Entropie et Distributions bibliométriques

Thierry Lafouge
Laboratoire Recodoc
Université Claude Bernard Lyon 1, Bat 721
43 Bd du 11 novembre 1918
69622 Villeurbanne Cedex, France.
lafouge@enssib.fr

Christine Michel
Laboratoire Recodoc
Université Claude Bernard Lyon 1, Bat 721
43 Bd du 11 novembre 1918
69622 Villeurbanne Cedex, France.
michel@dist.univ-lyon1.fr

Mots clés : distribution bibliométrique / entropie / loi du moindre effort / théorie de l'information

La théorie de l'information s'est particulièrement développée autour des années 50, en fait à partir du moment où C. Shannon a formalisé la circulation de l'information dans un modèle mathématique. On parle alors d'entropie ou de théorie statistique de l'information. Ce dernier utilisera celle-ci comme outil de traitement du signal (Shannon 1975). Nous allons rappeler brièvement les notions les plus importantes de cette théorie. Nous expliciterons ensuite le contexte d'utilisation de celle-ci en bibliométrie, qui est celui des distributions qui étudient les régularités statistiques observées dans le domaine de la production et de l'usage de l'information. Tous ces travaux ont confirmé l'existence de grandes similitudes entre les différentes distributions étudiées dans le domaine de l'information et donc l'existence de régularités et de rapports mesurables, qui autorisent de ce fait la prévision et le concept de lois.
Cet article prolonge des études qui ont montré qu'il existe un lien entre distribution et entropie plus particulièrement le lien qui existe entre le principe du "moindre effort" et la forme analytique d'une distribution Zipfienne.

1. Rappel sur l'entropie

Soit une source d'information, producteur de n événements aléatoires, de probabilités respectives $p_1, p_2 \ldots p_n$

$(\sum_{i=1}^{n} p_i = 1)$

on appelle entropie d'une telle source la fonction *H* suivante (Caumel 1988) :

$$H(p_1, p_2 \ldots p_n) = -\sum_{i=1}^{n} p_i Log_2(i)$$

Tout comme les systèmes étudiés en science physique, plus l'entropie de la source est élevée et moins le système est ordonné (ici prévisible). Shannon suppose que plus un système est ordonné et moins il ne produit d'information. Ainsi l'entropie H est maximum si tous les évènements sont équiprobables c'est à dire si on a : $\forall i \; p_i = \frac{1}{n}$, dans ce cas $H = Log_2(n)$

La fonction logarithme de base deux qui est couramment utilisée est cohérente avec le codage électronique de l'information par le bit (le bit est défini comme étant l'entropie maximum de la source binaire aléatoire).

Il est possible d'étendre et de généraliser la définition de l'entropie pour une loi de probabilité continue. Dans ce cas, on ne parlera plus de distribution mais de fonction de densité (notée ν) d'un phénomène aléatoire. Et on définira son entropie par la fonction :

$$\boxed{H(v) = -\int v(t) Log(v(t)) dt}$$

En bibliométrie les événements classiquement étudiés sont la production d'articles, la production de mots clés, ou les emprunts d'ouvrages; les sources respectivement prises en compte seront alors les auteurs, les références bibliographiques et les livres. Ces évènements sont remarquables car ils sont caractérisés par des régularités statistiques. Il est alors intéressant d'observer comment varient les entropies, c'est à dire comment varie la quantité d'information, de ces différentes sources en fonction du processus aléatoire qui les gouverne.

2. Rappel sur les distributions Bibliométriques

Les trois distributions bibliométriques classiques sont la loi de Lotka, relative à la production d'articles, la loi de Zipf, relative à la féquence d'apparition des mots dans un texte et celle de Bradford relative à la dispersion des articles dans des périodiques.

En fonction des phénomènes étudiés, ces lois vont s'énoncer soit de manière fréquentielle, soit de manière ordonnée.

