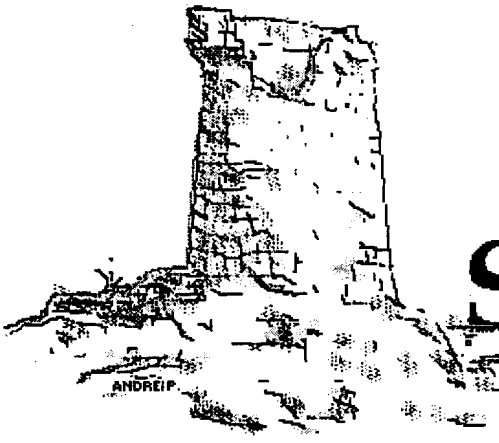
ILE ROUSSE

30 Mai - 2 Juin 1995

LES SYSTEMES D'INFORMATION ELABOREE

Bibliométrie - Linguistique - Information Stratégique - Veille Technologique
Intelligence économique

*



S.F.B.A.

PROGRAMME

*

Journées d'Etudes organisées par la Société Française de Bibliométrie Appliquée avec le concours du Ministère de l'Enseignement Supérieur et de la Recherche (DIST), du Secrétariat Général de la Défense Nationale (SGDN) et de la Direction de la Recherche et de la Technologie (DGA).

Sous le haut patronage de :

-la Municipalité d'Ile Rousse,
-l'Université de Corse,
-l'Université d'Aix-Marseille,
-l'Université de Marne-La-Vallée,
-le Centre de Documentation de l'Armement (CEDOCAR),
-le Centre d'Etudes Scientifiques de Défense (CESD).

COMITE SCIENTIFIQUE DU PROGRAMME

J.F. COPPOLANI (SGDN)
J.L. DALLEMAGNE (U.M.L.V.)
H. DOU (CRRM)
B. DOUSSET (IRIT)
J.E. DUBOIS (PARIS VII)
C. DUTHEUIL (SYNTHELABO)
A. GIRARD (IFP)
F. JAKOBIAK (EXISTRAT)
P. LACOSTE (CESD)
D. LAURENT (U.M.L.V.)
P. ANDREI (U.M.L.V.)
Y. LE COADIC (CNAM)
F. LEPINE (FED)
C. LONGEVIALLE (MIN. de l'INTERIEUR)
F. MARCOTORCHINO (IBM)
B. MARX (INPI)
C. PAOLI (CESD)
W. TURNER (CERESI)
E. VALENSI (CEDOCAR)

MARDI 30 Mai 1995

9h *ACCUEIL DES PARTICIPANTS*

9h15 *OUVERTURE DES JOURNEES*
-Pierre PASQUINI, Député, Maire d'Ile Rousse,
-Jean-François BERNARDINI, Président de l'Université de Corse,
-Pierre LACOSTE, Président du CESD,
-Henri DOU, Président de la SFBA.

10h *INTERVENTIONS INAUGURALES*
-Jacques-Emile DUBOIS, Président de CODATA INTERNATIONAL,
-Edouard VALENSI, Directeur du CEDOCAR,
-François JAKOBIAK, Consultant Existrat,
-Jean-François COPPOLANI, SGDN.

SESSIONS I : STRATEGIE ET CONTEXTE

Président : H. DOU

11h15 **-"Exemple d'une initiative de soutien national au profit des PME"** p.8
QUAZOTTI S. (Centre de Recherche Public Henri Tudor - Cellule Veille Technologique)

11h45 **-"Le SIFU, outil de représentation et de diagnostic de l'organisation de l'entreprise"** p.9
BEAUCHENE D. et MAIRE J.L. (Laboratoire de Logiciels pour la Productique)

12h15 **-"Structure et organisation de la veille dans le groupe Hutchinson"** p.11
BOUQUET V. (Hutchinson)

12h45 Flash-Poster: **-"Stratégie de réseaux et intelligence économique"** p.12
MOINET N. (Université de Poitiers)
-"Richesses et limites de l'interrogation multibases d'un serveur" p.13
BRACHET J.P. (IIRIAM)

SESSIONS II : STRATEGIE ET CONTEXTE (Suite)

Président : C. PAOLI

14h45 **-"Le rôle du SCIP en France"** p.14
MARTI Y.M. (SCIP France)

15h15 **-"La mise en place de la veille stratégique et le *génie organisationnel*"** p.15
DUMAS Ph. (Centre de recherche LePont, Toulon)

15h45 Flash-Poster: **-"Travail coopératif assisté par ordinateur et partage d'informations: vers une conception orientée processus"** p.17
CISSE A. et LINK-PEZET J. (Université de Toulouse)

16h15 **-"Création automatique de thesaurus pour un système d'aide à l'expertise"** p.18
DELOISON J.P., TRIGANO Ph. et LAMROUS S. (Université de Technologie de Compiègne)

16h45 **-"Un instrument d'aide au traitement des informations issues du processus de veille stratégique : fondements et validation dans les entreprises"** p.19
CARON M.L. et LESCA H. (Université Grenoble II)

17h15 **-"L'utilisation de réseaux relationnels pour déterminer des scénarii actuels dans le domaine de la communication électronique"** p.21
BOFF L.H. et VARGAS L. (Université fédérale du Rio Grande, Sud Brésil)

17h45 **-"L'appétit vient en lisant: livres et lecteurs dans le département du Var"** p.22
RODA J.C. (Bibliothèque universitaire de Toulon-Var)

MERCREDI 31 Mai 1995

SESSIONS III : METHODES ET OUTILS

Président : Y. LE COADIC

9h -**"Modélisation et manipulation des données stéréotypées. Spécifications pour un langage de reformatage"** p.23
LE MEUR A. (Laboratoire CRAIE, Université Rennes II)

9h30 -**"La statistique des lois de Zipf"** p.24
LHEN J., QUONIAM L., DOU H - LAFOUGE T. (CRRM - ENSIIB)

10h -**"Méthode d'analyse des thèmes et réseaux de la coopération scientifique nationale et internationale sur les céramiques dans la Base SCI"** p.25
ABD-EL-KADER M., MIQUEL J.F. - DORE J.C. (Ecole Centrale de Paris - URA 401 CNRS)

10h30 -**"Mesure relative de l'information utile dans des périodiques scientifiques"** p.26
LAFOUGE T. (Université Lyon I - RECODOC)

11h Flash-Poster: -**"Génération automatique de réseaux en bibliométrique"** p.27
BOUTIN E. et DUMAS P. (Univ. de Toulon),
et QUONIAM L., ROSTAING H. et DOU H. (CRRM)
-**"Analyse de la production scientifique du centre INRA d'Antibes"** p.28
BRANCA-LACOMBE G., COTTE F. et WAJNBERG E. (INRA)

11h30 -**"Diffusion de l'information et hypermédias sur Internet (ex: Résumé des communications des journées d'île Rousse)"** p.29
DOU J.M.(CRRM) et BALDIT P.(Pétronaphte)

12h -**"TETRAWEB : un serveur W3 pour TETRALOGIE"** p.30
DOUSSET B. et DKAKI T. (IRIT UPS)

Flash-Poster: -**"Les recherches sur les pathologies infectueuses et parasitaires à l'INSERM de 1990 à 1994"**
BARBEROUSSE N., BOUVIALA C. , DE BOISCHEVALIER B. et DE METZ S. (INSERM)

SESSIONS IV : ANALYSE DES TEXTES

Président : J. GUYAUX

15h -***"L'analyse de Dépêches de Presse,*** **une application industrielle d'analyse de données textuelles"** p.31
COUPET P, GOUTTAS C., HUOT C. et WARNESSON I. (IBM)

15h30 -**"Contribution de l'analyse linguistique à un système de veille technologique dans le domaine aéronautique"** p.32
ANDREI P., SIMONI J.L., FLUHR C., KRUMEICH C. et PAOLI C. (CESD-CEDOCAR-CEA).

16h -**"SERAPHIN, un système d'extraction automatique d'énoncés importants"** p.33
BERRI J., MALRIEU D., MINEL J.L. - LEROUX D. (EDF/DER - CAMS-CNRS)

16h30 Flash-Poster: -**"DATAREAD, un outil de transcodage automatique associé à une validation d'experts"** p.34
LA TELA A. et DOU H. (CRRM)
-**"Approche bibliométrique des résumés bibliographiques des Chemical Abstracts"** p.35
FAUCOMPRE P., QUONIAM L.,DOU H. (CRRM) et PARTYKA S. (LAMMI)

17h -**"Analyse de corpus textuels, construction de référentiels et indexation automatique"** p.36
MAUCERI C. (GSI-Erli)

17h30 -**"Une approche linguistico-infométrique au service de la veille scientifique et technologique"**
POLANCO X., GRIVEL L. et ROYAUTE J. (INIST-CNRS) p.37

18h -**"L'analyse linguistique automatique comme point de départ pour la recherche de tendances thématiques dans les publications scientifiques"** p.39
IBEKWE F. et LALLICH G. (CRISTAL-GRESEC, Université Grenoble III)

JEUDI 1er Juin 1995

SESSIONS V : APPLICATIONS

Président : F. MARCOTORCHINO

9h -**"La base de données scientifiques et bibliométriques de l'INSERM : un outil performant"** p.40
BARBEROUSSE N., BOUVIALA C., de METZ S. et TRECOURT C. (INSERM)

9h30 -**"Application du logiciel TEWAT à l'analyse du développement pharmaceutique pour le domaine des maladies neurodégénératives"** p.42
COUPET P., HUOT C. - GRANDJEAN N. (IBM - Inframonitor)

10h -**"Les indicateurs de la recherche médicale en Tunisie à travers leur cartographie"** p.43
TURCHI F., MAHMOUD S.(I.S.D. Tunisie) et HASSANALY P.

10h30 -**"La démarche de veille stratégique du CETIM"** p.45
DUMAS S. et DEVALAN P. (CETIM)

11h15 -**"Constitution d'un dossier de Veille Scientifique et Technologique sur la *Lutte biologique grâce aux insectes auxiliaires*"** p.46
BRANCA-LACOMBE G., BIJAOUI A. et COTTE F. (INRA)

11h45 -**"La veille technologique appliquée à la production de biocarburants liquides issus de biomasses"** p.47
CORONINI R. et de LOOZE M.A. (UPMF-IREPD et INRA)

SESSION V (Suite)

Président :F. LONGEVIALLE

14h15 -**"Les réseaux de compétences : application à l'élaboration d'un groupe de recherche en physique fondamentale"** p.49
SURAUD M.G. - QUONIAM L. et ROSTAING H. (LERASS,IUT - CRRM)

14h45 -**"Une étude du positionnement d'un département du CNRS dans le contexte de la littérature scientifique"** p.51
RING B. (CRN/IN2P3), VERGNES G. (IPN/IN2P3), REMY D. (CSD/Orsay)

15h15 -**"L'expérience de l'ARIST PACA-Corse en matière de veille technologique"** p.52
ARMANET F.

