Utilisation de la théorie mathématique de la communication en sciences de l'information

Jean-Bernard Marino

THÈSE

PRÉSENTÉE

À L' ÉCOLE DES HAUTES ÉTUDES EN SCIENCES SOCIALES

POUR OBTENIR

LE TITRE DE DOCTEUR EN 3e CYCLE

Spécialité: Sciences de l'information

par

Jean-Bernard MARINO

UTILISATION DE LA THEORIE MATHEMATIQUE DE LA COMMUNICATION EN SCIENCES DE L'INFORMATION

Soutenue le 12 janvier 1984 devant la commission d'examen:

J. ARSAC, Président

M. BARBUT
B. BOUCHON } Examinateurs
J. MEYRIAT

Je tiens à exprimer ma respectueuse reconnaissance à Monsieur le Professeur J. MEYRIAT, qui a accepté de diriger ce travail et auprès de qui j'ai constamment trouvé aide et encouragements.

Que Monsieur le Professeur J. ARSAC, qui a accepté la présidence du Jury et qui a bien voulu aiguiller mes premiers pas dans le monde de la recherche trouve ici l'expression de ma reconnaissance,

ainsi que Monsieur le Professeur M. BARBUT qui m'a fait l'honneur de juger ce travail.

Cette thèse n'aurait pas vu le jour sans l'aide attentive et constante de Mademoiselle B. BOUCHON, Chargée de Recherche, qui m'a permis de franchir nombre d'obstacles théoriques et pratiques, et de bénéficier du financement et des facilités de travail du Groupe de Recherche "Structures de l'information". Qu'elle trouve ici l'expression de ma pro-fonde gratitude.

Ma gratitude s'adresse également à Madame MASBOU, Ingénieur informaticien du Groupe de Recherche et à Monsieur BONNO, Maître-Assistant à la Faculté des Sciences de Reims, qui ont assuré la pro-grammation de la partie expérimentale.

Je tiens aussi à remercier tout particulièrement le S.R.I. de l'Agence Spatiale Européenne qui, en la personne de Monsieur P. LEQUAIN, a concouru au financement de la présente recherche.

Il m'est enfin agréable de remercier Madame A.-M. LAURENT, Bibliothécaire-adjointe responsable du prêt inter-bibliothèques à la B.U. section Sciences et Techniques de Reims qui plus d'une fois m'a procuré l'introuvable.

TABLE DES MATIERES

INTRODUCTION

Le développement des télécommunications a imposé aux industriels la mise au point de systèmes de transmission de l'information à la fois fidèles, rapides et peu coûteux.

Les recherches entreprises entre les deux guerres amenèrent à définir une mesure de l'information transmise. L'idée fut développée par Claude Elwood SHANNON pour déboucher, en 1948, sur une théorie mathématique de la communication.

Cette théorie se révéla d'emblée très féconde dans son domaine d'ori--gine. Mais très vite elle intéressa de nombreux chercheurs de toutes disciplines, comme en témoigne la tenue dès l'été 1950, à Londres, du premier symposium international sur la "théorie de l'information".

Ont concouru à ce succès :

- une formulation mathématique assez générale pour encourager l'appli--cation de la théorie à divers domaines, comme le suggéra W. WEAVER en 1949 ;
- une terminologie à la fois séduisante et ambiguë (information, entropie, redondance, etc.) semblant promettre une clé à diverses interrogations du monde scientifique.

Les professionnels de la documentation, de leur côté, cherchant à asseoir leur discipline sur des bases théoriques et méthodologiques solides, ont tenté d'appliquer la théorie mathématique de la communication à divers processus relatifs au traitement de l'information scientifique et technique. Les résultats n'ont cependant pas pleinement répondu à leurs attentes, sans pour autant que les obstacles rencontrés aient toujours été clairement perçus.

Le présent travail a pour objet

- de recenser et de commenter les utilisations faites jusqu'à maintenant de la théorie mathématique de la communication et de ses développe--ments en sciences de l'information,
- de tenter d'énoncer d'une façon générale les données du problème ainsi délimité,
- de proposer une problématique adaptée à la fois à la structure des bases de données et à l'examen de divers problèmes documentaires concrets,
- enfin d'aborder les questions de signification de l'information sous l'éclai--rage bien particulier des disciplines documentaires.

Le premier paragraphe sera consacré à un rappel des principaux éléments de la théorie mathématique de la communication.

Nous supposerons tout au long de ce travail que le lecteur est familiarisé avec le contexte de la bibliographie scientifique tant manuelle qu'automa--tisée.

QUELQUES ELEMENTS DE LA THEORIE MATHEMATIQUE DE LA COMMUNICATION

Nous nous limiterons, sauf exceptions, aux messages de type discret (discontinus), laissant de côté le cas continu. Nous utiliserons le plus souvent la notation adoptée par H. ATLAN dans son remarquable ouvrage de synthèse pour biologistes (5), notation inspirée de celle de H. QUASTLER (89).

I. DEFINITION DE LA QUANTITE D'INFORMATION

A. NOTION DE QUANTITE D'INFORMATION

Les travaux de C. E. SHANNON (98) sont nés de l'étude au sein de la Compagnie Bell de problèmes particuliers aux télécommunications (télégraphe, téléphone, radio, télévision). Ils aboutissent à une théorie de la communication essentiellement mécaniste, c'est-à-dire ne tenant compte en aucune façon de la signification des messages transmis. Le problème à résoudre est purement technique : quel codage optimal peut-on appliquer à des messages choisis dans un ensemble connu afin de les transmettre le plus fidèlement et le plus rapidement possible en présence de parasites ? SHANNON définit la quantité d'information contenue dans un message comme une fonction de la fréquence d'utilisation des différents symboles composant le message.

Pour des jeux de symboles suffisamment significatifs du point de vue statistique, il est d'usage d'assimiler la fréquence à la probabilité d'apparition des symboles.

Soient N symboles différents caractérisés chacun par une probabilité d'apparition p(i), i étant compris entre 1 et N. La quantité d'information moyenne par symbole d'un message utilisant ces N symboles est :

$$H = - \sum_{i=1}^{N} p(i) \log_2 p(i) \ , \quad \text{avec} \sum_{i=1}^{N} p(i) = 1,$$

que nous noterons plus simplement par la suite

$$- \sum_{i} p(i) \log_2 p(i).$$

H se présente comme la somme des quantités d'information spéci-fiques locales en i, $H(i) = - \log_2 p(i)$ attachées à chaque symbole, pondérées par la probabilité p(i) d'apparition de ce symbole :

$$H = \sum_i p(i)\, H(i).$$

La mesure de H ne traite donc que la probabilité d'apparition des symboles parmi l'ensemble des N symboles utilisables. Elle rend compte de l'homogénéité d'un choix statistique.

Première remarque :
La fonction H est analogue à la formule de l'entropie thermodyna-mique de GIBBS - BOLTZMANN (24) :

$$S = - k \sum_i p_i \log p_i$$

où p_i est la probabilité de présence d'une molécule de gaz dans un micro-état i de l'espace des phases comprenant 6 dimensions (3 pour la position, 3 pour la quantité de mouvement). L'analogie avec la fonction H a conduit SHANNON à baptiser H l'entropie du message.
Les équations étant les mêmes à une constante près *, les scienti-fiques se sont interrogés sur la nature de la parenté information - entropie : similitude formelle ou traduction d'une réalité physique ? Il en est résulté une foule fort embarrassante d'interprétations. Quoi qu'il en soit, l'identification pure et simple de l'information à l'entropie est exclue pour une raison d'unité : alors que H est une grandeur sans dimension, S a la dimension de la constante de BOLTZMANN k et s'exprime en énergie par unité de température. Cette différence est liée à une incompatibilité conceptuelle plus générale : M. MUGUR-SCHACHTER fait remarquer que H apparaît comme un concept probabiliste abstrait interprétable dans un deuxi-ème temps en fonction de diverses situations, tandis que S est lié dès l'origine à la description d'une classe de situations physiques (83).