Fréquence (Lotka):

L'approche fréquentielle est la plus ancienne (Lotka 1926). La probabilité d'apparition d'un événement est calculée en fonction de sa fréquence d'apparition. Un exemple de loi de distribution fréquentielle est la loi de Lotka ou l'analyse porte sur la production d'articles scientifiques par des chercheurs. Lotka propose d'écrire la distribution n_i du nombres de scientifiques qui ont écrit i articles par :

$$n_i = \frac{n_1}{i^2} \qquad i = 1,2.....i_{\max}$$

où $i_{\max}$ est le nombre maximum d'articles produits par un chercheur.

Cette loi est généralisée dans la formule : $n_i = \frac{K}{i^a}$ où K et α sont des constantes dépendants du domaine étudié.

Rang (Zipf):

L'énonciation d'une loi selon le rang suppose que l'on ait au préalable ordonné la source d'information selon sa production. Ces distributions par rang sont utilisées lorsque l'ordonnancement de la production de la source est inévitable pour marquer l'apparition d'une régularité. L'exemple le plus caractéristique d'une distribution par rang est celui de la loi de Zipf. Ce dernier observe la fréquence d'apparition des termes dans des textes anglais. Il va comptabiliser le nombre de termes qui ont une fréquence d'apparition identique. En ordonnant ces nombres d'une manière décroissante, il observa qu'il existe une relation inversement proportionnelle entre le rang de présentation d'un terme et sa fréquence d'apparition. Zipf traduit cette régularité par l'équation suivante :

$g(r) = \frac{K}{r^q}$ où g(r) désigne la fréquence de la forme de rang r.

De nombreux travaux ont montré les équivalences entre les distributions par rang et les distributions fréquentielles (Egghe 1988). Le choix de l'une ou l'autre des présentations dépend de l'analyse que l'on souhaite mener. Dans l'étude des distributions de mots clefs, on choisit de présenter la distribution par rang car elle est la plus significative (Quoniam 1992). Dans une publication récente (Lhen 1995), faite en collaboration avec le CRRM, nous avons montré l'intérêt de la théorie de l'entropie généralisée (Reyni 1960) pour manipuler les distributions de ce type.

Le cas continu (Pareto) :

L'analogue de la distribution de Lotka pour le cas continu est la distribution de Pareto qui s'écrit sous la forme :

$$v(t) = \frac{a}{to}\left(\frac{to}{t}\right)^{a+1} \quad t \geq to \qquad a \geq 0$$

Haitun (Haitum 1982) définit une distribution Zipfienne par la fonction de densité hyperbolique suivante $v(t)=\frac{C}{t^{1+a}}$ où t appartient à l'intervalle $[1,\infty)$ et où α et C sont des constantes positives.
Si a = 1 nous sommes dans le cas bien connu de la loi de Lotka .
Toutes les propriétés mathématiques de ce type de distributions ont été largement étudiées. SD Haitum oppose ce type de distribution aux distributions gaussiennes.

3. Entropie et distribution

3.1 Problématique

Soit une distribution bibliométrique, qu'on note (F, I) où F représente l'ensemble des formes qui désignerons une source bibliographique identifiée : ensemble d'auteurs, d'articles,…..Le mot forme désigne également dans un fichier téléchargé tout suite de caractères entourée de séparateurs. Une forme peut être constituée de un ou plusieurs mots et désigner des éléments aussi variés que: auteur, source, code, mot clé….
I désigne l'ensemble des items : nombre positif représentant l'occurrence (le nombre d'apparition) de la forme. Toutes ces valeurs constituent la production de la source.
Nous avons vu précédemment que plusieurs représentations d'une distribution sont possibles On suppose que cette source est gouvernée par un processus aléatoire stationnaire qu'on peut observer et qu'elle est caractérisée par une fonction qui traduit l'effort pour produire l'ensemble des différents items. Le problème étudié est alors le suivant: pour une quantité d'effort donnée quelle est la relation entre la distribution aléatoire de la source et la fonction d'effort si on cherche à rendre maximum la quantité d'information (au sens de Shannon) produite par la source . Ces techniques, qui consistent à maximiser l'entropie sont utilisés également pour évaluer des bases de données documentaires (Kantor 1998).