15h45 Flash-Poster: -**"L'information au service de la décision stratégique du laboratoire. Application à une technique pointue : la microscopie-IRTF"** p.53
Kister J., Mouillet V., Dou H., Meunier C. et Quoniam L. (GOAE -CRRM)

16h15 -**"Classification Internationale des Brevets : Etude de l'Innovation et Cohérence d'Affectation"**
DOS SANTOS R. et HASSANALY P. (CRRM) p.54

16h45 -**"La biotechnologie des corps gras : application des méthodes bibliométriques à l'analyse de l'information brevet"** p.55
CHARBONNIER G., GRAILLE J., PINA M., MONTET D. et MEUNIER J. (CIRAD)

17h15 -**"La montée de la violence dans la société allemande, étude dynamique à partir de la modélisation des flux d'information"** p.56
de SAINT LEGER M., TURNER W. - RENNER I. (CERESI/CNRS - IZ-Sozialwissenschaften)

17h45 -**"INTERNET au Japon"** p.57
EMERIC J.L., HAON H.(CEDOCAR)

18h15 Flash-Poster: -**"Le SCI journal citation reports ou comment positionner une revue dans l'environnement ISI"** p.59
MAGRI M.H., SOLARI A. et RERAT K. (INRA)

VENDREDI 2 JUIN

SESSION VI : SYSTEMES D'INFORMATION

Président : C. DUTHEUIL

9h -**"L'information élaborée pour intégrer la formation technique au tissu économique"** p.60
LIAUTARD D., RETOURNA C. et GIRAUD E. (CRDP et CRRM)

9h30 -**"Personnalisation des réponses dans un Système de Recherche d'Information"** p.62
RULQUIN V., DAVID A. et THIERY O. (CRIN)

10h -**"Système d'information stratégique pour le management : concepts et modèles"** p.63
NDIAYE S. et LINK-PEZET J. (LIS)

10h30 -**"Prospective des métiers et stratégie des groupes leaders dans le domaine des "info-routes", structure de l'information, méthode et outil"** p.65
VERNEUIL P. (CESD)

11h Flash-Poster: -**"Le pôle de recherche : structure d'organisation de la science en région"** p.66
BIZARD J.-B. (CERESI/CNRS)
-**"Réseau de Veille sur les Technologies d'Information (RVTI)"** p.67
DORE D. (Conseil de gestion)

11h30 -**"Représentations réactives et/ou représentations interactives : *Définition du coût cognitif*"** p.68
FAVIER N. (Laboratoire de recherche "Le Pont", Toulon)

12h -**"Acteurs, systèmes d'informations: cercle vertueux et cercle vicieux de la génération et de la circulation de l'information"**
FAYARD P. et JACQUES-GUSTAVE P. (Univ. de Poitiers; LABCIS-INTELCO)

12h30 -**"Génération de systèmes de Recherche d'information sur les autoroutes de l'information"** p.70
DUCLOY J. (CRIN-CNRS & INRIA Lorraine)

13h Flash-Poster: -**"Application de la méthode de choix multicritère ELECTRE à la Veille Technologique"** p.71
Ginting R., Dou H., Hassanaly P. (Institut Technologie d'Indonésie, CRRM)
-**"L'information en chimie. De la veille scientifique et technique à l'Intelligence économique. Les Alkylpolyglucosides."** p.72
Baretta A., Loigerot J., Dos Santos R., Dou H. (ENSSPICAM, CRRM)

TABLE RONDE

15h - **Débat : "L'information élaborée et les autoroutes de l'information".**
J. GUYAUX, F. JAKOBIAK, C. PAOLI, M. COPPOLANI, H. DOU, P. LACOSTE, E. VALENSI

Des démonstrations seront organisées pendant la durée des Journées par :

- le CEDOCAR
- le CESD
- le CRRM
- IBM
- l'IRIT
- la société MANOS

Les systèmes d'information élaborée
(Ile Rousse, 30 mai – 2 juin 1995)
Société française de bibliométrie appliquée

La « statistique » des lois de Zipf

*LHEN J., **LAFOUGE T.,***ELSKENS Y., *QUONIAM L., *DOU H.

*C.R.R.M. Centre Scientifique de St Jérôme F-13397 Marseille CEDEX 20
**ENSIIB, 78 rue du 11 Nov. 1918. F-69623 Villeurbanne CEDEX
***Equipe turbulence plasma, U.R.A. 773, C.N.R.S- Université de Provence, I.M.T. Château-Gombert F-13451 Marseille CEDEX 20

Mots clés: distributions bibliométriques, théorie de l'information, indicateurs

Résumé

Les lois de Zipf, Lotka et Bradford ont fait l'objet de nombreux travaux dans le domaine de la bibliométrie. La plupart de ces travaux ont porté sur une description analytique de ces courbes. D'autres travaux font état de l'utilisation de la théorie de l'information de Shannon.
Notre but ici est de décrire un certain nombre de ces indicateurs pour dégager une « statistique » opérationnelle capable de caractériser des lois de Zipf expérimentales, comme la moyenne arithmétique et la variance caractérisent des lois présentant une tendance centrale.
Les écoles de pensée française en bibliométrie sont les seules, à notre connaissance, à présenter une interprétation de ces lois en trois parties (Trivial, Information, Bruit). Cette façon d'interpréter nous permet de valoriser les signaux faibles quand l'innovation est recherchée ou au contraire d'éliminer un bruit (au sens statistique) quand des cartographies générales sont recherchées. Dans la littérature, nous trouvons des découpages en deux zones (le coeur et la dispersion) ou en plusieurs zones (dont le nombre fluctue à chaque expérimentation) conformément aux travaux de Lotka et Bradford.
Notre but sera donc aussi la détermination des seuils entre ces différentes zones, seuils qui pour l'instant sont déterminés de façon purement empirique, fondée sur l'expérience et l'expérimentation.
Tout au long de notre exposé, des exemples seront étudiés et nous conclurons par des perspectives de prolongement de ce travail.

Abstract

In our bibliometry field, there are many works on the laws of Zipf, Lotka and Bradford. Most of them explain analytic description of these curves. Other works show the use of the Shannon information theory.
Our aim in this paper will be to determine some criteria for the development of a piece of statistics able to characterize experimental Zipf laws, as the arithmetic average and variance characterize laws with a central tendency.
The french school of thinking, to our knowledge, seems to be the only one showing an interpretation in three points (trite, information, noise) of these curves. This way of interpretation makes low signals stand out whenever we look for breakthroughs and on the contrary suppresses the noise when general mappings are required. Among scientific writing we found some cuts of the curve in two zones (dispersion and core) or in several (the number of zones depends of the experimental laws) according to the work of Bradford and Lotka. So our aim will be to define thresholds between these different zones, thresholds which are currently defined by a rule of thumb, based on experimentation.
All along our lecture, some examples will be studied and we will conclude on opening on the subject.

1. Introduction

Notre discipline n'est pas la seule à posséder des lois « Zipfiennes »: la biologie, la géologie[1], la physique[2] et l'économie[3] en comptent de nombreux exemples. Ces lois, caractéristiques de diversité plutôt que de dispersion, ont souvent été liées à la théorie de l'entropie. Notre but va être ici de préciser des formulations déjà existantes dans la littérature, d'apporter des compléments utiles pour interpréter et comparer des distributions « Zipfiennes », et enfin de les appliquer à des distributions expérimentales. Les formulations que nous passerons en revue ne sont applicables que sur des séries expérimentales longues (Nous estimons que l'ordre de grandeur du nombre de références minimum à prendre en compte est de 500 à 1000 références).

2. Entropie et Diversité d'ordre a

Renyi propose en 1961 une approche unifiée pour décrire ces lois en définissant l'entropie d'ordre a :

$$\text{pour } a \neq 1 : H_a = \frac{1}{1-a} \log \sum_{i=1}^{n} (p_i)^a \quad \text{(logarithme népérien)}$$

$$\text{pour } a = 1 : H_1 = \lim_{a \to 1} H_a = -\sum_{i=1}^{n} p_i \log p_i$$

où n représente le nombre de formes distinctes dans un champ sur l'ensemble du corpus et p_i représente la probabilité ou fréquence d'apparition de la forme i dans le corpus. Soit $p_i = N_i / N$ avec N_i nombre d'occurrences de la forme i, i = 1,... n et

$N = \sum_i N_i$ nombre total d'occurrence de toutes les formes.

Par convention, nous ordonnerons les probabilitées : $1 \geq p_1 \geq ... \geq p_n \geq 0$.

La dénomination « forme » représente les formes graphiques constituées de caractères alphanumériques contenus entre deux séparateurs[4]. Par exemple les formes peuvent être des mots-clés multi- ou uni-termes, mais aussi des noms d'auteurs avec ou sans prénoms, ou des codes.

Hill[5, 6] propose une approche qui va définir une mesure de la composition en forme, une mesure de diversité. Les diversités d'ordre a selon Hill (cette nomenclature est due à Daget[6] 1980) sont:

$$D_a = \exp H_a$$

Selon Hill ces diversités regroupent plusieurs concepts qui sont complémentaires. C'est ainsi que nous les comprendrons. Notons que H_a est toujours positive et $D_a > 1$; H_a et D_a sont des fonctions décroissante de a.

3. Les différents ordres de l'entropie de Renyi et de la diversité de Hill

3.1 - Ordre 0:

A l'ordre 0 les expressions générales deviennent:

$$H_0 = \log n \qquad \text{et} \qquad D_0 = \exp H_0 = n$$

dit autrement la diversité à l'ordre 0 correspond au nombre de formes distinctes présentes dans un champ sur l'ensemble des références. Cette diversité rend donc compte de l'étalement de la courbe principalement dû aux basses fréquences. Dans l'interprétation classique d'une loi « Zipfienne », nous disons: « peu de formes à fortes fréquences, beaucoup de formes à faibles fréquences ». Le nombre de formes est donc plus un indicateur qualifiant les faibles fréquences, la partie étalée de la courbe, la diversité à faible fréquence, le coté bruit de la courbe.
Procaccia[2] appelle H_0 (qu'il note D_0) la dimension fractale, elle est liée aussi à la dimension de Hausdorff-Besicovitch[7, 8] qui donne une représentation fractale[8] entre autres. Certains auteurs ont déjà signalé la dimension fractale des données bibliométriques, par exemple A.F.J. Van Raan[9] et R. Fairthorne[17]. Quant à Hill, il note N_0 ce que nous appelons ici D_0.

3.2 - Ordre 1:

A l'ordre 1 les équations générales deviennent:

$$H_1 = -\sum_{i=1}^{n} p_i \log p_i \qquad \text{et} \qquad D_1 = \exp H_1$$

Cette formule correspond à celle de l'entropie thermodynamique ou l'information de Shannon[10].
La diversité (ou les diversités) est considérée pour Hill comme une « variance » des formes. Elle peut être définie comme la mesure de la composition en forme.
En physique[2] H_1 est définie comme la dimension d'information.

3.3 - Ordre 2:

A l'ordre 2 l'entropie de Renyi est associée à une grandeur nommée concentration par Hill et Daget.

$$C = \sum_{i=1}^{n} \left(\frac{n_i}{N}\right)^2 = \sum_{i=1}^{n} (p_i)^2$$

A l'ordre 2 les équations générales deviennent:

$$H_2 = -\log \sum_{i=1}^{n} (p_i)^2 = -\log (C) \qquad \text{et} \qquad D_2 = \frac{1}{C}$$

Procaccia[2] donne à H_2 le nom de dimension de corrélation. De plus Hill montre que D_2 est l'inverse de l'indice de Simpson ou concentration C[16].