* *J. MAX (75) présente l'entropie de GIBBS - BOLTZMANN comme négative ou nulle, la formule de S étant précédée du signe +.*

Enfin, si on admet l'hypothèse de BRILLOUIN (11) selon laquelle il y a variation d'entropie d'un système thermodynamique lors de l'acquisition d'information, les variations simultanées d'entropie et de quantité d'information sont de signes contraires, l'information apparaissant comme l'équivalent d'une "néguentropie".

Deuxième remarque :
SHANNON définit la grandeur "quantité d'information" sans définir la notion d'information. Il s'agit là d'une démarche qui peut sembler déroutante. Elle s'inscrit en fait dans la logique de la recherche scientifique, dont l'objet principal est l'étude des phénomènes et pour laquelle les questions du type "qu'est-ce que l'information ?" se révèlent absurdes, comme le montre M. MAZUR (76).

B. NECESSITE D'UNE FORME LOGARITHMIQUE

Quand on considère deux événements 1 et 2 indépendants de proba-bilité p(1) et p(2), la probabilité d'obtenir un couple 1 et 2 simulta-nément est le produit p(1) p(2).
La quantité d'information spécifique apportée par 1 est une fonction f[p(1)] ; de même f[p(2)] pour 2, et f[p(1,2)] pour 1 et 2 simultanés. Comme il est logique et pratique d'envisager une fonction f telle que f[p(1,2)] = f[p(1)] + f[p(2)], on a choisi $f = -\log_2 p$.

Première remarque :
La base 2 du logarithme permet d'obtenir H = 1 pour p = 1/2. Une telle valeur de H correspond à la survenue d'un événement parmi deux événements équiprobables (jeu de pile ou face, par exemple). H apparaît de ce point de vue comme une mesure de l'incertitude levée par la survenue de cet événement.

Deuxième remarque :
Le signe moins est nécessaire pour que H soit positif : une probabilité étant inférieure ou égale à 1, son logarithme est négatif.

C. CONDITION D'APPLICATION DE LA FONCTION H

La fonction H caractérise un ensemble de messages ayant en com-
-mun :
1) le même jeu de N symboles différents et indépendants les uns des autres,
2) la même distribution de probabilités p(i),
même s'ils diffèrent par leur longueur ou l'ordre de succession des symboles.
De plus, et sans entrer dans le détail, ATLAN (5) fait remarquer que "l'émission de suites de symboles par la source doit constituer un pro-
-cessus stochastique stationnaire et ergodique, ce qui signifie que le régime de probabilités est le même tout le long des séquences de sym-
-boles et aussi qu'il n'existe pas de variations surajoutées périodiques ou autres qui pourraient permettre de diviser l'ensemble de messages en processus indépendants".

II. TRANSMISSION DE L'INFORMATION

L'optimisation de la transmission de l'information par une voie de télé-
-communications est à l'origine des travaux aboutissant à la théorie ma-
-thématique de la communication (T.M.C.). Il est donc logique de trouver dans cette théorie une formulation purement objective et technique des phénomènes et objets étudiés. W. WEAVER (109) définit trois types de problèmes dans le domaine des communications :

1) Les problèmes techniques : avec quelle exactitude les symboles utilisés peuvent-ils être transmis ?
2) Les problèmes sémantiques : avec quelle précision les symboles trans-
-mis véhiculent-ils la signification recherchée ?
3) Les problèmes d'efficacité : avec quelle efficacité le message porteur de signification une fois reçu affecte-t-il le destinataire de la façon recherchée ?

Seuls les problèmes techniques sont abordés par SHANNON, particulière--ment sous l'angle du codage.

A. VOIE DE COMMUNICATION

Les messages sont transmis d'une source à un destinataire à travers une voie de communication représentée schématiquement ci-dessous :

Le message, ne pouvant en général être transmis tel quel, est codé afin de parcourir la voie puis décodé afin d'être restitué au desti--nataire.

B. CODAGE

L'opération du codage permet de représenter d'une façon biunivoque le système de symboles du message par un autre système de symboles. Pour des raisons de technique électronique, la représentation se fait en général en système binaire, avec les symboles conventionnels 0 et 1. Les messages codés apparaissent ainsi comme des suites de 0 et de 1.

SHANNON démontre que H représente le nombre minimum moyen de symboles binaires à utiliser par symbole de départ du message pour effectuer le codage (théorème du codage sans bruit).

III. TRANSMISSION DE L'INFORMATION DANS UNE VOIE AVEC BRUIT

Dans l'hypothèse plus générale où la transmission du message se fait d'une façon imparfaite - avec bruit - il y a perte d'information et la formule H doit tenir compte des probabilités conditionnelles de tran--sition entre les symboles du message d'entrée et ceux du message de sortie.

A. PROBABILITES CONDITIONNELLES

Si x_i est un symbole du message d'entrée X et y_j un symbole du message de sortie Y, on définit la probabilité de transition $p(j|i)$ de x_i à y_j comme la probabilité de trouver y_j dans le message de sortie sachant que x_i se trouve à la place correspondante dans le message d'entrée.

Dans le cas d'une transmission sans bruit :
$p(j|i) = p(i|j) = 1$ pour $i = j$,
$p(j|i) = p(i|j) = 0$ pour $i \neq j$.

On peut relier la probabilité conditionnelle à la probabilité conjointe de trouver à la fois x_i et y_j :
$p(i,j) = p(i)\ p(j|i) = p(j)\ p(i|j)$.
Dans le cas particulier de variables indépendantes, on retrouve le produit $p(i,j) = p(i)\ p(j)$, avec $p(j|i)$ égal à $p(j)$ et $p(i|j)$ égal à $p(i)$.

B. QUANTITE D'INFORMATION

Soit un message d'entrée X et un message de sortie Y.
La quantité d'information du message d'entrée est :

$$H(X) = - \sum_i p(i) \log_2 p(i),$$

celle du message de sortie est :

$$H(Y) = - \sum_j p(j) \log_2 p(j).$$

On peut définir une quantité d'information sur X et Y à la fois :

$$H(X,Y) = - \sum_{i,j} p(i,j) \log_2 p(i,j).$$

Si les variables sont indépendantes, $H(X,Y) = H(X) + H(Y)$.
En présence de bruit, $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
$< H(X) + H(Y)$,

avec $H(Y|X) = - \sum_{i,j} p(i)\, p(j|i) \log_2 p(j|i)$

et $H(X|Y) = - \sum_{i,j} p(j)\, p(i|j) \log_2 p(i|j)$.

L'incertitude attachée à l'occurrence conjointe de X et Y est égale à celle attachée à l'un des messages plus l'incertitude conditionnelle attachée à l'autre quand le premier est connu.

C. FONCTIONS CARACTERISTIQUES D'UNE VOIE AVEC BRUIT

Le déficit d'information ainsi provoqué par la dépendance des variables permet de définir plusieurs fonctions :

1) Quantité d'information transmise dans la voie :

$$T(X;Y) = T(Y;X) = H(X) - H(X|Y)$$
$$= H(Y) - H(Y|X)$$
$$= H(X) + H(Y) - H(X,Y).$$

T mesure le déficit d'incertitude sur Y quand X est connu. C'est une relation symétrique exprimant une interdépendance du message d'entrée et du message de sortie.