Des études ont montré qu'il existe un lien entre distribution et entropie. Yablonsky, a en particulier démontré le lien qui existe entre le principe du "moindre effort" et la forme analytique d'une distribution Zipfienne.
Yablonsky (Yablonsky 1980) se place dans le cas de l'analyse de la production d'articles des chercheurs. L'effort nécessaire pour produire un article est modélisé par la fonction noté E, $E(t)=kLog(t)$, où k est le coefficient de proportionnalité. t = 1 correspond à "l'état minimal du scientifique" qui a produit un article; cette fonction $E(t)$ appelée fonction d'effort traduit le fait que la production d'un article supplémentaire requière de la part du chercheur moins d'effort. Cette modélisation est connue sous le nom de "loi du moindre effort": la fonction de densité qui répond à la question de la maximisation de l'entropie, sous contrainte d'un effort, est la fonction Zipfienne. Nous allons formuler ce problème sous forme mathématique en utilisant une représentation continue de la source.

Plus précisément on cherche une fonction $v(t)$, t étant défini sur l'intervalle [a, ∞[, telle que :
$v(t)$ ³ 0 sur [a, ∞[$(a \geq 0)$ vérifiant:

(a) $\int_a^{\infty} v(t)dt = 1$ (v fonction de densité)

(b) $\int_a^{\infty} E(t)v(t) = E$ (contrainte d'un effort constant)

et maximisant l'entropie H de v : $H(v) = -\int_a^{\infty} v(t) Log(v(t)) dt$

$v(t)$ est une distribution bibliométrique de nature variée : géométrique, Zipfienne, binomiale négative, exponentielle…
$E(t)$ est la fonction d'effort,
E est une constante positive qui correspond à l'effort moyen,
H est l'entropie.

3.2 Résultat de Yablonsky : le cas Zipfien

Yablonsky (Yablonsky 1980) montre que la maximisation de l'entropie pour la fonction d'effort $E(t) = kLog(t)$ est obtenue pour une densité dont la forme analytique est :

$$v(t) = \frac{C}{t^{1+a}} \quad \text{où} \quad \boldsymbol{a} = \frac{k}{E} \quad \text{et} \quad t \geq 1$$

Le calcul de l'entropie en fonction de α donne : $\boxed{H(\boldsymbol{a}) = -Log(\boldsymbol{a}) + \frac{1}{\boldsymbol{a}} + 1}$

α étant par définition positif il est facile de montrer que l'entropie est une fonction décroissante de α. On retrouve l'interprétation classique de la loi de Lotka, à savoir que plus **a** est grand, et plus le fossé entre le nombre de chercheurs qui produit beaucoup et le nombre de chercheurs qui produit peu est grand (sachant qu'il y a peu de chercheurs qui produisent beaucoup par rapport au nombre de chercheurs qui produisent peu).

3.3 Le cas géométrique

Très souvent la distribution géométrique est utilisée pour quantifier certaines régularités observées en bibliométrie en particulier dans le domaine des usages de documents en bibliothèque par exemple (Lafouge 1997).

Une distribution géométrique s'écrit sous la forme :

$$\boxed{G(q)(i) = (1-q)q^{i-1} \quad i = 1,2....\infty \quad 0 < q < 1}$$

Si on écrit l'équivalent continue de cette distribution on obtient la distribution exponentielle suivante :

$$\boxed{v(t) = pq^t = pe^{t \log q} = pe^{-t \log(\frac{1}{q})}} \quad (1\text{-}q = p)$$

Nous avons choisi de prendre la fonction d'effort linéaire $\boxed{E(t) = k(t-1)}$ et montré que la fonction qui répond au problème variationnel précédant est la distribution exponentielle:

$$\boxed{v(t)=\alpha e^{-\alpha(t-1)} \quad où \quad \alpha=\frac{k}{E} \quad \text{et} \quad t\geq 1}$$

Le cas t =1 comme précédemment correspond à l'état minimal du scientifique qui a produit un article.