3.4 - Ordre -∞:

A cet ordre, nous mettons en évidence la forme qui a la probabilité d'apparition minimum. Information triviale en soi, mais qui est bien un indicateur de la diversité. Par analogie avec la nomenclature de Daget et de Hill nous pouvons dire que $D_{-\infty}$ identifie les occurrences (le nombre total d'occurrence). Il est l'inverse de la probabilité la plus faible.

$$D_{-\infty} = 1 / (p_i)_{min}$$

ou $(p_i)_{min} = (\inf N_i) / N$ avec le minimum théorique $(p_i)_{min} = 1 / N$

par construction $(p_i)_{min} = p_n$

La grandeur $(p_i)_{min}$ est l'inverse de la contribution spécifique la plus faible, laquelle ne peut être inférieure à 1/N, mais peut être plus élevée si la fréquence absolue minimale est supérieure à 1. Il est clair que $(p_i)_{min}$ soit inférieur ou égal à 1/n (avec égalité dans le cas de diversité maximum). Nous avons donc une position extrême pour l'asymptote supérieure dans le graphe de D_a en fonction de a que nous pouvons interpréter en disant qu'elle traduit le nombre maximal « d'espèces possibles » avec N présences.

3.5 - Ordre +∞:

A cet ordre, c'est la fréquence maximum d'apparition qui est révélée. Information triviale en soi, mais qui est aussi un indicateur de la diversité. Encore une fois par analogie avec Daget et Hill nous pouvons dire que $D_{+\infty}$ identifie la forme dominante ou les formes codominantes. Il représente l'inverse de la probabilité la plus forte.

$$D_{+\infty} = 1 / (p_i)_{max}$$

avec $(p_i)_{max} = (\sup N_i) / N$ et par construction: $\sup N_i = N_1$.

La grandeur $(p_i)_{max}$ (supérieure ou égale à 1/n) est l'inverse de la contribution spécifique la plus forte. Nous avons donc une position de l'asymptote inférieure dans le graphe de D_a en fonction de a que nous pouvons interpréter en disant qu'elle traduit le nombre minimal « d'espèces possibles », pour Daget.

Pour Hill les diversités ont un intérêt théorique car elles peuvent être reliées à la stabilité, la maturité, la productivité et l'évolution temporelle d'un système. Elles sont une mesure de paramètre dont les valeurs observées peuvent être expliquées par diverses théories selon la discipline: Théorie cybernétique, information de Shannon, entropie thermodynamique, théorie du chaos.

4. Les rapports entre les différents termes de l'entropie d'ordre a

Nous pouvons ajouter comme Legendre[1] un autre critère pour définir la forme de la « Zipfienne »: la régularité R représentant l'information relative contenue dans la forme de la distribution. Elle a été introduite pour la première fois par Pielou 1966 et se définit en terme de Hill et Daget (en fait Hill note la régularité J) par :

$$R = \left|\frac{H_1}{H_0}\right|$$

Remarque : La valeur absolue n'est pas nécessaire, mais la régularité a été introduite pour la première fois comme cela nous l'avons donc conservée.

La régularité est le rapport des entropies correspondant aux nombres de diversité d'ordre 0 et 1. R joue le même rôle qu'un coefficient de variation, ce qui permet de comparer deux distributions à l'aide d'un coefficient « normalisé ». Il est clair que $0 \le R \le 1$. De plus, R = 1 si et seulement si toutes les probabilités sont égales : $p_i = 1/n$ (cas de diversité maximale, pour laquelle $H_a = \log n$ et $D_a = n$, pour tout a).

5. Indicateurs de tendance centrale

L'étude des fréquences des formes ne donne pas d'indication sur les relations entre celles-ci. A priori, nous ne connaissons pas la structure de l'espace des formes indicées par i. Nous cherchons cependant à identifier, approximativement, les formes qui appartiennent aux fortes fréquences, aux faibles fréquences ou à la partie centrale de la distribution des fréquences. Nous pouvons alors définir à partir des entropies de Renyi la fréquence moyenne $\overline{p}$ et la variance des fréquences V_p qui sont deux indicateurs classiques.

$$\overline{p} = \frac{1}{n}\sum_{i=1}^{n} p_i = \frac{1}{D_0}$$

$$V_p = \frac{1}{n}\sum_{i=1}^{n} p_i^2 - \overline{p}^2 = \frac{D_0 - D_2}{D_0^2 D_2}$$

Ces deux indicateurs ne caractérisent cependant pas directement les formes, mais seulement les fréquences. A ce titre, ils ne nous intéressent pas directement dans la suite de l'exposé. Mais ce n'est pas une voie à négliger définitivement.

Pour définir une répartition des formes, nous associons à chacun un rang r_i, par ordre décroissant des fréquences. Le choix le plus simple est : $r_i = i$. Lorsque plusieurs formes ont la même fréquence, nous leur attriburons des rangs successifs, dans un ordre arbitraire. A l'aide de ce rang, nous pouvons calculer une valeur moyenne sur les formes. S'il y a peu de sens à calculer une moyenne arithmétique des fréquences ou des occurrences, il est plus naturel de calculer une valeur moyenne « géométrique » (par opposition à analytique, c'est à dire en utilisant la distance) comme barycentre de l'ensemble des formes en se rapprochant de Camel$_{11}$. Pour ce faire nous sommons les rangs des formes pondérés par leur occurrence et nous divisons par le nombre de formes.
Donc le rang moyen est :

$$f_m = \sum_{i=1}^{n} (i * N_i) / N$$

Selon les principes de la théorie des probabilités, il est également possible de déterminer ce que nous appellerons le rang médian de l'ensemble des formes, qui coupe la somme des occurrences des formes rangées par occurrence décroissante en 2 parties égales.

6. Indicateurs pour la comparaison et le découpage d'une loi

Fort des indicateurs décrits dans la littérature, notre intérêt est de les utiliser pour présenter une interprétation des lois bibliométriques en trois parties, en obtenant des coupures en fonction d'indices, les plus objectifs possibles, ou du moins reproductibles car cette séparation nous permet de valoriser les signaux faibles quand l'innovation est recherchée ou au contraire d'éliminer un bruit (au sens statistique) quand des cartographies générales sont recherchées. Deux familles d'indices sont présentées. Les premiers sont des indicateurs d'évaluation de l'importance de chacune des trois parties, les autres seront des indicateurs de séparation ou de seuil afin d'assurer le découpage des distributions.

6.1 Indicateurs de l'importance de chacune des trois parties

Nous voulons mesurer l'importance relative des différentes parties de la courbe Trivial (T), Bruit (B), Information (I), en quelque sorte définir le « poids » de chaque partie. Comme

chacun des indices définis ci-dessus recouvre ces concepts, nous les dégageons en combinant ces différents indices.

$$L_T = H_1 - H_2$$
$$L_I = (H_1 + H_2) / H_1$$
$$L_B = H_0 - H_1$$

Ainsi L_T, L_I et L_B sont respectivement des indicateurs de l'importance relative des différentes parties, permettant de comparer les "Zipfiennes" entre elles. Ces indices ont été déterminés heuristiquement et leur pertinence vérifiée expérimentalement.
Notons que $L_T \geq 0$, $L_B \geq 0$ et $1 \leq L_I \leq 2$ d'aprés nos définitions.

6.2 Critères de séparation

De même que la mesure de l'importance relative de chacune des parties, nous avons voulu pouvoir séparer ces différentes parties « physiquement » de façon systématique, quel que soit le corpus, à partir de la statistique que nous venions d'établir.

6.2.1 Coupure entre la partie intéressante et la partie bruit

Nous la definissons par :

$$C_b = D_0 - D_0\left[\frac{1}{R} * \frac{L_B}{L_B + L_T + L_I}\right]$$

Ainsi le rang de la forme qui sépare l'ensemble « trivial plus information » du bruit est la partie entière de la coupure :

$$i = E(C_b)$$

Ce rang correspond à une fréquence. Etant donné qu'à l'intérieur d'une gamme de fréquence il est impossible statistiquement de différencier une forme d'une autre, nous prendrons donc le rang inferieur de cette fréquence comme rang de coupure bruit.

$$r_b = \inf rang(f_i)$$

Nous pouvons remarquer dans le cas équiprobable, à savoir p_i = constante = 1/n, que R = 1 et C_b = n.

6.2.2 Coupure entre la partie triviale et la partie intéressante

Elle peut être définie par :

$$C_t = [D_2]$$

Ainsi le rang de la forme qui sépare le trivial de l'information est la partie entière de la coupure :

$$i = E(C_t)$$

Mais ce rang correspond à une fréquence. Etant donné qu'à l'interieur d'une gamme de fréquence il est impossible statistiquement de différencier une forme d'une autre, nous prendrons donc le rang supérieur de cette fréquence comme rang de coupure trivial.

$$r_t = \sup rang(f_i)$$

Application à des distributions expérimentales

Nous allons passer en revue les différents termes en expliquant pour certain ordre a de l'entropie ce qu'il apporte dans la compréhension d'une loi « Zipfienne » et comment s'en servir comme indicateur « utile » pour comparer, interpréter des distributions expérimentales. Pour ce faire nous allons nous appuyer sur trois distributions expérimentales. Ces trois distributions sont issues d'un même télédéchargement de notices. Notre choix de comparer les champs d'un même télédéchargement est dû au fait que les comparaisons de distributions Zipfiennes peuvent être sensibles au nombre de références et pour le moment nous voulions nous affranchir de ce problème. Deux sont extraites du même champ (descripteur) mais une fois en considérant les multitermes (nous l'appellerons MT), une fois en considérant les unitermes par choix de séparateurs différents (nous l'appellerons UT). Ces deux distributions sont représentées Figure 1 et 2. La dernière de nos distributions est issue du champ auteur et est représentée Figure 3 (nous l'appellerons AU). Notre exemple à trois champs dont deux « dépendants » a pour objet d'illustrer l'incidence des modifications de distribution sur les paramètres calculés.