2) Ambiguïté :

$H(Y|X)$ mesure la quantité d'information de la sortie quand l'entrée est déterminée.

3) Equivocation :

$H(X|Y)$ mesure la quantité d'information de l'entrée quand la sortie est connue.

Remarque :

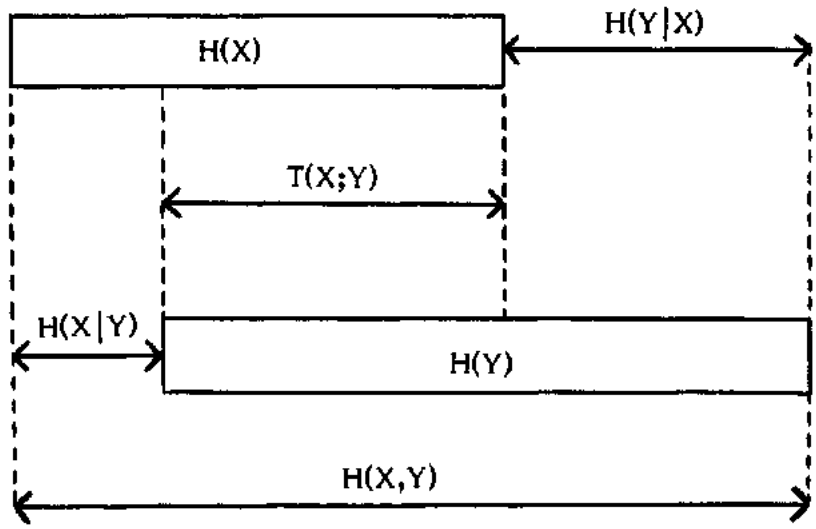
Dans le cas d'une transmission sans bruit,

$H(Y) = H(X) = H(X,Y) = T(X;Y)$

$H(Y|X) = H(X|Y) = 0$.

Il y a transmission intégrale du message d'entrée.

Dans le cas général, les différentes fonctions décrivant la transmission peuvent être représentées graphiquement selon le schéma de H. QUASTLER :



D. REDONDANCE

La création d'une redondance est susceptible de pallier le déficit d'information dû au bruit. L'opération consiste à ajouter au message d'entrée des symboles supplémentaires permettant de détecter ou de contrebalancer une altération du message. La quantité d'informa--tion du message par symbole est ainsi réduite. La redondance mesure la perte relative :

$$R = \frac{H_{max} - H_r}{H_{max}} = 1 - \frac{H_r}{H_{max}} ,$$

H_r étant la quantité d'information du message redondant et H_{max} la quantité d'information du même message avant traitement.

E. CAPACITE D'UNE VOIE

La capacité C d'une voie est définie comme la quantité d'information maximum que peut transmettre cette voie, c'est-à-dire le maximum de la fonction T. Si la voie est sans bruit, C = H(X).

SHANNON démontre que si l'on veut transmettre un message H dans une voie de capacité C, il existe une méthode de codage optimale telle que :

- si $H \leqq C$, on peut faire tendre l'équivocation vers zéro,
- si $H > C$, on peut faire tendre l'équivocation vers H - C.

Si le "théorème du codage avec bruit" démontre qu'on peut transmet--tre avec une erreur tendant vers 0 un message de quantité d'infor--mation $H \leqq C$ dans une voie parasitée, il n'indique pas quel codage permet d'y arriver. Cependant, l'apport de ce théorème est important dans la mesure où il fixe une limite à H en deçà de laquelle la pré--sence de bruit, contrairement à ce que l'on pensait auparavant, ne rend pas les erreurs de transmission inévitables.

UTILISATIONS FAITES DE LA THEORIE MATHEMATIQUE DE LA COMMUNICATION EN SCIENCES DE L'INFORMATION

On peut diviser grossièrement en six groupes les travaux auxquels la T.M.C. a donné lieu dans le domaine des sciences de l'information :

1) Hypothèses d'application.
2) Conception des cartes perforées.
3) Evaluation des performances d'un système documentaire.
4) Indexation automatique.
5) Diversité d'une population bibliographique.
6) Stockage de données en ordinateur.

Il conviendra d'ajouter un septième groupe de travaux divers couvrant certains domaines plus ou moins étroitement connexes à la T.M.C..

Ce chapitre consacré aux utilisations faites jusqu'à maintenant de la T.M.C. en sciences de l'information n'entend pas se limiter à un strict inventaire. La démarche que nous avons adoptée est double :

- Passer en revue les diverses utilisations selon un classement logique, avec le constant souci de ne pas alourdir l'exposé par une ré-écriture des tra--vaux analysés. Cet examen a pour but de mettre en lumière les points principaux des travaux, points plus ou moins clairement explicités par les auteurs eux-mêmes.
- Emettre sur ces travaux un certain nombre d'éclaircissements et de critiques personnels rendus nécessaires par certaines lacunes dont la plus répandue nous semble être le contraste entre la longueur des explications apportées à chaque étape de raisonnement et l'importance de l'étape.

I. HYPOTHESES D'APPLICATIONS

Nombreuses sont les allusions, dans les publications de sciences de l'infor--mation, à la théorie de SHANNON. Il n'en découle pas automatiquement que les auteurs de ces publications traitent de la T.M.C.. Il semble bien qu'on puisse attribuer ce décalage à un malentendu terminologique dû à l'emploi de l'expression "théorie de l'information" qui, mal précisée, peut aussi bien désigner la théorie de SHANNON et ses développements que des études théoriques sur la notion d'information. Dans ce dernier cas, la référence à la T.M.C. sert de balisage afin d'assurer le lecteur

que l'existence de cette théorie particulière a bien été perçue et prise en compte dans la réflexion.

Seules les hypothèses axées en propre sur la T.M.C. sont ici passées en revue.

A. LES HYPOTHESES DE FAIRTHORNE

1) Les trois applications possibles.

Lors d'un symposium (36), R. A. FAIRTHORNE présenta en 1960 une communication développant, à la lumière de la T.M.C., un certain nombre de questions théoriques relevant, au sens large, de la théorie de l'information.

Le premier mérite de l'auteur est de comparer un système docu--mentaire à un système thermodynamique ouvert modélisable en une succession de systèmes fermés presque identiques.

Une telle analogie ouvre ensuite la voie à un inventaire des objets auxquels on pourrait appliquer "des considérations relevant de la théorie de l'information" - formule volontairement prudente :

a) Des séquences historiques de documents, établies semble-t-il par des relations de citation.

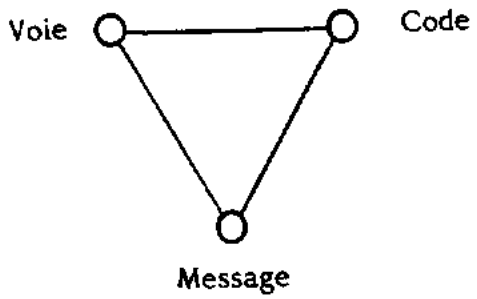
b) Des caractéristiques "spatiales", établies semble-t-il par examen des caractéristiques textuelles communes à plusieurs membres.

c) Des flux de caractéristiques informationnelles apparaissant lors de la conversion de séquences en ensembles stockés.

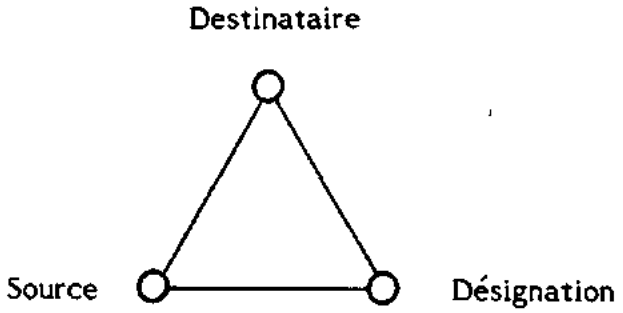
2) Les flux d'information.