Dans ce cas $\frac{1}{\alpha}$ est la moyenne de v. La contrainte (b) correspond donc à fixer l'espérance. Si nous nous plaçons comme précédemment dans le cas de la production scientifique, la fonction linéaire *E(t)* traduit que lorsque l'on applique le principe du maximum de l'entropie avec effort constant, la fonction résultante est une distribution géométrique.

Ces résultats sont démontrés en utilisant des techniques de calcul variationnel.·

Démonstration

Montrons que la fonction :

$$w(t)=\alpha e^{-\alpha(t-1)} \qquad \text{où} \qquad \alpha=\frac{k}{E} \quad \text{et} \quad t\geq 1$$

vérifie les conditions (1) (2) et (3)

$v \geq 0$ sur $[1, \infty[$ (1)

$\int_1^\infty v(t)dt=1$ (2)

$\int_1^\infty k(t-1)v(t)dt=E$ (3)

et maximise la fonction: $H(v)=-\int_1^\infty v(t)Log(v(t))dt$

On montre facilement que la fonction w vérifie les conditions (1) (2) et (3). Montrons que w maximise bien l'entropie. Nous allons montrer que H atteint son minimum pour la fonction w.

Soit *F* la fonction suivante:

$F(t,v)=v\,Log(v)+\lambda v+\alpha(t-1)v$:

où λ est une constante qui a pour valeur : $\lambda=-1-\text{Log}(\alpha)$

On a : $\frac{\partial}{\partial v}F(t,v)=Log(v)+1+\lambda+\alpha(t-1)$

On montre facilement que cette dérivée s'annule pour w.

Pour *t* fixé on a : $\frac{\partial}{\partial v}F(t,w)=0$

Pour t fixé on a : $\frac{\partial^2}{\partial^2 v}F(t,v)=\frac{1}{v}\geq 0$

F étant convexe et s'annulant en w pour toute valeur de t fixé on peut écrire :

$\forall v \quad F(t,v) \geq F(t,w)$

soit : $\forall v \quad v\,Log(v) + \lambda v + \alpha(t-1)v \geq w\,Log(w) + \lambda w + \alpha(t-1)w$

$\int_1^\infty (\nu\,Log(\nu) + \lambda\nu + \alpha(t-1)\nu)dt \geq \int_1^\infty (w\,Log(w) + \lambda w + \alpha(t-1)w)dt$

Soit pour toute fonction ν vérifiant les conditions de normalisation (2) et (3)

$\int_1^\infty \nu\,Log(\nu)dt \geq \int_1^\infty w\,Log(w)dt$ d'où le résultat cherché.

Le calcul de l'entropie en fonction de α donne :

$$\boxed{H(\alpha) = 1 - Log(\alpha)}.$$

Nous avons le même résultat que précédemment pour la variation de H en fonction de α. On remarque d'autre part que l'entropie calculée avec une loi d'effort linéaire est toujours inférieure à une entropie calculée avec une loi d'effort logarithmique et que cette différence varie de façon inversement proportionnelle à α. La dispersion, et donc l'entropie, est plus forte dans le cas Zipfien. Ce résultat justifie à posteriori le choix de l'entropie d'ordre 1 pour caractériser la diversité d'une distribution Zipfienne (Lhen 1995).

Remarque

Nous pouvons démontrer le résultat précédant de Yablonsky en utilisant la même technique avec la fonction F suivante :

$$F(t,v) = vLog(v) + \lambda v + (1+\alpha)vLog(t) \text{ où } \lambda = -Log(\alpha) - 1$$

3.4 Le cas binomial négatif

Une autre distribution est très souvent utilisée (Lafouge 1999) pour modéliser les distributions d'usage, la distribution binomiale négative qui s'écrit sous la forme :

$$Bn(q,r)(1) = (1-q)^r$$

$$Bn(q,\ r)(i) = r(r+1)....(r+i-1).\frac{q^{i-1}.(1-q)^r}{i!} \qquad i = 2....\infty \ ;\ 0 < q < 1$$

Si r =1 on reconnaît l'équation d'une distribution géométrique. Il n'existe pas à notre connaissance comme pour le cas géométrique de loi strictement équivalente dans le cas continu. Nous allons utiliser des techniques de convolution pour construire une nouvelle distribution qu'on appellera ici « pseudo binomiale ».