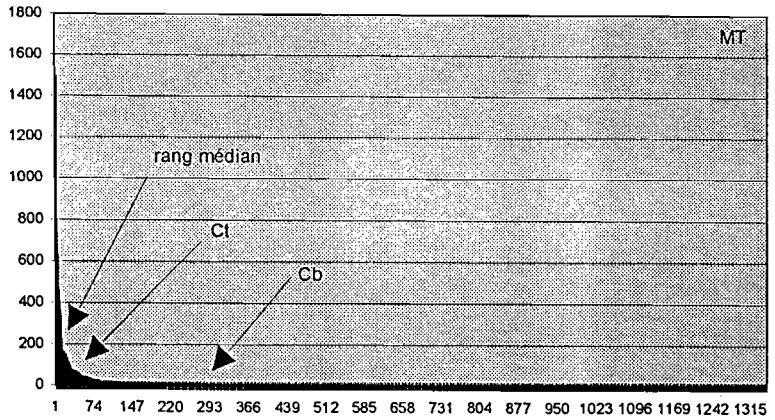


Figure 1 Distribution expérimentale champ descripteur multiterme

(en ordonnée le nombre d'occurrences, en abscisse le rang de la forme)

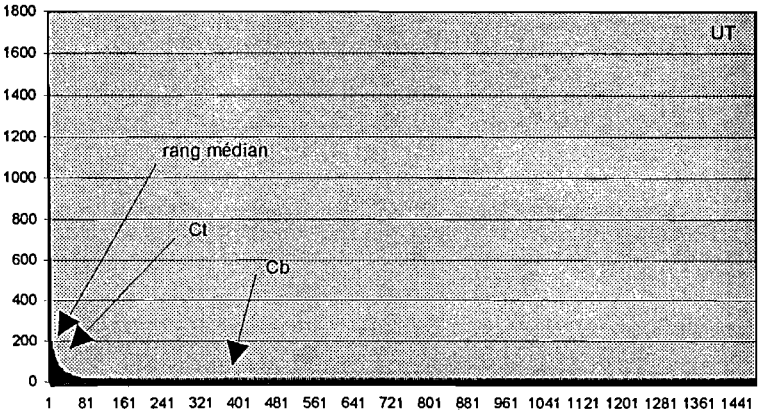


Figure 2 Distribution expérimentale champ descripteur uniterme

(en ordonnée le nombre d'occurrences, en abscisse le rang de la forme)

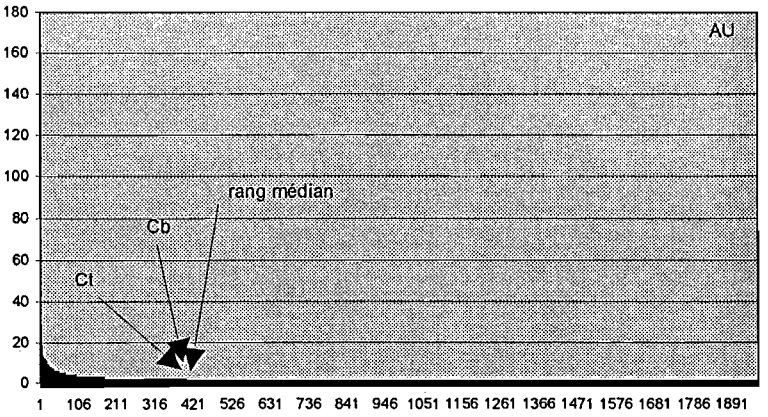


Figure 3 Distribution expérimentale champ auteur

(en ordonnée le nombre d'occurrences, en abscisse le rang de la forme)

6.3 -Indicateurs de description et de comparaison de lois de Zipf

Il est clair que tous ces indicateurs ne peuvent avoir d'intérêt que s'il est possible de les « valider » avec des séries expérimentales réelles. C'est la raison pour laquelle nous avons développé un programme capable de calculer tous les paramètres décrits ci-dessus sur des données issues du logiciel DATAVIEW (@C.R.R.M.). Les résultats issus du traitement des trois séries sont reportés dans le Tableau 1.

Tableau 1 Indices calculés sur les séries expérimentales

Indicateur\Distributions expérimentales	MT	UT	AU
Entropie d'ordre a			
H_0	7.21	7.3	7.58
H_1	4.64	4.63	7.13
H_2	3.37	3.05	5.57
Diversité d'ordre a			
D_0	1346	1481	1964
D_1	103.82	102.36	1245.15
D_2	29.15	21.19	262.94
p_i min	4.6e-5	7.3e-5	3.2e-4
p_i max	8.4e-2	1.3e-1	5.3e-2
Régularité R	0.64	0.63	0.94
rang moyen fm	88	116	659
rang médian	12	9	408
Indicateurs d'intensité des parties			
Triviale (L_T)	1.27	1.58	1.56
Information (L_I)	1.73	1.66	1.78
Bruit (L_B)	2.56	2.67	0.46
Indicateurs de rupture en trois parties			
rang coupure Trivial/Intéressant Ct	29	21	397
rang coupure Intéressant/Bruit Cb	311	369	398

Dans notre exemple, nous remarquons peu de différence entre MT et UT. En effet ces distributions proviennent du même corpus, étant extraites du même champ en considérant soit les multitermes, soit les unitermes.
Nous pouvons malgré tout observer que le trivial est plus fort pour UT que pour MT (plus de formes, relativement, à haute occurrence) comme le bruit (plus de formes).
Par contre, l'information est plus forte pour la distribution MT qui contient relativement plus de formes à occurrence intermédiaire.
Lorsque nous regardons les associations de formes (paires de formes) de la partie information, nous voyons apparaître une décroissance du nombre de paires pour UT vis-à-vis de MT. C'est un résultat que nous employons empiriquement pour diminuer le nombre d'association de mots clef multitermes non thésaurés qui génère un réseau de paires fortement connexe trop grand pour être interprétable.
La régularité (donnant R = 0.63 pour UT et R = 0.64 pour MT) nous montre que ces courbes ne sont pas en pente douce, contrairement à la distribution AU (qui a une régularité de 0.94) où nous ne pouvons par conséquent que difficilement séparer l'information du bruit. Ceci nous donne des valeurs pour H_0 et H_1 très voisines.
Ceci est normal sur un champ auteur où nous constatons en règle générale que quelques auteurs publient beaucoup et que beaucoup d'auteurs publient peu dans un domaine (cf la loi de Lotka). Il y a souvent très peu d'auteurs qui publient à fréquences intermédiaires, et ces auteurs sont difficiles à isoler.
Ceci est bien sûr complètement différent dans un champ descripteur où des sous-domaines apparaissent à des fréquences intermédiaires.
Dans ce cas, le bruit n'est pas marginalisé puisque la forme la moins fréquente représente une occurrence de ~ 0.66% de l'occurrence maximale, d'où la faible valeur relative du paramètre bruit L_B = 0.46 et l'impossibilité de dégager la partie information qui se confond avec le bruit.
Cette extension de la courbe avec une faible différence d'ordre de grandeur (un facteur deux) pour les occurrences nous est aussi mise en évidence par la forme moyenne qui est de 660 pour 1964 formes. Alors qu'elle est de 88 pour 1346 formes et 117 pour 1481 formes sur des corpus ou les occurrences varient sur trois ordres de grandeur.

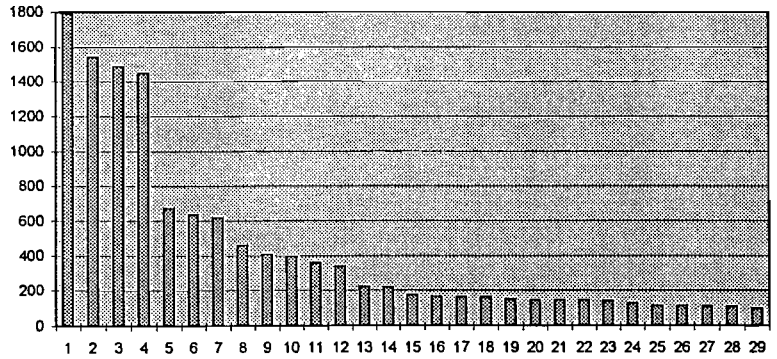
Le rang médian corrobore cette observation : il est de 408 pour la distribution AU alors qu'il vaut environ 10 pour les deux autres.

6.4 -Indicateurs de ruptures de lois de Zipf

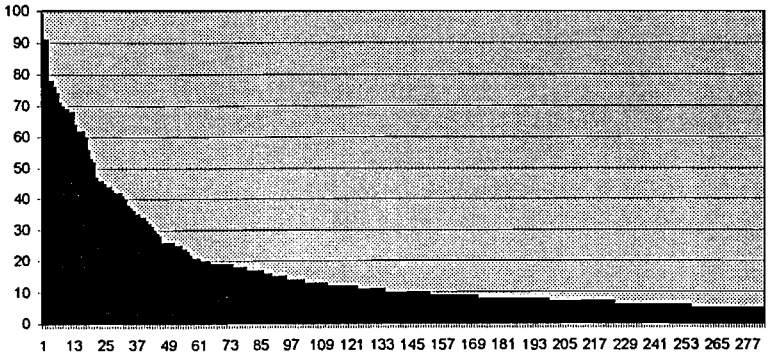
Dans nos exemples la coupure du « bruit » sépare les formes qui ont une occurrence inférieure à 0.2% du nombre de références. Il apparaîtra alors légitime de considérer toute forme ayant cette occurrence ou une occurrence inférieure comme faisant partie du bruit.
Dans nos exemples la coupure du « trivial » sépare les formes qui ont une occurrence supérieure à 4% du nombre de références. Il apparaîtra alors légitime de considérer toute forme ayant cette occurrence ou une occurrence supérieure comme faisant partie du trivial.
Ceci n'est pas vrai pour la distribution AU car la différence relative entre les occurrences n'est pas suffisante. Ainsi le trivial et l'information sont quasiment contenus dans ce qui constitue habituellement la coupure trivial, C_t. C_t nous donne ici tout ce qui est supérieur ou égal à 1.5% de l'occurrence maximale.

De ces trois distributions, nous pouvons faire une représentation graphique avec une coupure en trois parties conformément à la Figure 4.

Partie "triviale" de MT



Partie "information"



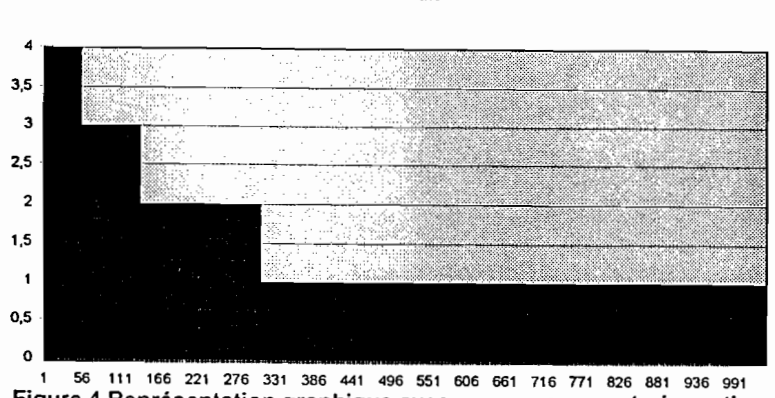


Figure 4 Représentation graphique avec une coupure en trois parties
(en abscisse le nombre de formes de chaque partie et en ordonnée le nombre d'occurrences)

La partie triviale représente les termes à forte fréquence, fortement redondants. Nous l'avons qualifiée de triviale en ce sens que le contenu informatif de ces formes est souvent faible et relève le plus souvent d'informations que tout le monde connaît.
La partie qualifiée de bruit représente ce qui au sens statistique n'a pas de poids. Quand une cartographie générale d'un domaine est recherchée, ils constituent les termes qui rendent creuse la matrice[12] et sont donc à éliminer. En validant les termes un à un d'un point de vue informatif, avec l'aide d'un expert du domaine, c'est dans cette partie, précisément que peuvent être trouvés des termes à très haut poids innovant au milieu de beaucoup de termes aberrants. Ils constituent donc un bruit statistique, mais sont les plus riches en contenu informatif.
La partie intermédiaire, qualifiée d'intéressante contient des termes qui peuvent constituer des sous thèmes dans un corpus. C'est cette partie qui est la plus précieuse dans des cartographies globales. C'est celle qui de façon automatique, sans l'aide d'un expert, a le plus de probabilité de fournir des résultats intéressants.