FAIRTHORNE reviendra en détail quelques années plus tard sur la troisième hypothèse : dans le contexte des sciences de l'infor--mation, il développe un certain nombre de considérations à la fois théoriques et qualitatives sur la modélisation des processus de communication (37).

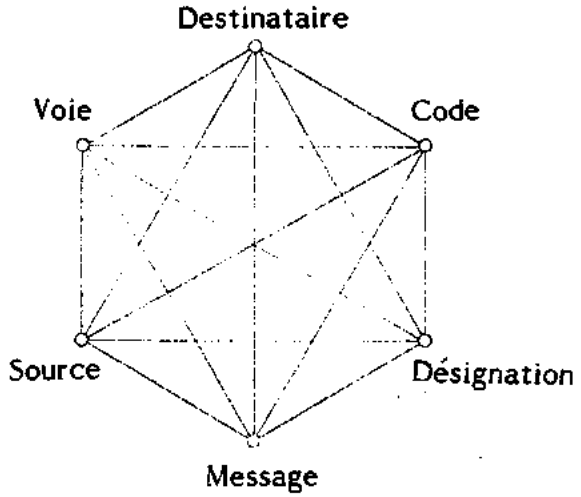
Il limite le modèle de SHANNON au phénomène de "transfert de signal", dans lequel n'interviennent que 3 éléments : la voie, le code, le message.



Ce modèle est complété par une seconde triade symbolisant le "discours" et composée de la source, du destinataire et de la dési-gnation.



La superposition de ces deux triades permet de visualiser un pro-cessus de communication type dans le domaine documentaire, que FAIRTHORNE désigne par "notification" :



Par ce processus, un utilisateur reçoit des messages après spéci-
-fication à un intermédiaire de ses besoins en termes de caracté-
-ristiques de document.
Par soustraction de 3 sommets de ce processus-type, FAIRTHORNE identifie un certain nombre de processus élémentaires :

Acheminement : Voie - Source - Destinataire,
Transmission : Voie - Source - Code,
Réception : Voie - Destinataire - Code,
Classement : Voie - Message - Désignation,
Classification : Code - Message - Désignation,
Attribution : Code - Message - Source,
Adressage : Code - Message - Destinataire,
D.S.I. : Désignation - Message - Destinataire,
Paternité auteur : Désignation - Message - Source,

Si on reconnaît la réalité de ces distinctions, il apparaît que la correspondance établie entre ces processus élémentaires et leur interprétation graphique est conditionnée par le caractère arbitraire et simplificateur des prémisses : cherchant à dépouiller les proces-
-sus au maximum, FAIRTHORNE élimine par exemple la source et la destinataire dans le modèle de transfert de signal. De même, la seconde triade (discours) isole la désignation du message. Ces coupures, tout naturellement, se retrouvent dans le tableau typo-
-logique des processus élémentaires et, éventuellement, dans d'au-
-tres graphes résultant de l'accolement de 2 processus élémentaires ayant un ou plusieurs éléments communs.
FAIRTHORNE ne cherche pas à quantifier les relations au sein des différents processus élémentaires - ce qui, cependant, pourrait justifier des simplifications assez gênantes quand on demeure sur le plan qualitatif. Il se borne simplement à signaler la validité possible de la mesure de type entropique, avec deux restrictions de principe :

a) La validité formelle n'entraîne pas nécessairement l'utilité ;
b) Ce qui serait traité et mesuré ne serait pas forcément de l'infor-
-mation au sens courant, mais une quantité pouvant renseigner selon le cas sur différents phénomènes.

B. CLASSIFICATION AUTOMATIQUE

M. E. MARON, de la Rand Corporation, propose une méthode de classification automatique basée sur un calcul de probabilités condi--tionnelles et justifiée par une fonction entropique (72). Une telle étude peut être rattachée à la deuxième hypothèse de FAIRTHORNE. L'expérimentation de MARON est appliquée à une collection de résumés analytiques dans le domaine de l'informatique - automatique.

1) Indexation manuelle.

La première phase de l'opération consiste à ranger un lot de 260 "documents" (notices) dans une ou plusieurs "catégories de sujets" (32 classes simples de classification).

Des mots-clés sont attribués aux documents selon certaines règles arbitraires :

a) la forme singulière et la forme plurielle d'un même mot consti--tuent deux mots différents,

b) deux orthographes différentes d'un même mot déterminent également deux mots distincts,

c) sont éliminés les articles, prépositions et conjonctions,

d) sont éliminés les mots banals dans le domaine étudié (ordinateur, système, donnée, machine, etc.),

e) sont éliminés les mots n'apparaissant dans le corpus entier qu'une ou deux fois.

2) Spectre de mots.

La deuxième phase fait apparaître pour chacun des 1000 mots retenus le nombre d'occurrences dans chaque catégorie. Pour chaque mot se dessine un "pic" correspondant à un grand nombre de présences dans une catégorie. Un mot sans pic est rejeté comme non représentatif. Ne subsistent après appariement à l'ensemble des catégories que 90 mots-clés.

3) Prévisibilité de classement.

MARON définit l'incertitude quant au classement d'un document dans une catégorie par deux fonctions H et H' :

$$H = - \sum_{j=1}^{32} p(C_j) \log_2 p(C_j)$$

où $p(C_j)$ est la probabilité a priori pour un document d'être classé dans la catégorie C_j ;

$$H' = - \sum_{j=1}^{32} p(C_j | W_i) \log_2 p(C_j | W_i)$$

où $p(C_j | W_i)$ est la probabilité, dans le cas où le i ème mot apparaît dans le document, que celui-ci appartienne à la catégorie C_j. L'incertitude levée par l'attribution d'un mot-clé est déterminée par la différence entre H' et H. On peut ainsi, de deux mots W_1 et W_2, connaître celui qui lève le plus d'incertitude et, donc, constitue un meilleur mot-clé pour la catégorie donnée.

Ces formules sont présentées par MARON pour appuyer son raisonnement et justifier l'usage de probabilités conditionnelles. Elles ne sont cependant pas utilisées pour le calcul de concordance entre mots et catégories. Cette concordance est en effet obtenue simplement par la probabilité qu'un document contenant les mots-clés W_k, W_m, ..., W_s appartiennent à la catégorie C_j : nombre d'attribution :

$$p(C_j | W_k . W_m . \ldots . W_s) \# k \; p(C_j) \; p(W_k | C_j) \; p(W_m | C_j) \ldots p(W_s | C_j)$$

avec $\sum_{j=1}^{32} p(C_j | W_k . W_m . \ldots . W_s) = 1$

et $p(W_i | C_j) = \dfrac{p(C_j | W_i) \; p(W_i)}{p(C_j)}$ (théorème de BAYES).

Les $p(C_j)$ sont obtenus en comptant le nombre de documents en j ème catégorie, divisé par le nombre total de documents.

Les $p(W_i | C_j)$ sont obtenus en comptant le nombre d'apparitions du i ème mot qui appartient aux documents indexés en j ème catégorie, divisé par le nombre total d'apparitions de mots dans tous les documents de la j ème catégorie.

4) Résultats.

Les nombres d'attribution calculés sur la base du lot de 260 docu--ments assignent une bonne catégorie dans 85 % des cas. Sur un lot différent de celui ayant servi de base de calcul, on tombe à 52 %, ce qui ne permet pas de conclure étant donné la faiblesse de l'échantillon de départ.

5) Extension de la démarche de MARON.

