revue est dite demandée ou circule[22]. Pour la distribution d'usage (V) les sources secondaires sont situées : on parle de la collection de revues de l'Inist dans un domaine, comme de la collection la Pleiade à la bibliothèque municipale de Bordeaux. Dans cette étude l'Inist fonctionne suivant le même modèle qu'une bibliothèque.

Comment vérifier un tel modèle ? En d'autre terme comment utiliser les mathématiques.

Approche positiviste

Lorsque nous avons à comparer entre elles des distributions observées, la méthode la plus élégante consiste, chaque fois que les données s'y prêtent, à ajuster chaque distribution empirique, par une distribution théorique d'un type donné. Cette méthode doit être préconisée chaque fois que nous avons un modèle explicatif raisonnable.
C'est ce que l'on a fait (Lafouge, 1998, chapitre 4) avec les commandes d'articles à l'Inist: on a explicité précédemment un échantillon des données que l'on a analysé (voir exemple 4).
Pour cette expérimentation la distribution d'usage des revues ($V(i)\ i = 1,2...$) est connue (On est ici dans le cas d'un champ infométrique fermé, pas de « no se ») et la distribution de contenu est relative au volume.
La probabilité qu'un article ne soit jamais demandé notée p_o et la distribution de contenu notée G sont inconnues; par contre on connaît en partie la distribution d'usage des articles. Nous utilisons ici une voie classique en statistique, qui consiste à ajuster chaque distribution empirique observée à une distribution théorique d'un type donné dont la forme est générée par l'équation [1] .

Approche mathématique[23]

[22] Dans le cadre de l'Inist la revue circule car c'est à partir d'elle qu'on fait une photocopie de l'article commandé.
[23] Le terme mathématique ici peut sembler curieux. Il aurait été peut -être préférable de parler d'approche constructiviste. C'est une autre manière d'utiliser les mathématiques que nous proposons ici .

Ce n'est pas exactement cette voie que nous avons suivie dans nos travaux (Lafouge, 2001). Nous avons voulu découvrir les propriétés que « cachait » ce modèle. Plus précisément nous avons voulu donner un sens à ce modèle en passant à la limite sans nous préoccuper des données observées pour l'instant.

Dans notre modèle (voir encadré ci-dessus):

- La distribution d'usage des articles n'est pas prise directement en compte; elle est remplacée par une série de lois de probabilité dépendantes de j et p_o ($Pj(p_o): j = 1,2.....$) j est le nombre d'articles par revues (c'est donc obligatoirement un entier strictement positif) et p_o est la proportion d'articles n'ayant jamais été commandés.

- La distribution de contenu (G) quantifie le nombre d'articles par revue. Nous faisons une autre hypothèse plus audacieuse: nous supposons que cette distribution[24] de contenu qui est nécessairement discrète quantifie « l'information » des revues .

La question que l'on se pose est alors la suivante. Comment se comporte la distribution d'usage (V) lorsque la proportion d'articles commandés ($1-p_o$) devient de plus en plus petite ($(1-po) \to 0$) et qu'en parallèle le nombre moyen d'articles par revues ($E(G)$) lui devient de plus en plus grand ($E(G) \to \infty$) ? Avant de répondre à cette question nous devons préciser deux point:

1) Quel est le sens autre que mathématique de ces conditions limites[25]?
2) Que voulons-nous dire par : « comment se comporte la distribution d'usage (V) »?

1) Sens des conditions limites

La première est liée, nous semble-t-il, au concept d'obsolescence de l'information: celui-ci est apparu lorsque l'on s'est intéressé à l'usage des travaux passés. Un article n'est plus commandé (ou plus cité) au delà d'un certain délai depuis sa date de parution.

[24] Nous supposons qu'elle possède au minimum un moment d'ordre 1 que nous noterons $E(G)$: cette hypothèse mathématique est restrictive.

[25] Le passage à la limite est possible uniquement si la distribution de contenu est relative aux revues.

La deuxième est liée à l'explosion de la quantité d'information. Le nombre d'articles parus ne cesse d'augmenter alors que le nombre de revues reste stable. Ceci est valable sur des périodes de temps restreintes et pour des domaine de recherche stable.
Ces deux concepts sont liés. C''est parce que la quantité d'information augmente très vite que le taux dl'obsolescence augmente également ; attention, cela ne signifie pas forcément que les articles plus anciens perdent de la valeur scientifique mais bien que les articles plus récents reçoivent un surplus de citations ou de commandes. Aussi il n'est pas ridicule de supposer que obsolescence et quantité d'information sont liées: plus précisément on fait l'hypothèse que $E(G).(1-p_o)$ tend vers une limite finie, ce qui signifie que l'on est dans un cas stationnaire.

2) Comportement de la distribution V

On observe très souvent des distributions d'usage dans les bibliothèques qui s'ajustent suivant des lois simples de type poisson, géométrique, binomiale négative. La question que nous nous posons est alors la suivante. N'est-ce-pas parce que la distribution de contenu (G) est d'un type particulier que la distribution d'usage qu'on observe est du même type ?

3) Résultats

Nous avons démontré mathématiquement que les distributions engendrées par la formule [1] correspondent (Lafouge et Lainé-Cruzel, 1997) (Lafouge et Guinet, 1999) (Lafouge 2001) à des lois de circulation stationnaires classiques (Poisson, géométrique, binomiale négative) que nous avons vues précédemment, lorsque l'on passe à la limite dans les conditions décrites ci-dessus. Pour ce faire, il a fallu faire des hypothèses sur la loi de la distribution de contenu. Nous avons montré que si la forme de la distribution de contenu est de type Poisson, géométrique ou binomiale négative alors la distribution d'usage a la même forme. Il a été nécessaire de faire des hypothèse sur la série des lois de probabilité ($P_j(p_o)$) : deux hypothèses ont été formulées, elles conduisent au même résultat, ce qui donne plus de solidité au modèle.