7. Conclusion

Notre volonté était de montrer que notre discipline n'est pas un cas particulier et que des indicateurs peuvent être construits pour caractériser ses distributions. Notre discipline possède bien quelques indicateurs spécifiques tel que le coefficient multiplicateur de Bradford. Il nous semblait intéressant d'entreprendre une démarche unificatrice avec d'autres disciplines. Cette première étape se voulait essentiellement didactique quant à certains indicateurs. De plus nous voulions valider le fait que ces indicateurs soient le réel reflet d'interprétation « empiriques » que nous avions, même s'il est clair que notre convention (r_i = i) et notre définition de C_b et C_l comportent une part d'arbitraire. Nous l'avons montré à l'aide de nos trois exemples.
Notre prochain but sera d'unifier les indicateurs spécifiques de notre discipline en démontrant qu'ils ne sont que des cas particuliers d'indicateurs plus globaux existant par ailleurs. Nous comptons, à partir du modèle Zipf-Mandelbrot[13], rendre compte des lois de Lotka et Bradford[14]. Cette étape franchie, nous comptons établir les dimensions fractales[9, 7] et topologiques[15] de ces lois pour nous permettre de nouvelles représentations signifiantes des corpus en terme d'information. Ces représentations plus proches du respect des propriétés de l'information, devraient être plus performantes en terme de détection de zones potentiellement innovantes.

Bibliographie

1 Legendre P., Legendre L.
Ecologie numérique (2ième éd.)
Masson, Paris et les Presses de l'Université du Québec,1984, p187-242.

2 Mayer-Kress G.,(ed.)
Dimensions and Entropies in Chaotic Systems
Proc. Internat. Workshop Pecos River Ranch, New Mexico, (11-16 sept 1985), Part II
Springer-Verlag, Berlin, 1986.

3 Voll M.
Analyse des données (3ième ed.)
Economica, 1985, Paris.

4 Lebart L., Salem A.
Statistique textuelle
Dunod, 1994, Paris.

5 Hill M.O.
Diversity and Evenness: a unifying notation and its consequences,
Ecology,1973, Vol.54,No.2 p427-433.

6 Daget P.
Le nombre de diversité de Hill : un concept unificateur dans la théorie écologique,
Oecol. Gener., 1980, Vol.1, No.1 p51-70.

7 Mandelbrot B.
Fractals-Forms, Chance and Dimension
Freeman (ed.), 1977, San-Francisco.

8 Bergé P., Pomeau Y., Vidal Ch.
L'Ordre dans le Chaos
Hermann (ed.), 1988, Paris.

9 Van Raan A.F.J.
Fractal dimension of cocitation,
Nature, 18 october 1990, Vol.347 , p626.

10 Yaglom A.M., Yaglom I.M.
Probabilité et Information
Traduction Mercouroff W.
Dunod, 1959, Paris.

11Camel Y.
Probabilités Théorie et Application
Chap 9, p.109,
Eyrolles, 1988, Paris.

12 Quoniam L.
Bibliométrie sur des références bibliographiques: méthodologie
in La Veille Technologique (Dou H. Ed.), Dunod, 1992, Paris.

13 Mandelbrot B.
Contribution à la théorie mathématique des jeux de communication,
Publ. Inst. Statist. Univ., 1953, Paris 2, p1-124.

14 Rostaing H.
Veille Technologique et bibliométrie : Concepts, Outils, Applications
Thèse soutenue le 13 janvier 1993, Univ. Aix-Marseille III.

15 Badii R.
Conservation laws and thermodynamic formalism for dissipative dynamical systems,
Thèse 1987 universität Zürich.

16 Egghe L.
Bridging the gap: conceptual discussions on infometrics,
Scientometrics, may 1994, Vol.30, N°1, p35.

17 Fairthorne R.
Citation classic - empirical hyperbolic distributions (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction,
Current contents/Social & Behavioral Sciences, 1987, N°3, p16.

La « statistique » des lois de Zipf

LHEN J., **LAFOUGE T.,ELSKENS Y., *QUONIAM L., *DOU H.**

*****C.R.R.M. Centre Scientifique de St Jérôme F-13397 Marseille CEDEX 20
******ENSIIB, 78 rue du 11 Nov. 1918. F-69623 Villeurbanne CEDEX
*******Equipe turbulence plasma, U.R.A. 773, C.N.R.S- Université de Provence, I.M.T. Château-Gombert F-13451 Marseille CEDEX 20

Mots clés: distributions bibliométriques, théorie de l'information, indicateurs

Résumé

Les lois de Zipf, Lotka et Bradford ont fait l'objet de nombreux travaux dans le domaine de la bibliométrie. La plupart de ces travaux ont porté sur une description analytique de ces courbes. D'autres travaux font état de l'utilisation de la théorie de l'information de Shannon.
Notre but ici est de décrire un certain nombre de ces indicateurs pour dégager une « statistique » opérationnelle capable de caractériser des lois de Zipf expérimentales, comme la moyenne arithmétique et la variance caractérisent des lois présentant une tendance centrale.
Les écoles de pensée française en bibliométrie sont les seules, à notre connaissance, à présenter une interprétation de ces lois en trois parties (Trivial, Information, Bruit). Cette façon d'interpréter nous permet de valoriser les signaux faibles quand l'innovation est recherchée ou au contraire d'éliminer un bruit (au sens statistique) quand des cartographies générales sont recherchées. Dans la littérature, nous trouvons des découpages en deux zones (le coeur et la dispersion) ou en plusieurs zones (dont le nombre fluctue à chaque expérimentation) conformément aux travaux de Lotka et Bradford.
Notre but sera donc aussi la détermination des seuils entre ces différentes zones, seuils qui pour l'instant sont déterminés de façon purement empirique, fondée sur l'expérience et l'expérimentation.
Tout au long de notre exposé, des exemples seront étudiés et nous conclurons par des perspectives de prolongement de ce travail.

Abstract

In our bibliometry field, there are many works on the laws of Zipf, Lotka and Bradford. Most of them explain analytic description of these curves. Other works show the use of the Shannon information theory.
Our aim in this paper will be to determine some criteria for the development of a piece of statistics able to characterize experimental Zipf laws, as the arithmetic average and variance characterize laws with a central tendency.
The french school of thinking, to our knowledge, seems to be the only one showing an interpretation in three points (trite, information, noise) of these curves. This way of interpretation makes low signals stand out whenever we look for breakthroughs and on the contrary suppresses the noise when general mappings are required. Among scientific writing we found some cuts of the curve in two zones (dispersion and core) or in several (the number of zones depends of the experimental laws) according to the work of Bradford and Lotka. So our aim will be to define thresholds between these different zones, thresholds which are currently defined by a rule of thumb, based on experimentation.
All along our lecture, some examples will be studied and we will conclude on opening on the subject.

1. Introduction

Notre discipline n'est pas la seule à posséder des lois « Zipfiennes »: la biologie, la géologie[1], la physique[2] et l'économie[3] en comptent de nombreux exemples. Ces lois, caractéristiques de diversité plutôt que de dispersion, ont souvent été liées à la théorie de l'entropie. Notre but va être ici de préciser des formulations déjà existantes dans la littérature, d'apporter des compléments utiles pour interpréter et comparer des distributions « Zipfiennes », et enfin de les appliquer à des distributions expérimentales. Les formulations que nous passerons en revue ne sont applicables que sur des séries expérimentales longues (Nous estimons que l'ordre de grandeur du nombre de références minimum à prendre en compte est de 500 à 1000 références).

2. Entropie et Diversité d'ordre a

Renyi propose en 1961 une approche unifiée pour décrire ces lois en définissant **l'entropie d'ordre a** :

$$\text{pour } a \neq 1 : H_a = \frac{1}{1-a} \log \sum_{i=1}^{n} (p_i)^a \quad \text{(logarithme népérien)}$$

$$\text{pour } a = 1 : H_1 = \lim_{a \to 1} H_a = -\sum_{i=1}^{n} p_i \log p_i$$

où **n** représente le nombre de formes distinctes dans un champ sur l'ensemble du corpus et $\mathbf{p_i}$ représente la probabilité ou fréquence d'apparition de la forme i dans le corpus.
Soit $\mathbf{p_i = N_i / N}$ avec $\mathbf{N_i}$ nombre d'occurrences de la forme i, i = 1,... n et

$\mathbf{N} = \sum_{\mathbf{i}} \mathbf{N_i}$ nombre total d'occurrence de toutes les formes.

Par convention, nous ordonnerons les probabilitées : $1 \geq p_1 \geq \ldots \geq p_n \geq 0$.

La dénomination « forme » représente les formes graphiques constituées de caractères alphanumériques contenus entre deux séparateurs[4]. Par exemple les formes peuvent être des mots-clés multi- ou uni-termes, mais aussi des noms d'auteurs avec ou sans prénoms, ou des codes.

Hill[5, 6] propose une approche qui va définir une mesure de la composition en forme, une mesure de diversité. **Les diversités d'ordre a selon Hill** (cette nomenclature est due à Daget[6] 1980) sont**:**

$$D_a = \exp H_a$$

Selon Hill ces diversités regroupent plusieurs concepts qui sont complémentaires. C'est ainsi que nous les comprendrons. Notons que H_a est toujours positive et $D_a > 1$; H_a et D_a sont des fonctions décroissante de a.

3. Les différents ordres de l'entropie de Renyi et de la diversité de Hill

3.1 - Ordre 0:

A l'ordre 0 les expressions générales deviennent:

$$H_0 = \log n \qquad \text{et} \qquad D_0 = \exp H_0 = n$$

dit autrement la diversité à l'ordre 0 correspond au nombre de formes distinctes présentes dans un champ sur l'ensemble des références. Cette diversité rend donc compte de l'étalement de la courbe principalement dû aux basses fréquences. Dans l'interprétation classique d'une loi « Zipfienne », nous disons: « peu de formes à fortes fréquences, beaucoup de formes à faibles fréquences ». Le nombre de formes est donc plus un indicateur qualifiant les faibles fréquences, la partie étalée de la courbe, la diversité à faible fréquence, le coté bruit de la courbe.
Procaccia[2] appelle H_0 (qu'il note D_0) la dimension fractale, elle est liée aussi à la dimension de Hausdorff-Besicovitch[7, 8] qui donne une représentation fractale[8] entre autres. Certains auteurs ont déjà signalé la dimension fractale des données bibliométriques, par exemple A.F.J. Van Raan[9] et R. Fairthorne[17]. Quant à Hill, il note N_0 ce que nous appelons ici D_0.