Rappel

Soient deux variables aléatoires continues indépendantes X_1 et X_2 ayant respectivement pour fonctions de densité F_1 et F_2 définies sur l'intervalle $]-\infty, +\infty[$. On montre que la variable

aléatoire $X_1 + X_2$ a pour densité la fonction F1*F2 qu'on appelle produit de convolution de F1 et F2 :

$$F_1 * F_2(x) = \int_{-\infty}^{x} F_1(y).F_2(x-y)dy$$

On généralise cette définition pour un produit de convolution d'ordre j. Soit Xj une suite finie de variables aléatoires indépendantes de densité F on montre que la variable aléatoire $\sum_j X_j$ a pour fonction de densité la fonction F_j définie par le produit de convolution suivant :

$F_1 = F$

$F_2 = F * F \qquad F_2(x) = \int_{-\infty}^{x} F(y).F(x-y)dy$

$F_j = F_{j-1} * F \qquad j = 2,3....$

Ces techniques de convolution ont été utilisées pour donner une nouvelle interprétation de la loi de Lotka (Egghe 1994).
On sait que dans le cas discret (Calot 1984) la somme de j variables indépendantes de loi géométrique G(q) est une variable binomiale négative Bn(j q) . Aussi nous allons définir la fonction de densité v_j comme étant la convolution de j distributions exponentielles ; celle-ci sera la distribution, dite « pseudo binomiale négative », qui remplacera la distribution binomiale négative.

Soit la distribution exponentielle de densité:

$$v(t) = \alpha e^{-\alpha t} \quad t \geq 0$$

Un calcul simple montre que le produit de convolution d'ordre j de la fonction v s'écrit :

$$v_j = v * v.....v \ : v_j(t) = \alpha^j \frac{t^{j-1}}{(j-1)!} e^{-\alpha t} \qquad t \geq 0$$

Egghe a montré (Egghe 1994) les propriétés de stabilité pour la distribution géométrique en utilisant ce produit de convolution.

Si la distribution originelle est de type exponentielle, nous pouvons interpréter cette distribution vj de différentes manières :

- dans un contexte de pro*duction d'articles*, $v_j(i)$ est la proportion d'auteurs ayant écrit i articles, chaque article ayant exactement j auteurs.
- dans un contexte de *distributions de mots clefs de références bibliographiques*, $v_j(i)$ est la proportion de mots utilisés i fois, chaque référence ayant exactement j mots clés.

Le problème que nous cherchons à résoudre est alors le suivant : quelle est la nature (de type linéaire, logarithmique…) de la distribution d'effort (notée EF) qui est liée à un processus aléatoire de type « pseudo binomial négatif » (cf définition précédente) et qui maximise la quantité

d'information au sens de Shannon si on fixe la quantité d'effort . (noté E). Ce problème se formule mathématiquement de la façon suivante :

Soit la distribution suivante : $v_j(t) = \alpha^j \frac{t^{j-1}}{(j-1)!} e^{-at}$ $t \geq 0$ (j entier strictement positif)

quelle est la nature de la distribution d'effort EF qui vérifie les conditions suivantes :

$\int_0^\infty EF(t) v_j(t) = E$ (contrainte d'un effort constant)

et telle que l'entropie H: $-H(v) = -\int_0^\infty v(t) Log(v(t))dt$ pour toute fonction v vérifiant les conditions :

$v(t) \geq 0$ sur $[0, \infty[$

$\int_0^\infty v(t)dt = 1$

$\int_0^\infty EF(t) v(t)dt = E$

soit maximum pour la fonction $v_j(t)$.

Nous allons voir si la fonction d'effort EF suivante $\boxed{EF(t) = at - (j-1)\log(t)}$ est valide pour résoudre la problématique précédente.