5. Conclusion

Cet article qui s'appuie sur les concepts classiques de la science de l'information (Lecoadic, 1994) nous permet d'avoir un regard neuf sur ses lois . Bradford et Lotka sont réunis par le nouveau concept de dualité . La singularité de la loi de Zipf nous fait douter pour l'instant, non pas de sa réalité mais de sa pertinence comme étant une loi relevant de l'infométrie . Nous avons utilisé des outils classiques probabilistes pour définir le modèle d'un champ infométrique. Nous

pensons qu'il ne faut pas nous arrêter là . La définition d'un CIP avec le concept de dualité et l'introduction de la distribution de contenu doit nous inciter à utiliser d'autres outils mathématiques en infométrie (géométrique, algébrique) de la même façon que nous avons procédé en ayant une nouvelle approche et cette fois-ci en oubliant la formulation probabiliste de ces lois.

Références bibliographiques

Barbut M.,1990
Distribution de type paretien et interprétation des inégalités. Marc Barbut, p.15-35.
Dans : La modélisation confluent des Sciences
Edition du CNRS 1990.

Bradford S. C., 1934
Sources of information on specific subjects. S. C. Bradford
Engineering p. 85-86, 26 janvier 1934.

Burrel Q. L. 1998
Predictive aspects of some bibliometric process . Q. L. Burrel. Informetrics 87/88 : Select proceedings of the first international conference on bibliometric and theorical aspects of information retrieval. Elsevier, Amsterdam 1998.

Burrel Q. L. 2001
« Ambiguity » and Scientometric Measurement : a Disenting view.
Journal of the American Society for Information Science and Technology, 52(12) p.1075-1080, 2001.

Egghe, L., 1988
On the classification of the classical bibliometric laws.
Journal of Documentation, Vol 44, N°1, p.53-62 ,1988.

Egghe, L., 1990
The duality of informetrics systems with applications to the empirical law
Journal of Information Science, Vol 16, p17-27 1990.

Egghe, L, 2000
New informetric aspects of the Internet: some reflexions – many problems.
Journal of Information Science, 26 (5), p. 329-335, 2000.

Fondin H., 2001
La Science de l'information : posture épistémologique et spécificité disciplinaire.
Documentaliste Sciences de l'information Vol 38 , N°2 p.112-122, 2001.

Haitum, S. D., 1982
Stationary Scientometric Distribution.
Scientometrics N°4, Part I p. 5-25, Part II p. 89-104, Part III p.181-194. 1982.

Lada A. Adamic, 2000
Dernière modification 10-04-2000
Zipf, Power-Laws, and Pareto – a ranking tutorial
Internet Ecologies Area
Xerox Palo Alto Research Center
Palo Alto, CA 94304
http://www.parc.xerox.com/istl/groups/iea/papers/ranking/ranking.html

Lafouge, T., 2001
A mathematical Model of Documents circulation: Use Distribution, Utility Distribution, Content Distribution : example of scientific Articles Circulation in Journals
Proceedings of the eigth conference of the international Society for Scientometrics and Informetrics. Sydney Australia 2001, p.327-337.

Lafouge T, Michel C., 2001
Links between information construction and information gain. Entropy and bibliometric distributions.
Journal of information Science, 27(1) p 39-49, 2001.

Lafouge, T., Guinet E., 1999
A new explanation of the negative binomial law and the Poisson law with regard to library circulation data.
Journal of Information Science, 25(1), p.89-93, 1999.

Lafouge, T., 1998
Mathématiques du document et de l'information, Bibliométrie distributionnelle.
Mémoire d'habilitation.
http://193.51.109.173/memoires/ThierryLafouge_ext.pdf

Lafouge, T, Lainé-Cruzel S, 1997
A new explanation of the geometric law in the case of library circulation data.
Information Processing and Management, Vol 33, No 4, p. 523-527, 1997.

Lafouge, T., 1989
Etude comparée des différents modèles de circulation dans une bibliothèque. Revue Française de Bibliométrie, N°4 , p. 179-190, 1989.

Leydesdorff L., 2001
The challenge of Scientometric
The Development, Measurement, and Self-Organisation of Scientific Communications
2001
Universel Publishers / uPUBLISH.com
http://www.upublish.com/books/leydesdorff-sci.htm

Lecoadic Y. F., 1994
La Science de l'Information.
Paris PUF, 1994 (Que sais je)

Loose M., 2001
Term Dependance : A Basis for Luhn and Zipf Models
Journal of the American Society for Information Science and Technology 52(12) p1019-1025
2001.

Lotka A. J., 1960
The frequency distribution of scientific productivity
Journal of the Washington Academy of Sciences, 16 p317-323, 1960.

Price D. S., 1976
A general theory of bibliometric and other cumulative advantage process .
Journal of the American Society for Information Science., Vol 27, N°5, 1976, p. 292-306.

A. Reyni A., 1966
Calcul des probabilités
Editions Jacques Gabay 1992 (Réimpression Dunod Paris 1966)

Rostaing H., 1996
La bibliométrie et ses techniques chapitre 2
Co-édition Sciences de la Société-CRRM, 1996.

J. M Salaün J.M, Lafouge, C. Boukacem, 2001
Trading in ideas, articles and journals : a document case study
Scientometrics, Vol 47, N°3, p. 561-588, 2001.

Sengupta L.n., 1992
Bibliometrics, Scientometrics and librametrics: an overview..
Libri, Vol 42, N° 2 p. 75-98, 1992.

Zipf, G. K., 1935
The form and behavior of words
The psycho-biology of language Boston : Houghton, 1935 p. 20-48.