3.2 - Ordre 1:

A l'ordre 1 les équations générales deviennent:

$$H_1 = -\sum_{i=1}^{n} p_i \log p_i \qquad \text{et} \qquad D_1 = \exp H_1$$

Cette formule correspond à celle de l'entropie thermodynamique ou l'information de Shannon[10].
La diversité (ou les diversités) est considérée pour Hill comme une « variance » des formes. Elle peut être définie comme la mesure de la composition en forme.
En physique[2] H_1 est définie comme la dimension d'information.

3.3 - Ordre 2:

A l'ordre 2 l'entropie de Renyi est associée à une grandeur nommée concentration par Hill et Daget.

$$C = \sum_{i=1}^{n} \left(\frac{n_i}{N} \right)^2 = \sum_{i=1}^{n} \left(p_i \right)^2$$

A l'ordre 2 les équations générales deviennent:

$$H_2 = -\log \sum_{i=1}^{n} \left(p_i \right)^2 = -\log \left(C \right) \qquad \text{et} \qquad D_2 = \frac{1}{C}$$

Procaccia[2] donne à H_2 le nom de dimension de corrélation. De plus Hill montre que D_2 est l'inverse de l'indice de Simpson ou concentration C[16].

3.4 - Ordre $-\infty$:

A cet ordre, nous mettons en évidence la forme qui a la probabilité d'apparition minimum. Information triviale en soi, mais qui est bien un indicateur de la diversité. Par analogie avec

la nomenclature de Daget et de Hill nous pouvons dire que $D_{-\infty}$ identifie les occurrences (le nombre total d'occurrence). Il est l'inverse de la probabilité la plus faible.

$$D_{-\infty} = 1 / (p_i)_{min}$$

ou $(p_i)_{min} = (\inf N_i) / N$ avec le minimum théorique $(p_i)_{min} = 1 / N$

par construction $(p_i)_{min} = p_n$

La grandeur $(p_i)_{min}$ est l'inverse de la contribution spécifique la plus faible, laquelle ne peut être inférieure à 1/N, mais peut être plus élevée si la fréquence absolue minimale est supérieure à 1. Il est clair que $(p_i)_{min}$ soit inférieur ou égal à 1/n (avec égalité dans le cas de diversité maximum). Nous avons donc une position extrême pour l'asymptote supérieure dans le graphe de D_a en fonction de a que nous pouvons interpréter en disant qu'elle traduit le nombre maximal « d'espèces possibles » avec N présences.

3.5 - Ordre $+\infty$:

A cet ordre, c'est la fréquence maximum d'apparition qui est révélée. Information triviale en soi, mais qui est aussi un indicateur de la diversité. Encore une fois par analogie avec Daget et Hill nous pouvons dire que $D_{+\infty}$ identifie la forme dominante ou les formes codominantes. Il représente l'inverse de la probabilité la plus forte.

$$D_{+\infty} = 1 / (p_i)_{max}$$

avec $(p_i)_{max} = (\sup N_i) / N$ et par construction: $\sup N_i = N_1$.

La grandeur $(p_i)_{max}$ (supérieure ou égale à 1/n) est l'inverse de la contribution spécifique la plus forte. Nous avons donc une position de l'asymptote inférieure dans le graphe de D_a en fonction de a que nous pouvons interpréter en disant qu'elle traduit le nombre minimal « d'espèces possibles », pour Daget.

Pour Hill les diversités ont un intérêt théorique car elles peuvent être reliées à la stabilité, la maturité, la productivité et l'évolution temporelle d'un système. Elles sont une mesure de paramètre dont les valeurs observées peuvent être expliquées par diverses théories selon la discipline: Théorie cybernétique, information de Shannon, entropie thermodynamique, théorie du chaos.

4. Les rapports entre les différents termes de l'entropie d'ordre a

Nous pouvons ajouter comme Legendre[1] un autre critère pour définir la forme de la « Zipfienne »: la régularité R représentant l'information relative contenue dans la forme de la distribution. Elle a été introduite pour la première fois par Pielou 1966 et se définit en terme de Hill et Daget (en fait Hill note la régularité J) par :

$$R = \left| \frac{H_1}{H_0} \right|$$

Remarque : La valeur absolue n'est pas nécessaire, mais la régularité a été introduite pour la première fois comme cela nous l'avons donc conservée.

La régularité est le rapport des entropies correspondant aux nombres de diversité d'ordre 0 et 1. R joue le même rôle qu'un coefficient de variation, ce qui permet de comparer deux distributions à l'aide d'un coefficient « normalisé ». Il est clair que $0 \leq R \leq 1$. De plus, R = 1 si et seulement si toutes les probabilitées sont égales : $p_i = 1/n$ (cas de diversité maximale, pour laquelle $H_a = \log n$ et $D_a = n$, pour tout a).

5. Indicateurs de tendance centrale

L'étude des fréquences des formes ne donne pas d'indication sur les relations entre celles-ci. A priori, nous ne connaissons pas la structure de l'espace des formes indicées par i. Nous cherchons cependant à identifier, approximativement, les formes qui appartiennent aux fortes fréquences, aux faibles fréquences ou à la partie centrale de la distribution des fréquences. Nous pouvons alors définir à partir des entropies de Renyi la fréquence moyenne $\overline{p}$ et la variance des fréquences V_p qui sont deux indicateurs classiques.

$$\overline{p} = \frac{1}{n}\sum_{i=1}^{n} p_i = \frac{1}{D_0}$$

$$V_p = \frac{1}{n}\sum_{i=1}^{n} p_i^2 - \overline{p}^2 = \frac{D_0 - D_2}{D_0^2 D_2}$$

Ces deux indicateurs ne caractérisent cependant pas directement les formes, mais seulement les fréquences. A ce titre, ils ne nous intéressent pas directement dans la suite de l'exposé. Mais ce n'est pas une voie à négliger définitivement.

Pour définir une répartition des formes, nous associons à chacun un rang r_i, par ordre décroissant des fréquences. Le choix le plus simple est : $r_i = i$. Lorsque plusieurs formes ont la même fréquence, nous leur attriburons des rangs successifs, dans un ordre arbitraire.
A l'aide de ce rang, nous pouvons calculer une valeur moyenne sur les formes. S'il y a peu de sens à calculer une moyenne arithmétique des fréquences ou des occurrences, il est plus naturel de calculer une valeur moyenne « géométrique » (par opposition à analytique, c'est à dire en utilisant la distance) comme barycentre de l'ensemble des formes en se rapprochant de $Camel_{11}$. Pour ce faire nous sommons les rangs des formes pondérés par leur occurrence et nous divisons par le nombre de formes.
Donc le rang moyen est :

$$f_m = \sum_{i=1}^{n} \left(i * N_i \right) \Big/ N$$

Selon les principes de la théorie des probabilités, il est également possible de déterminer ce que nous appellerons le rang médian de l'ensemble des formes, qui coupe la somme des occurrences des formes rangées par occurrence décroissante en 2 parties égales.

6. Indicateurs pour la comparaison et le découpage d'une loi

Fort des indicateurs décrits dans la littérature, notre intérêt est de les utiliser pour présenter une interprétation des lois bibliométriques en trois parties, en obtenant des coupures en fonction d'indices, les plus objectifs possibles, ou du moins reproductibles car cette

séparation nous permet de valoriser les signaux faibles quand l'innovation est recherchée ou au contraire d'éliminer un bruit (au sens statistique) quand des cartographies générales sont recherchées. Deux familles d'indices sont présentées. Les premiers sont des indicateurs d'évaluation de l'importance de chacune des trois parties, les autres seront des indicateurs de séparation ou de seuil afin d'assurer le découpage des distributions.

6.1 Indicateurs de l'importance de chacune des trois parties

Nous voulons mesurer l'importance relative des différentes parties de la courbe Trivial (T), Bruit (B), Information (I), en quelque sorte définir le « poids » de chaque partie. Comme chacun des indices définis ci-dessus recouvre ces concepts, nous les dégageons en combinant ces différents indices.

$$L_T = H_1 - H_2$$
$$L_I = (H_1 + H_2) / H_1$$
$$L_B = H_0 - H_1$$

Ainsi L_T, L_I et L_B sont respectivement des indicateurs de l'importance relative des différentes parties, permettant de comparer les "Zipfiennes" entre elles. Ces indices ont été déterminés heuristiquement et leur pertinence vérifiée expérimentalement.

Notons que $L_T \geq 0, \quad L_B \geq 0 \quad \text{et} \quad 1 \leq L_I \leq 2$ d'aprés nos définitions.

6.2 Critères de séparation

De même que la mesure de l'importance relative de chacune des parties, nous avons voulu pouvoir séparer ces différentes parties « physiquement » de façon systématique, quel que soit le corpus, à partir de la statistique que nous venions d'établir.

6.2.1 Coupure entre la partie intéressante et la partie bruit

Nous la definissons par :

$$C_b = D_0 - D_0\left[\frac{1}{R} * \frac{L_B}{L_B + L_T + L_I}\right]$$

Ainsi le rang de la forme qui sépare l'ensemble « trivial plus information » du bruit est la partie entière de la coupure :

$$i = E(C_b)$$

Ce rang correspond à une fréquence. Etant donné qu'à l'intérieur d'une gamme de fréquence il est impossible statistiquement de différencier une forme d'une autre, nous prendrons donc le rang inferieur de cette fréquence comme rang de coupure bruit.

$$r_b = \inf rang(f_i)$$

Nous pouvons remarquer dans le cas équiprobable, à savoir p_i = constante = 1/n, que R = 1 et C_b = n.

6.2.2 Coupure entre la partie triviale et la partie intéressante

Elle peut être définie par :

$$C_t = [D_2]$$

Ainsi le rang de la forme qui sépare le trivial de l'information est la partie entière de la coupure :

$$i = E(C_t)$$

Mais ce rang correspond à une fréquence. Etant donné qu'à l'interieur d'une gamme de fréquence il est impossible statistiquement de différencier une forme d'une autre, nous prendrons donc le rang supérieur de cette fréquence comme rang de coupure trivial.

$$r_t = \sup rang(f_i)$$

Application à des distributions expérimentales

Nous allons passer en revue les différents termes en expliquant pour certain ordre a de l'entropie ce qu'il apporte dans la compréhension d'une loi « Zipfienne » et comment s'en servir comme indicateur « utile » pour comparer, interpréter des distributions expérimentales. Pour ce faire nous allons nous appuyer sur trois distributions expérimentales. Ces trois distributions sont issues d'un même télédéchargement de notices. Notre choix de comparer les champs d'un même télédéchargement est dû au fait que les comparaisons de distributions Zipfiennes peuvent être sensibles au nombre de références et pour le moment nous voulions nous affranchir de ce problème. Deux sont extraites du même champ (descripteur) mais une fois en considérant les multitermes (nous l'appellerons MT), une fois en considérant les unitermes par choix de séparateurs différents (nous l'appellerons UT). Ces deux distributions sont représentées Figure 1 et 2. La dernière de nos distributions est issue du champ auteur et est représentée Figure 3 (nous l'appellerons AU). Notre exemple à trois champs dont deux « dépendants » a pour objet d'illustrer l'incidence des modifications de distribution sur les paramètres calculés.