On trouve, une douzaine d'années plus tard, dans les publications de A. ANDREEWSKY, C. FLUHR et J. RAMBOUSEK (2) (3) une application analogue portant sur des probabilités conditionnelles de "tirer" un document - et non plus un code de classification - si un mot donné est présent.

Les études entreprises par ces auteurs déboucheront sur une fonc--tion de poids sémantique d'un mot M_i pénalisant les mots généraux présents dans l'ensemble des documents et donc peu discriminants (40). Cette fonction est d'autant plus élevée que l'entropie $H(M_i)$, calculée sur l'ensemble des N documents de la collection, est plus faible :

$$H(M_i) = - \sum_{j=1}^{N} p(D_j | M_i) \log_2 p(D_j | M_i).$$

Outre l'application de l'hypothèse de MARON au rapport mot / document, ces études présentent l'avantage d'aborder en même temps le problème sous un angle linguistique. Elles permettent actuellement de mettre en oeuvre un système documentaire opéra--tionnel et évolutif - SPIRIT.

C. LOI DU MOINDRE EFFORT

Les réflexions de P. ZUNDE (112) prennent leur source dans les tra--vaux de linguistique mathématique de B. MANDELBROT (71) qui utilisent la théorie de l'information et aboutissent, sous réserve de certaines hypothèses, à une relation entre la fréquence relative d'un mot x_i et son rang dans la liste des fréquences décroissantes :

$f(x_i) = C\,[r(x_i) + V]^{-B}$; C, V, B sont des constantes, V peut être nul.

Partant de cette application particulière, l'auteur soulève une question générale : peut-on unifier en un modèle unique la théorie de l'infor--mation et diverses lois empiriques du type hyperbolique rencontrées en sciences de l'information et dans certains domaines connexes ? A savoir :

1) Loi de ZIPF : occurrence de mots.

2) Loi de BRADFORD : dispersion de la littérature périodique.
   Remarque : Traditionnellement, la loi de BRADFORD n'est pas directement présentée sous une forme hyperbolique dans la mesure où interviennent non pas une fréquence et un rang mais un "multi--plicateur" et un nombre de périodiques.
   Si une collection de périodiques est divisée en 3 groupes constitués d'un noyau de n_1 titres, d'un groupe "intermédiaire" de n_2 titres et d'un groupe "lointain" de n_3 titres offrant le même nombre d'arti--cles dans le domaine considéré, on constate que, grossièrement, $n_3 = s_b\, n_2 = s_b^2\, n_1$, s_b étant le multiplicateur.

3) Loi de LOTKA : productivité des chercheurs.

4) Loi de SKINNER : association de mots.

5) Loi de taille du vocabulaire : nombre de mots différents d'un texte.

6) Loi de temps de réponse : réaction à des signaux.

7) Loi d'exhaustivité d'indexation : influence du nombre d'indexeurs.

D'autres exemples de lois empiriques analogues sont exposés dans l'article de revue de FAIRTHORNE (38).
D'après l'auteur, un approfondissement des études consacrées aux divers aspects de la loi du moindre effort devrait permettre d'affermir les bases empiriques des processus d'information et d'établir des rapports avec la théorie de l'information.

Le souhait de ZUNDE de voir progresser la compréhension des diverses lois du type hyperbolique semble en effet tout à fait légitime. Un des premiers obstacles à surmonter tient à la diversité des modes de présentation, d'un auteur à l'autre, de certaines de ces lois. Un effort de formalisation permettrait de mieux exploiter une masse abondante d'études.

II. CONCEPTION DE CARTES PERFOREES

E. GARFIELD (44) a expérimenté dans les années 1950 l'indexation de documents biomédicaux à l'aide de cartes perforées.
Afin d'optimiser le traitement des cartes, l'auteur s'est appuyé sur le principe du codage de l'information transmise dans une voie sans bruit : il fait dépendre le nombre de perforations de la fréquence d'utilisation p(i) des descripteurs dans l'ensemble de la collection. Les descripteurs les plus fréquents reçoivent ainsi le codage le plus court. La formule :

$$H = - \sum_i p(i) \log_2 p(i)$$

fixe le nombre minimum de symboles binaires à utiliser par descripteur.

Remarque :
L'étude de GARFIELD répond au souci d'améliorer un système à présent obsolète. Elle est toutefois intéressante dans la mesure où elle pose avec bon sens le problème général de l'utilisation de la T.M.C. en sciences de l'information, dans la perspective tout à fait actuelle d'utilisation d'une base de données.

III. EVALUATION DES PERFORMANCES D'UN SYSTEME DOCUMENTAIRE

A. PREMIERE APPROCHE

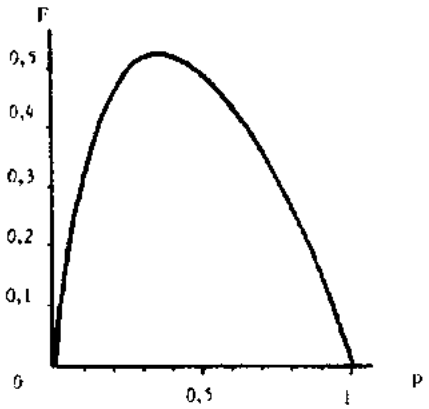
La communication de R. M. HAYES (52) peut être considérée comme un premier essai de délimitation du problème : une probabilité p(x) attachée au document x est introduite sans être nettement précisée ; un paramètre de pertinence r(x) comparable à un facteur d'utilité (cf. p. 115-116) et égal à 1 dans le cas d'une transmission sans bruit com--plète cette probabilité dans l'équation de "signification moyenne" S :

$$S(X) = -\sum_{x=1}^{N} r(x)\ p(x)\ \log\ p(x).$$

Bien que ne débouchant sur aucune application concrète, l'étude de HAYES a le mérite de proposer une évaluation quantitative de perfor--mance d'un système documentaire à l'aide de la T.M.C..

HAYES précisera ultérieurement (54) la probabilité p(x) afin de tenter d'établir une relation entre l'utilisation de documents et leur "quantité d'information". Pour cela, r(x) est la pertinence du docu--ment correspondant à la notice x et p(x) la probabilité d'apparition a priori de cette notice.
HAYES reconnaît que les objectifs initiaux recherchés par l'emploi de ce modèle n'ont pas été atteints.
Remarque :
Il apparaît intuitivement que, pour chaque document pris isolément, une relation peut unir pertinence et probabilité d'apparition a pos--teriori. On voit mal cependant comment une fonction entropique pondérée eût pu donner des résultats significatifs pour une collection de documents.
En premier lieu, sur quelle base solide à la fois sur le plan théorique et sur le plan de la pratique documentaire établir des probabilités a priori ?
En second lieu, et plus fondamentalement, la formule entropique est particulièrement adaptée à la description globale de la dispersion d'un ensemble d'objets, c'est-à-dire des disparités qu'on peut observer

au sein d'une population, comme nous le verrons en IV. L'entropie maximum correspond au désordre maximum, les objets étant dispersés uniformément dans tous les états disponibles. Cependant lorsqu'on passe du global au particulier pour considérer chaque document l'un après l'autre, la forme convexe du produit - p(x) log p(x) fausse tout essai de mise en rapport biunivoque entre pertinence et probabilité. Deux valeurs très différentes de p(x) peuvent apporter un produit de même valeur, comme le montre le graphe de la courbe
$F(p) = - p \log_2 p$:



Le produit - r(x) p(x) log p(x) ne pouvant avoir de signification con--crète claire, la sommation ne peut à plus forte raison traduire une réalité observable.