Nous vérifions facilement par récurrence que $\int_0^\infty v_j(t)dt = 1$

Montrons que l'effort $\int_0^\infty EF(t) v_j(t)dt$ est constant.

Posons $K(\alpha, j) = \frac{\alpha^j}{(j-1)!}$

$$E = \int_0^\infty EF(t) v_j(t) = K(\alpha, j)\left(\int_0^\infty \alpha t^j e^{-at} dt - \int_0^\infty (j-1)\log(t) t^{j-1} e^{-at} dt \right)$$

$$\int_0^\infty t^j e^{-at} dt = \left[t^j e^{-at} \right]_0^\infty - \int_0^\infty j t^{j-1} \frac{e^{-at}}{-\alpha} dt$$

or nous avons $\left[\alpha t^j e^{-at} \right]_0^\infty = 0$ et nous avons vu que $\int_0^\infty K(\alpha, j) t^{j-1} e^{-at} dt = 1$

Donc $\int_0^\infty \alpha t^j e^{-at} dt = \frac{j}{\alpha} \frac{1}{K(\alpha, j)}$

Soit $E = j-(j-1).K(\alpha,\ j).\int_0^\infty \log(t)t^{j-1}e^{-at}dt$

Or $\int_0^\infty \log(t)t^{j-1}e^{-at}dt$ est de la forme $\frac{A_j + B_j \log(\alpha)}{\alpha^j}$. Pour faire ce calcul on utilise des techniques liés au calcul intégral.

Par exemple pour j = 2 on a : $\int_0^\infty \log(t).t.e^{-at}dt = \frac{-1+g+Log(\alpha)}{\alpha^2}$

Pour j = 3 on a $\int_0^\infty \log(t).t^2.e^{-at}dt = \frac{-3+2g+2Log(\alpha)}{\alpha^3}$

(γ est la constante d'Euler :0.5772….)

Donc $\boxed{E = j-\frac{A_j + B_j \log(\alpha)}{(j-2)!}}$

E est bien constant en fonction de t.

Montrons à présent que EF maximise bien l'entropie. Nous allons montrer que H atteint son minimum pour la fonction $\nu_j(t)= K(\alpha, j)t^{j-1}e^{-at}$.

Soit F la fonction suivante:

$F(t,v)=v\,Log(v)+\ \lambda v+EF(t)v$:

où λ est une constante qui a pour valeur : $\lambda = -Log\big(K(a, j)\big)-1$

On a : $\frac{\partial}{\partial v}F(t,v)= Log(v)+1+\ \lambda+EF(t)$

Posons $\nu_j(t)= K(\alpha, j)t^{j-1}e^{-at}$

$\frac{\partial}{\partial v}F(t,\nu_j)= Log\big(K(\alpha, j)\big)+(j-1)log(t)-\alpha t+1+\ \lambda+\alpha t-(j-1)log(t)$

$\frac{\partial}{\partial v}F(t,\nu_j)= Log\big(K(\alpha, j)\big)+1+\ \lambda$

Si on remplace λ par sa valeur alors cette dérivée s'annule pour ν_j .

Pour t fixé on a : $\frac{\partial^2}{\partial^2 v}F(t,v)=\frac{1}{v}\geq 0$

F étant convexe et s'annulant en $\nu_j(t)$ pour toute valeur de t fixé on peut écrire :

$$\forall v \qquad F(t,v)\geq F(t\ ,\nu_j(t))$$

soit : $\forall v \qquad v\,Log(v)+\ \lambda v+\alpha(t-1)v \geq \nu_j(t)Log(\nu_j(t))+\ \lambda\nu_j(t)+\alpha(t-1)\nu_j(t)$

$$\forall v \qquad \int_0^\infty \left(v\,Log(v) + \lambda v + \alpha(t-1)v\right)dt \geq \int_0^\infty \left(\nu_j(t)Log(\nu_j(t)) + \lambda \nu_j(t) + \alpha(t-1)\nu_j(t)\right)dt$$

Soit pour toute fonction v vérifiant les conditions de normalisation (2) et (3)

$\forall v \qquad \int_0^\infty \left(v\,Log(v)\right)dt \geq \int_0^\infty \left(\nu_j(t)Log(\nu_j(t))\right)dt$ d'où le résultat cherché.