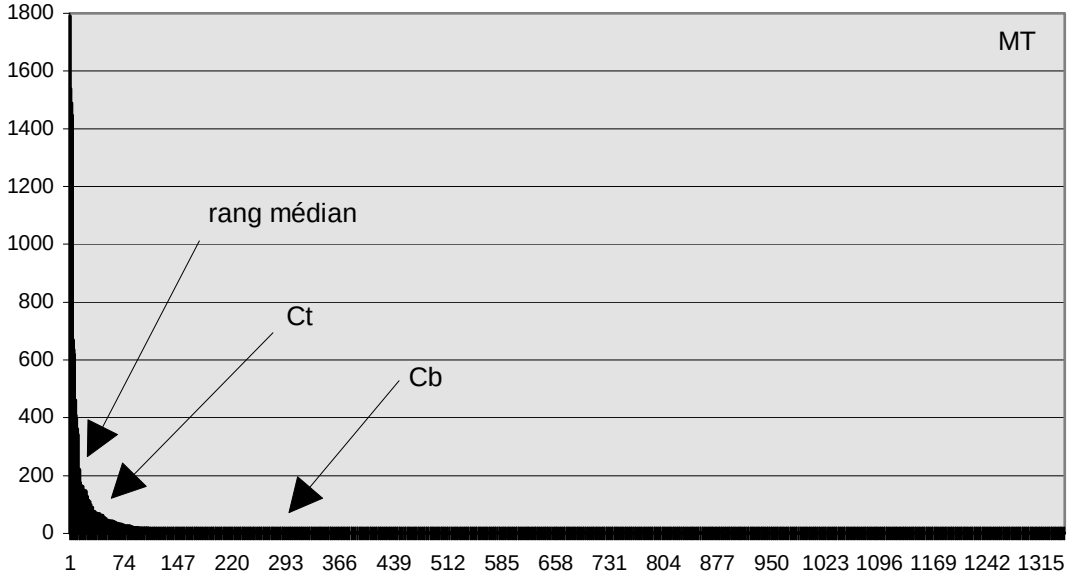


Figure 1 Distribution expérimentale champ descripteur multiterme

(en ordonnée le nombre d'occurrences, en abscisse le rang de la forme)

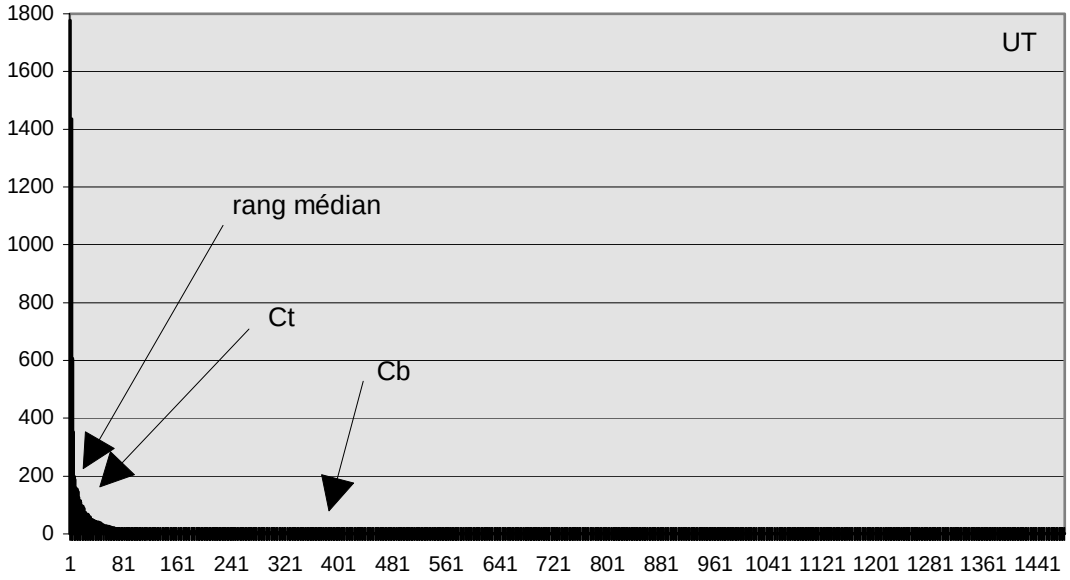


Figure 2 Distribution expérimentale champ descripteur uniterme

(en ordonnée le nombre d'occurrences, en abscisse le rang de la forme)

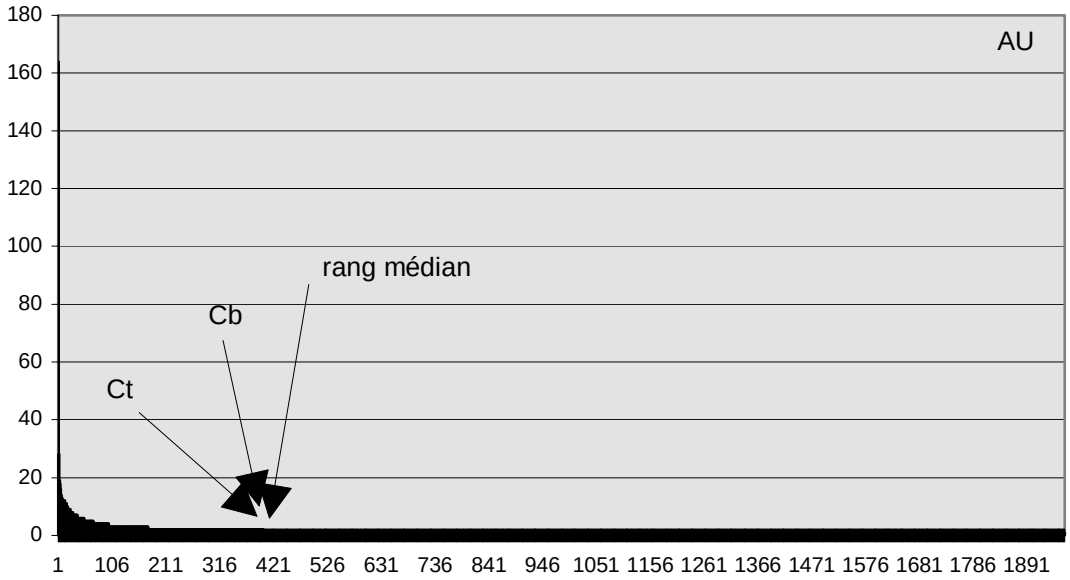


Figure 3 Distribution expérimentale champ auteur

(en ordonnée le nombre d'occurrences, en abscisse le rang de la forme)

6.3 -Indicateurs de description et de comparaison de lois de Zipf

Il est clair que tous ces indicateurs ne peuvent avoir d'intérêt que s'il est possible de les « valider » avec des séries expérimentales réelles. C'est la raison pour laquelle nous avons développé un programme capable de calculer tous les paramètres décrits ci-dessus sur des données issues du logiciel DATAVIEW (@C.R.R.M.). Les résultats issus du traitement des trois séries sont reportés dans le Tableau 1.

Tableau 1 Indices calculés sur les séries expérimentales

Indicateur\Distributions expérimentales	MT	UT	AU
Entropie d'ordre a			
H_0	7.21	7.3	7.58
H_1	4.64	4.63	7.13
H_2	3.37	3.05	5.57
Diversité d'ordre a			
D_0	1346	1481	1964
D_1	103.82	102.36	1245.15
D_2	29.15	21.19	262.94
p_i min	4.6e-5	7.3e-5	3.2e-4
p_i max	8.4e-2	1.3e-1	5.3e-2
Régularité R	0.64	0.63	0.94
rang moyen fm	88	116	659
rang médian	12	9	408
Indicateurs d'intensité des parties			
Triviale (L_T)	1.27	1.58	1.56
Information (L_I)	1.73	1.66	1.78
Bruit (L_B)	2.56	2.67	0.46
Indicateurs de rupture en trois parties			
rang coupure Trivial/Intéressant Ct	29	21	397
rang coupure Intéressant/Bruit Cb	311	369	398

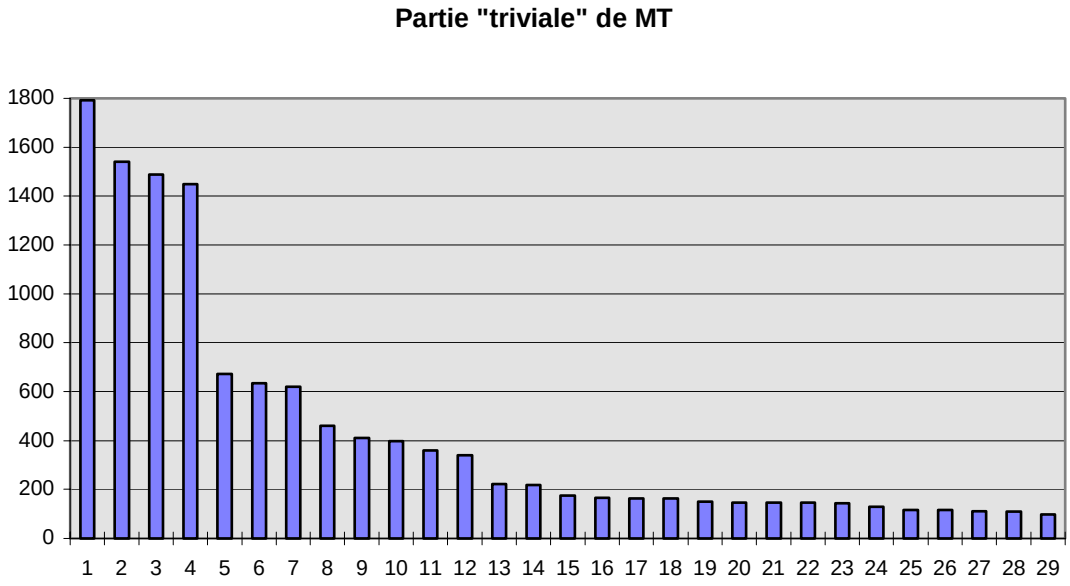
Dans notre exemple, nous remarquons peu de différence entre MT et UT. En effet ces distributions proviennent du même corpus, étant extraites du même champ en considérant soit les multitermes, soit les unitermes.
Nous pouvons malgré tout observer que le trivial est plus fort pour UT que pour MT (plus de formes, relativement, à haute occurrence) comme le bruit (plus de formes).
Par contre, l'information est plus forte pour la distribution MT qui contient relativement plus de formes à occurrence intermédiaire.
Lorsque nous regardons les associations de formes (paires de formes) de la partie information, nous voyons apparaître une décroissance du nombre de paires pour UT vis-à-vis de MT. C'est un résultat que nous employons empiriquement pour diminuer le nombre d'association de mots clef multitermes non thésaurés qui génère un réseau de paires fortement connexe trop grand pour être interprétable.
La régularité (donnant R = 0.63 pour UT et R = 0.64 pour MT) nous montre que ces courbes ne sont pas en pente douce, contrairement à la distribution AU (qui a une régularité de 0.94) où nous ne pouvons par conséquent que difficilement séparer l'information du bruit. Ceci nous donne des valeurs pour H_0 et H_1 très voisines.
Ceci est normal sur un champ auteur où nous constatons en règle générale que quelques auteurs publient beaucoup et que beaucoup d'auteurs publient peu dans un domaine (cf la loi de Lotka). Il y a souvent très peu d'auteurs qui publient à fréquences intermédiaires, et ces auteurs sont difficiles à isoler.
Ceci est bien sûr complètement différent dans un champ descripteur où des sous-domaines apparaissent à des fréquences intermédiaires.
Dans ce cas, le bruit n'est pas marginalisé puisque la forme la moins fréquente représente une occurrence de ~ 0.66% de l'occurrence maximale, d'où la faible valeur relative du paramètre bruit L_B = 0.46 et l'impossibilité de dégager la partie information qui se confond avec le bruit.
Cette extension de la courbe avec une faible différence d'ordre de grandeur (un facteur deux) pour les occurrences nous est aussi mise en évidence par la forme moyenne qui est

de 660 pour 1964 formes. Alors qu'elle est de 88 pour 1346 formes et 117 pour 1481 formes sur des corpus ou les occurrences varient sur trois ordres de grandeur.
Le rang médian corrobore cette observation : il est de 408 pour la distribution AU alors qu'il vaut environ 10 pour les deux autres.