B. ETUDE D'UNE REPARTITION OPTIMALE DE DESCRIPTEURS

La publication de P. ZUNDE et V. SLAMECKA ([113]) utilise l'équation de transmission d'information dans une voie sans bruit.
Un traitement mathématique assez élaboré permet de déterminer une distribution optimale des descripteurs d'un système documentaire par rapport au nombre de documents auxquels ils renvoient.
Pour cela, les auteurs définissent une probabilité p(t) d'occurrence d'un groupe de descripteurs ayant en commun la propriété de renvoyer à un même nombre t de documents.
Le système d'équations :

$$\begin{cases} H(T) = -\sum_{t=1}^{n} p(t) \log p(t) \\ \sum_{t=1}^{n} p(t) = 1 \\ \sum_{t=1}^{n} t\, p(t) = \bar{t} \quad \text{(nombre moyen de numéros d'attribution par descripteur = constante)} \end{cases}$$

permet d'obtenir, par dérivation, une valeur

$$p(t) = \frac{(1 - \frac{1}{\bar{t}})^t}{(\bar{t} - 1)}$$

correspondant à H(T) maximum.

ZUNDE et SLAMECKA en déduisent des courbes d'évaluation de différents systèmes documentaires caractérisés par des distributions différentes des valeurs p(t) et montrant un écart plus ou moins grand de H(T) avec H(T)max.

Remarque :
Les auteurs ne précisent pas le mode de détermination de la valeur de chaque p(t) dans le système d'équations, ce qui permet d'interpréter cette probabilité au moins de deux façons différentes. Ce défaut de rigueur permet de douter de la compatibilité des deux contraintes qui accompagnent l'équation H(T), comme l'a souligné B. C. BROOKES (14).

C. PERFORMANCE DE RESSAISIE

Diverses études ont été entreprises afin d'évaluer les performances d'un système de ressaisie* par codage binaire des documents symbolisant la pertinence et la non pertinence.
Toutes utilisent le concept de transmission de l'information dans une voie avec bruit.

* *Ressaisie doit ici être considéré comme l'équivalent du terme anglais "retrieval". L'expression "information retrieval" est habituellement traduite par "recherche de l'information" (d'où le descripteur "recherche information" attribué par les indexeurs de PASCAL). Nous avons cependant préféré le néologisme "ressaisie" au terme "recherche" qui, même accolé à information, demeure insuffisamment lié au contexte des techniques documentaires. On pourra trouver une justification à cette démarche dans le fait que les services documentaires de l'Agence spatiale européenne ont la double dénomination "Information Retrieval Service - Service de Ressaisie de l'Information" bien que le sigle IRS soit le plus souvent employé.*

1) Etude de A. R. MEETHAM (77).

q questions posées au système comprenant n documents produisent l'émission de q n symboles binaires 1 ou 0, qui représentent les documents pertinents et non pertinents.

Le message émis correspond à l'évaluation d'un sélecteur parfait. Le système de ressaisie étant considéré comme une voie avec bruit, le message reçu est constitué d'une autre séquence de 1 et de 0 résultant d'un tri imparfait.

La correspondance entre les deux messages, qui caractérise la transmission, est décrite par un tableau de contingence :

	Pertinent	Non pertinent
Ressaisi	$A = \sum_q a$	$B = \sum_q b$
Non ressaisi	$C = \sum_q c$	$D = \sum_q d$

avec A + B + C + D = N = q n. Pour une question : a est le nombre de documents ressaisis et pertinents, b le nombre de documents ressaisis et non pertinents ; il reste dans le système c documents pertinents et d documents non pertinents non ressaisis.

MEETHAM définit arbitrairement, sur cette base, une mesure de l'information par message apparentée au "rayon informatif" de JARDINE et SIBSON (60) :

$$I = A \log \frac{NA}{(A+B)(A+C)}$$ information apportée par les documents pertinents ressaisis,

$$+ D \log \frac{ND}{(B+D)(C+D)}$$ information apportée par les documents non pertinents non ressaisis,

$$+ C \log \frac{NC}{(A+C)(C+D)}$$ information apportée par les documents pertinents non ressaisis,

$$+ B \log \frac{NB}{(A+B)(B+D)}$$ information apportée par les documents non pertinents ressaisis.

I est d'autant plus élevé que le système est plus performant. Partant des tables d'observation dressées par CLEVERDON (22), l'auteur calcule la valeur de I en fonction du "niveau de coordi--nation" (nombre minimum de termes d'indexation de la question devant être comparés aux descripteurs du document pour que celui-ci soit ressaisi).

Pour chaque langage d'indexation, un niveau de coordination opti--mum apportant le maximum d'information est déterminé.

Le calcul de I est également appliqué à la ressaisie de documents indexés par couplage bibliographique, et permet de dresser un tableau de correspondance avec la "force de couplage", c'est-à-dire le nombre de citations identiques dans le document d'entrée et le document ressaisi.

Remarques :

On définit l'unité de couplage bibliographique comme "un article de référence utilisé par deux publications" (63).

L'étude conclut à une qualité de ressaisie comparable à celle des langages conventionnels les plus performants.

2) Etude de J. BELZER (9).

BELZER s'attache à quantifier l'efficacité de l'évaluation de la pertinence d'un document à partir d'un produit de remplacement de ce document.

Cinq types de produits de remplacement sont proposés par une équipe de bibliothécaires à une population de 70 chercheurs, en réponse à une question bibliographique par chercheur :

1 - Citation bibliographique simple,
2 - Résumé analytique,
3 - Premier paragraphe,
4 - Dernier paragraphe,
5 - Premier et dernier paragraphes.

Les documents ressaisis par les bibliothécaires sont divisés en 5 groupes de tailles égales et proposés aux chercheurs sous forme d'un des 5 produits de remplacement.
Les chercheurs assignent à chacun de ces produits une estimation P = pertinent, $\overline{P}$ = non pertinent.
Quelques jours plus tard, les documents eux-mêmes sont proposés aux chercheurs qui évaluent définitivement leur pertinence par R = pertinent, $\overline{R}$ = non pertinent.
Un tableau peut être dressé décrivant la correspondance entre le nombre de documents estimés pertinents ou non et le nombre de documents retenus comme pertinents ou non, et ceci pour chacun des 5 groupes :

	P	$\overline{P}$
R	PR	$\overline{P}$R
$\overline{R}$	P$\overline{R}$	$\overline{P}\overline{R}$

Un deuxième tableau donnant les mêmes résultats sous forme de pourcentages assimilés à des probabilités sert de base au calcul de H(P), H(R) et H(R,P) avec :

$H(P) = - p(P) \log p(P) - p(\overline{P}) \log p(\overline{P})$,
$H(R) = -p(R) \log p(R) - p(\overline{R}) \log p(\overline{R})$,
$H(R,P) = - p(PR) \log p(PR) - p(\overline{P}\overline{R}) \log p(\overline{P}\overline{R})$
$- p(\overline{P}R) \log p(\overline{P}R) - p(P\overline{R}) \log p(P\overline{R})$.

L'auteur en déduit pour chaque produit de remplacement une valeur de $T(R;P) = H(R) + H(P) - H(R,P)$ qui permet de classer, dans l'ordre décroissant, les produits de remplacement selon : 5 - 4 - 3 - 2 - 1.

Remarques :
L'auteur conclut d'une façon fort optimiste que T mesure la quan--tité d'information significative des produits de remplacement.