Calcul de l'entropie $-H(v) = \int_0^\infty v(t)Log(v(t))\,dt$

$$-H(v_j) = \int_0^\infty K(\alpha, j)t^{j-1}e^{-at}Log\left(K(\alpha, j)t^{j-1}e^{-at}\right)dt$$

$$-H(v_j) = \int_0^\infty K(\alpha, j)t^{j-1}e^{-at}\left(Log(K(\alpha, j)) + (j-1)log(t) - \alpha t\right)dt$$

$$-H(v_j) = \int_0^\infty K(\alpha, j)Log(K(\alpha, j))t^{j-1}e^{-at}dt + \int_0^\infty K(\alpha, j)(j-1)t^{j-1}e^{-at}\log(t)dt - \int_0^\infty \alpha K(\alpha, j)t^{j}e^{-at}dt$$

Or $\int_0^\infty K(\alpha, j)t^{j-1}e^{-at}dt = 1$ et $\int_0^\infty t^{j-1}e^{-at}\log(t)dt$ est de la forme $\frac{A_j + B_j\log(\alpha)}{\alpha^j}$

Donc $-H(v_j) = Log(K(\alpha, j)) + K(\alpha, j)(j-1)\frac{A_j + B_j\log(\alpha)}{\alpha^j} + \alpha K(\alpha, j)\int_0^\infty \frac{j}{\alpha}t^{j-1}e^{-at}dt$

$$-H(v_j) = jLog(\alpha) - \log((j-1)!) + \frac{1}{(j-2)!}\left(A_j + B_j\log(\alpha)\right) + j$$

$$\boxed{-H(v_j) = \left(j + \frac{B_j}{(j-2)!}\right)\log(\alpha) + \frac{A_j}{(j-2)!} + j - \log((j-1)!)}$$

Conclusion

Notre problématique était de trouver un lien entre l'entropie d'une source et sa distribution bibliométrique. Nous somme partis du travail de Yablonsky sur le principe du "moindre effort" et de la forme analytique d'une distribution Zipfienne. Nous avons essayé d'appliquer ce même principe de recherche à d'autres distributions, en particulier les distributions géométriques et distributions binomiales négatives.

Notre problématique est la suivante :

- Si *v(t)* est la forme analytique d'une distribution bibliométrique de nature variée : géométrique, Zipfienne, binomiale négative, exponentielle, *t* étant défini sur l'intervalle [a, ∞[, telle que : *v(t)³ 0* sur [a, ∞[$(a \geq 0)$
- si $E(t)$ est la fonction d'effort de production de la source (E est une constante positive qui correspond à l'effort moyen),
- et si *H* est l'entropie.

Nous cherchons les couples ($E(t)$, $v(t)$) vérifiant :

(a) $\int_a^{\infty} v(t)dt = 1$

(b) $\int_a^{\infty} E(t)v(t) = E$ (contrainte d'un effort constant)

(c) $H(v) = -\int_a^{\infty} v(t)Log(v(t))dt$ est maximum.

Dans le cas d'une **distribution Zipfienne**, la fonction de densité est de la forme $v(t)=\frac{C}{t^{1+a}}$

Yablonsky montre que la problématique est résolue pour le couple :

$$\left(E(t)=kLog(t) \quad , \quad v(t)=\frac{C}{t^{1+a}}\right), \text{ où } \qquad \alpha=\frac{k}{E} \quad \text{et} \quad t\geq 1$$

Le calcul de l'entropie en fonction de α donne : $H(\alpha)=-Log(\alpha)+\frac{1}{\alpha}+1$ (1)

Dans le cas d'une **distribution géométrique**, la fonction de densité est de la forme $v(t)=\alpha e^{-\alpha(t-1)}$.