6.4 -Indicateurs de ruptures de lois de Zipf

Dans nos exemples la coupure du « bruit » sépare les formes qui ont une occurrence inférieure à 0.2% du nombre de références. Il apparaîtra alors légitime de considérer toute forme ayant cette occurrence ou une occurrence inférieure comme faisant partie du bruit.
Dans nos exemples la coupure du « trivial » sépare les formes qui ont une occurrence supérieure à 4% du nombre de références. Il apparaîtra alors légitime de considérer toute forme ayant cette occurrence ou une occurrence supérieure comme faisant partie du trivial.
Ceci n'est pas vrai pour la distribution AU car la différence relative entre les occurrences n'est pas suffisante. Ainsi le trivial et l'information sont quasiment contenus dans ce qui constitue habituellement la coupure trivial, C_t. C_t nous donne ici tout ce qui est supérieur ou égal à 1.5% de l'occurrence maximale.

De ces trois distributions, nous pouvons faire une représentation graphique avec une coupure en trois parties conformément à la Figure 4.



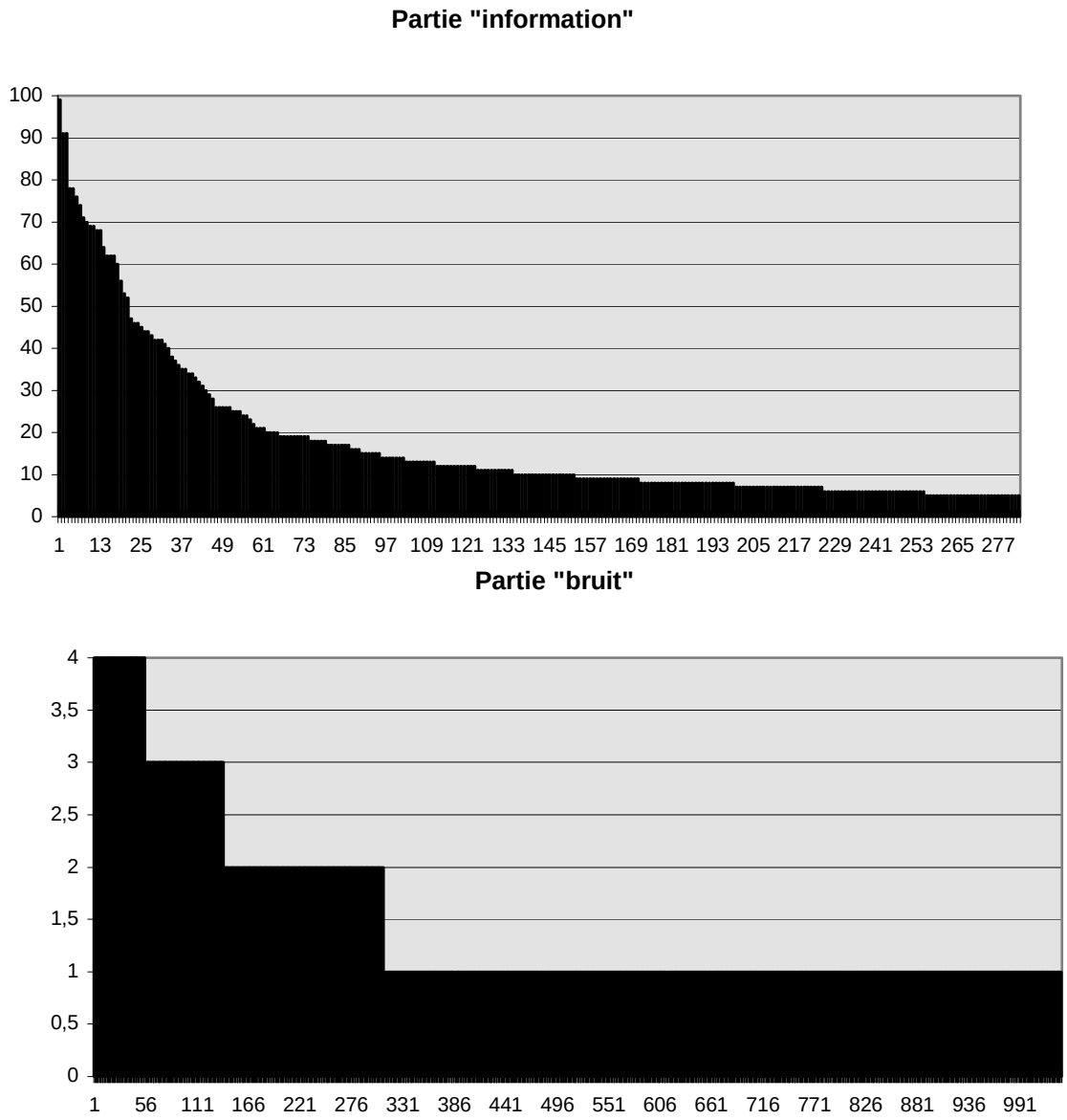


Figure 4 Représentation graphique avec une coupure en trois parties
(en abscisse le nombre de formes de chaque partie et en ordonnée le nombre d'occurrences)

La partie **triviale** représente les termes à forte fréquence, fortement redondants. Nous l'avons qualifiée de triviale en ce sens que le contenu informatif de ces formes est souvent faible et relève le plus souvent d'informations que tout le monde connaît.
La partie qualifiée de **bruit** représente ce qui **au sens statistique** n'a pas de poids. Quand une cartographie générale d'un domaine est recherchée, ils constituent les termes qui rendent creuse la matrice$_{12}$ et sont donc à éliminer. En validant les termes un à un d'un point de vue informatif, avec l'aide d'un expert du domaine, c'est dans cette partie, précisément que peuvent être trouvés des termes à très haut poids innovant au milieu de beaucoup de termes aberrants. Ils constituent donc un bruit statistique, mais sont **les plus riches en contenu informatif**.
La partie intermédiaire, qualifiée d'**intéressante** contient des termes qui peuvent constituer des sous thèmes dans un corpus. C'est cette partie qui est la plus précieuse dans des cartographies globales. C'est celle qui de façon automatique, sans l'aide d'un expert, a le plus de probabilité de fournir des résultats intéressants.

7. Conclusion

Notre volonté était de montrer que notre discipline n'est pas un cas particulier et que des indicateurs peuvent être construits pour caractériser ses distributions. Notre discipline possède bien quelques indicateurs spécifiques tel que le coefficient multiplicateur de Bradford. Il nous semblait intéressant d'entreprendre une démarche unificatrice avec d'autres disciplines. Cette première étape se voulait essentiellement didactique quant à certains indicateurs. De plus nous voulions valider le fait que ces indicateurs soient le réel reflet d'interprétation « empiriques » que nous avions, même s'il est clair que notre convention ($r_i = i$) et notre définition de C_b et C_t comportent une part d'arbitraire. Nous l'avons montré à l'aide de nos trois exemples.
Notre prochain but sera d'unifier les indicateurs spécifiques de notre discipline en démontrant qu'ils ne sont que des cas particuliers d'indicateurs plus globaux existant par ailleurs. Nous comptons, à partir du modèle Zipf-Mandelbrot[13], rendre compte des lois de Lotka et Bradford[14]. Cette étape franchie, nous comptons établir les dimensions fractales[9, 7] et topologiques[15] de ces lois pour nous permettre de nouvelles représentations signifiantes des corpus en terme d'information. Ces représentations plus proches du respect des propriétés de l'information, devraient être plus performantes en terme de détection de zones potentiellement innovantes.

Bibliographie

1 Legendre P., Legendre L.
Ecologie numérique (2ième éd.)
Masson, Paris et les Presses de l'Université du Québec,1984, p187-242.

2 Mayer-Kress G.,(ed.)
Dimensions and Entropies in Chaotic Systems
Proc. Internat. Workshop Pecos River Ranch, New Mexico, (11-16 sept 1985), Part II
Springer-Verlag, Berlin, 1986.

3 Voll M.
Analyse des données (3ième ed.)
Economica, 1985, Paris.

4 Lebart L., Salem A.
Statistique textuelle
Dunod, 1994, Paris.

5 Hill M.O.
Diversity and Evenness: a unifying notation and its consequences,
Ecology,1973, Vol.54,No.2 p427-433.

6 Daget P.
Le nombre de diversité de Hill : un concept unificateur dans la théorie écologique,
Oecol. Gener., 1980, Vol.1, No.1 p51-70.

7 Mandelbrot B.
Fractals-Forms, Chance and Dimension
Freeman (ed.), 1977, San-Francisco.

8 Bergé P., Pomeau Y., Vidal Ch.
L'Ordre dans le Chaos
Hermann (ed.), 1988, Paris.

9 Van Raan A.F.J.
Fractal dimension of cocitation,
Nature, 18 october 1990, Vol.347 , p626.

10 Yaglom A.M., Yaglom I.M.
Probabilité et Information
Traduction Mercouroff W.
Dunod, 1959, Paris.

11Camel Y.
Probabilités Théorie et Application
Chap 9, p.109,
Eyrolles, 1988, Paris.

12 Quoniam L.
Bibliométrie sur des références bibliographiques: méthodologie
in La Veille Technologique (Dou H. Ed.), Dunod, 1992, Paris.

13 Mandelbrot B.
Contribution à la théorie mathématique des jeux de communication,
Publ. Inst. Statist. Univ., 1953, Paris 2, p1-124.

14 Rostaing H.
Veille Technologique et bibliométrie : Concepts, Outils, Applications
Thèse soutenue le 13 janvier 1993, Univ. Aix-Marseille III.

15 Badii R.
Conservation laws and thermodynamic formalism for dissipative dynamical systems,
Thèse 1987 universität Zürich.

16 Egghe L.
Bridging the gap: conceptual discussions on infometrics,
Scientometrics, may 1994, Vol.30, N°1, p35.

17 Fairthorne R.
Citation classic - empirical hyperbolic distributions (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction,
Current contents/Social & Behavioral Sciences, 1987, N°3, p16.