C'est jouer sur les mots. La quantification permet certes de classer les produits de remplacement, mais ce classement s'appuie sur une observation extérieure. Il ne découle que des estimations du jury. On peut en effet penser qu'un premier paragraphe contient plus d'information qu'un résumé analytique, mais en cette matière l'usage fait de la T.M.C. n'apporte rien : elle ne donne qu'un reflet d'une estimation subjective. Le message est émis par le jury et reçu par le jury. Si on supprime l'observation du jury, il n'y a plus de quantification possible car les événements pris en compte n'appartiennent pas en propre aux documents étudiés. L'article est toutefois intéressant pour plusieurs raisons :

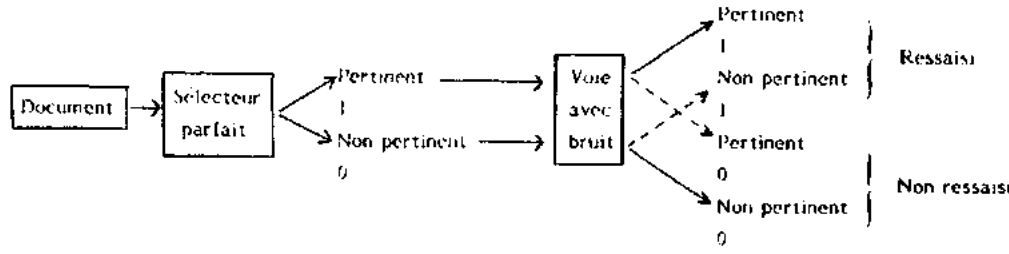
a) La méthode d'expérimentation est soigneusement décrite.

b) On peut se rendre compte de la quantité énorme de travail, de moyens et de bonnes volontés nécessaires à la réalisation d'une expérience basée sur la T.M.C.

c) Le classement par ordre d'efficacité des produits de remplace--ment recoupe l'expérience quotidienne des indexeurs : on cons--tate en effet fort souvent un décalage entre le contenu du résumé et celui du corps de la publication.

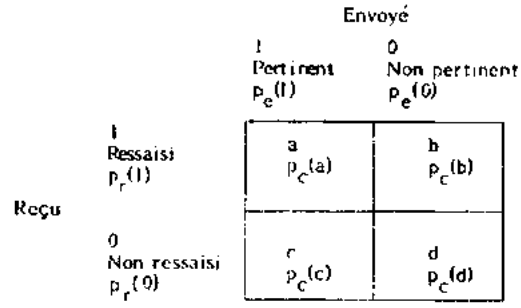
3) Etude de A. E. CAWKELL (19).

Comme dans l'étude de MEETHAM, le système de ressaisie est considéré comme une voie avec bruit, selon la figure ci-dessous :



Le message envoyé à la suite d'une question se présente également comme une suite de 1 et de 0 attribués par un sélecteur parfait.

Le message reçu est constitué d'une autre séquence résultant du tri imparfait schématisé ci-dessus.
La correspondance entre les deux messages est décrite par un ta-bleau de contingence faisant apparaître les probabilités de chaque situation :



L'information transmise est donnée par l'équation :
$T(X;Y) = H(X) + H(Y) - H(X,Y)$, soit

$$\begin{aligned} T(X;Y) = &- p_e(1) \log p_e(1) - p_e(0) \log p_e(0) \\ &- p_r(1) \log p_r(1) - p_r(0) \log p_r(0) \\ &+ p_c(a) \log p_c(a) + p_c(b) \log p_c(b) \\ &+ p_c(c) \log p_c(c) + p_c(d) \log p_c(d). \end{aligned}$$

T est d'autant plus élevé que le système est plus performant. Un rapport est établi entre T, le rappel $\frac{a}{a+c}$ et la précision $\frac{a}{a+b}$ sous forme de courbes caractéristiques "statiques" et "dynamiques".

Remarque :
Il s'agit là d'une des publications les plus rigoureuses sur la ques-tion. Notons que CAWKELL est ingénieur, membre en 1975 de l'équipe d'I.S.I. (18).

4) Etude de M. GUAZZO (50).
Partant des mêmes bases que MEETHAM et CAWKELL, GUAZZO définit une mesure de l'information apportée par une session de ressaisie équivalente au facteur N près à celle de MEETHAM.

I est d'autant plus élevé que le système est plus performant, avec comme limite supérieure la quantité d'information contenue dans le jugement du sélecteur parfait, H(X).
Les valeurs prises par la fonction I permettent d'évaluer l'accord entre deux types d'indexation (manuelle et automatique par exem--ple) à partir d'une table de co-occurrence.

IV. INDEXATION AUTOMATIQUE

Les travaux de L. L. BRINER (12) (13) ont pour but l'identification et l'évaluation de mots-clés dans un texte.

A. IDENTIFICATION

Les relations entre mots-clés peuvent être directe : A rel. B, ou indirectes : A rel. B , A rel. C ⇒ B rel. C.
De multiples associations entre mots-clés conduisent BRINER à poser que "les mots-clés sont des noms dont la signification est fixée par des relations croisées multiples ou un usage grammatical redondant, et nous pouvons identifier les mots-clés sur la base de leur usage grammatical multiple ou de leur redondance au sein de propositions". Les mots généraux en tant que tels sont rejetés, sauf s'ils jouent un rôle de modifieur d'un autre mot. L'analyse de texte permet de les éliminer, et de ne conserver que les sujets, les objets, les compléments et les modifieurs que l'auteur classe en 3 rôles fonctionnels principaux :

- les sujets,
- les élaborateurs (compléments et objets directs),
- les clarificateurs (objets indirects et modifieurs).

B. EVALUATION

L'identification ayant servi à déterminer quels noms représentent le mieux le sujet traité, la quantification de la dispersion des mots-clés sert ensuite à estimer l'importance relative d'un mot-clé au sein d'un message.

Pour cela, l'auteur utilise la formule de SHANNON calculant la capa-cité d'une voie - dans le cas continu, que nous n'avons pas abordé - avec redéfinition très libre et arbitraire des variables :

$$C = B \log (1 + \frac{S}{N})$$

B = largeur de bande,
S = intensité du signal,
N = intensité du bruit.

1) Analogies.

a) Variable "largeur de bande" → "largeur de message" M :
M est défini comme le nombre de paragraphes où le mot-clé joue un des trois rôles (sujet, élaborateur, clarificateur).

b) Variable "intensité du signal" → "redondance multifonctionnel-le" R :
R apparaît comme une quantification relative de l'usage multi-fonctionnel :

$$R = \frac{SE + SC + EC}{K}$$

S = comptage du rôle "sujet",
E = comptage du rôle "élaborateur",
C = comptage du rôle "clarificateur",
K = comptage d'occurrence du mot-clé.

c) Variable "intensité du bruit" → "comptage du vocabulaire" V :
L'analogie provient de l'hypothèse qu'on peut comparer le bruit à la difficulté de traitement d'un texte, difficulté proportion-nelle à l'abondance du vocabulaire. Pour un mot-clé isolé, V = 1.

Compte tenu de ces analogies assez arbitraires, la capacité infor-mative d'un mot-clé est exprimée par la fonction :

$$X = M \log (1 + \frac{SE + SC + EC}{K})$$

2) Termes-clés.

Les mots-clés composés ("termes-clés") obtenus par juxtaposition de 2 ou 3 mots-clés simples expriment dans certains cas plus fidèlement que les mots simples pris un à un le sens voulu par l'auteur.

Dans ce cas, la capacité d'information de ces termes-clés est déterminée par une sommation pondérée des capacités K_N de chaque mot-clé, qui avantage la fin du terme-clé, conformément aux particularités de la langue anglaise :

$$X = \sum_{N=1}^{W} K_N \log \left(1 + \frac{W - N + 1}{W}\right)$$

K_N = capacité du N^e mot,
W = nombre de mots dans le terme-clé,
N = numéro du mot, de droite à gauche.