Nous avons montré que la problématique est résolue pour le couple :

$$\left(E(t)=k(t-1) \quad , \quad v(t)=\alpha e^{-\alpha(t-1)}\right) où \ \alpha=\frac{k}{E} \quad \text{et} \quad t\geq 1$$

Le calcul de l'entropie en fonction de α donne :

$$H(\alpha)=1-Log(\alpha). \text{ (2)}$$

Dans le cas de la **distribution binomiale négative** , la distribution de densité correspondante choisie est $v_j(t)=\alpha^j\frac{t^{j-1}}{(j-1)!}e^{-\alpha t}$ t ≥ 0

Nous avons montré que la problématique est résolue pour le couple :

$$\left(EF(t)=\alpha t-(j-1)\log(t) \quad , \quad v_j(t)=\alpha^j\frac{t^{j-1}}{(j-1)!}e^{-\alpha t}\right) \qquad où \quad t\geq 0$$

Le calcul de l'entropie en fonction de j et α donne :

$$-H(j,\alpha)=\left(j+\frac{B_j}{(j-2)!}\right)\log(\alpha)+\frac{A_j}{(j-2)!}+j-\log\left((j-1)!\right)$$

Nous pouvons observer que dans le cas d'une distribution Zipfienne, la fonction d'effort de production de la source est logarithmique, dans le cas géométrique elle est linéaire, dans le cas d'une distribution "pseudo-binomiale négative" est composée de la somme de deux fonctions : une linéaire (effort constant) et l'autre logarithmique (loi du moindre effort).

Nous pouvons observer que dans les trois cas l'entropie est une fonction décroissante de α. Dans les deux premier cas; le cas Zipfien et géométrique l'entropie décroît d'une manière logarithmique. Dans le cas "pseudo-binomiale négatif", la fonction d'effort conserve sa composante logarithmique mais est beaucoup plus complexe puisqu'elle dépend aussi du paramètre j. Tout l'intérêt est alors à présent de calculer les termes A_j et B_j et ainsi définir complètement le calcul de l'entropie.

Références bibliographiques

EGGHE L.(1988) : *On the classification of the classical bibliometric laws.* Jounal of Documentation, Vol 44(81), 1988, p. 53-62.

EGGHE L. (1994) : *Special features of the author publication relationship and a new explanation of Lotka's law based on convolution theory.* Journal of the American Society for Information Science, Vol 45(6),1994, p. 422-427.

CALOT G. (1984) : Cours de calcul de probabilité. Chapitre 12. Dunod décision 1984, 476 pages.

CAUMEL Y. (1988) : Probabilités, théorie et applications. Chapitre 8. Eyrolles, 1988, 290 pages.

KANTOR Paul B., JUNG Jin Lee : *Testing the Maximum Entropy Principle for information Retrieval.* Journal of the American Society for Information Science, Vol 49 (6), 1998, p 557-556.

LAFOUGE T., LAINE CRUZEL S. (1997) : *A new explication of the geometrical law in the case of library circulation data* . Information Processing and Management, Vol 33 (4), 1997, p. 523-527.

LAFOUGE T.,GUINET E. (1999) : *A new expanation of the negative binomial law and the Poisson law with regard to library circulation data.* Journal of Information Science Vol 25 (1), 1999, p 89-93.

LHEN J., LAFOUGE T., ELSKENS Y., QUONIAM L., DOU H. (1995) : *La « statistique des lois de Zipf*. Revue Française de Bibliométrie N°14, 1995, p. 135-146.

QUONIAM L. (1992) : *Bibliométrie sur les références bibliographiques: méthodologie,* p. 244, 261.La Veille technologique; l'information scientifique, technique et industrielle. Dunod, 1992, 436 pages.

SHANNON C.,WEAVER W..(1975) : Théorie mathématique de la communication. Bibliothèque du CEPL, 1975, 188 pages.

YABLONSKY A.L (1980) : *On fundamental regularities of the distribution of scientific productivity.* Scientometrics, Vol 2,(1), 1980, p. 3-34.