Une normalisation de ces formules doit être prévue en fonction de la taille du texte. Pour cela, la capacité de chaque mot ou terme-clé est divisée par une capacité globale de l'ensemble du texte, calculée selon le même principe que pour un mot-clé simple.

C. RESULTATS

Le programme informatisé SYNTRAN permet de traiter, selon les principes énoncés par BRINER, les textes de longueur moyenne. L'auteur reconnaît toutefois une limite aux possibilités du système : la faiblesse du procédé quand on désire indexer soit des textes courts, soit des textes longs. On peut remarquer, de plus, au vu des résultats reproduits en (12) une fâcheuse tendance des valeurs de capacité à avantager les mots et termes très généraux ("keyword", "text key--word", "text", par exemple) au détriment de notions plus fines (com--me "information capacity value") permettant mieux d'individualiser une publication.

Le rejet, au départ, des mots généraux sur des bases syntaxiques (les trois rôles fonctionnels) n'est peut-être pas assez restrictif. Bien que l'objet de l'expérience soit une indexation automatique "matières", on aboutit à l'extraction d'un vocabulaire de base peut-être mieux adapté à un programme de classification automatique qu'à un programme d'indexation "matières".

En outre, l'analogie artificielle entre capacité d'une voie avec bruit et valeur informative, telle qu'elle est présentée, semble traiter trop rapidement la notion de bruit.

On peut en effet se poser la question : Le message informatif est parsemé de signaux non discriminatoires, le tout formant un texte à indexer. Ne convient-il pas de faire entrer d'une façon ou d'une autre l'abondance des signaux non discriminatoires dans l'intensité du bruit ?

V. DIVERSITE D'UNE POPULATION BIBLIOGRAPHIQUE

A. MESURES EN ECOLOGIE QUANTITATIVE

La diversité d'une population est une notion particulièrement utile en écologie quantitative. Elle dépend à la fois du nombre d'espèces dans une collection et de l'abondance relative de chaque espèce. Sa quantification s'applique à l'étude d'une communauté, mais permet aussi de comparer la diversité de deux communautés. Les biologistes disposent d'un éventail de mesures de diversité, chacune adaptée à un type de problème écologique. Deux de ces mesures sont dérivées de la théorie de l'information : la mesure dite "de BRILLOUIN" et la mesure entropique ([65]) ([87]).

La mesure de BRILLOUIN ([11]) permet de caractériser le nombre de complexions possibles pour N objets distribués en s espèces différentes de population N_i :

$$H_B = \frac{K}{N} \log \frac{N!}{N_1! \dots N_i! \dots N_s!}$$, K étant une constante.

H_B dépend du nombre d'espèces, de l'abondance de chaque espèce et du nombre N d'individus. La comparaison de la diversité de deux popu--lations différentes se fait par le calcul d'une diversité relative :

$$H_{rel} = \frac{H_B}{H_{Bmax}},$$

H_{Bmax} correspondant à une répartition uniforme $N_i = \frac{N}{s}$. Pour des nombres N_i importants, l'approximation de STIRLING conduit à la formule de SHANNON, avec $p(i) = \frac{N_i}{N}$.

La mesure de SHANNON est utilisée sous sa forme habituelle - qui ne dépend pas du nombre d'individus -

$$H_s = - \sum_{i=1}^{s} p(i) \log p(i)$$

ou sous une forme corrigée en fonction du nombre d'individus :

$$E(H) = H_s - \frac{s - 1}{2N} .$$

La mesure de BRILLOUIN est plutôt employée dans le cas d'échan--tillons réduits. Celle de SHANNON est plutôt employée quand on dispose d'échantillons "sûrs" et quand le nombre total d'individus intervient peu. Cette distinction est cependant toute théorique : comme le fait remarquer G. F. ESTABROOK, cité par LEGENDRE (65), "les deux formules donnent des résultats identiques aux dernières décimales près, sauf lorsque les échantillons sont tellement petits que, de toute façon, on ne voudrait pas les utiliser pour des calculs de diversité spécifique".

B. DIVERSITE DE CO-REDACTION

La mesure H_B non normalisée, proposée par J. L. DOLBY (31) :

$$I = k \log \frac{N!}{N_1! \ldots N_i! \ldots N_s!}$$

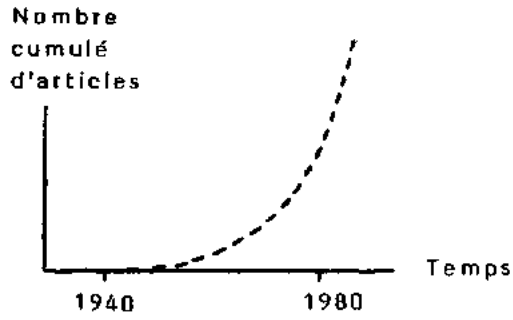
est appliquée par W. M. SHAW JR. (99) (101) au calcul du degré de cohésion d'une collection de N = 131 articles prenant en compte la pro--priété de rédaction par plusieurs auteurs (co-rédaction). Les N arti--cles sont groupés en s = 87 classes de co-rédaction selon le tableau ci-dessous :

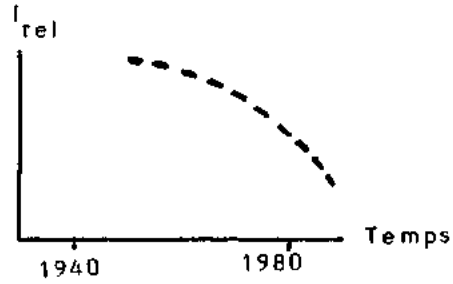
Nombre de classes	Nombre d'auteurs par classe	Nombre d'auteurs
1	20	20
1	18	18
1	10	10
3	6	18
6	5	30
9	4	36
13	3	39
25	2	50
28	1	28
Total : 87 classes		Total : 249 auteurs

SHAW tire de ces données deux séries de valeurs transcrites sur des graphiques :

1) Diversité relative de la population d'auteurs en fonction du temps :

$I_{rel} = \frac{I}{I_{max}}$; I_{max} correspond à N = s (un objet par espèce), $I_{max} = k \log N!$





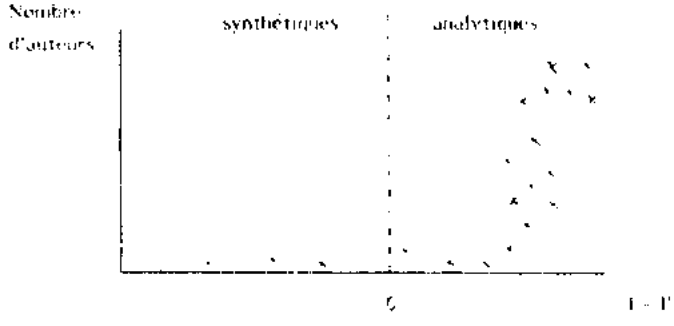
Remarque : SHAW présente les deux graphiques en un seul par sur--impression. Nous les avons séparés afin d'éviter tout risque de confusion.

2) Contribution de chaque auteur à la cohésion de la collection. Cette contribution est mesurée par l'écart I - I', I mesurant la diversité de la population entière et I' celle d'une population d'où l'auteur a été retiré.

Si I - I' > 0, le retrait produit moins de diversité : l'auteur joue un rôle analytique.

Si I - I' < 0, le retrait produit plus de diversité : l'auteur joue un rôle synthétique.

SHAW constate que 3 auteurs sur 249 jouent un rôle synthétique dans la population :



Ces calculs débouchent sur plusieurs questions et suggestions dis--cutables se rapportant à la sociologie de la communication scien--tifique, voire à la politique de financement de la recherche, dont :

a) Les articles publiés par les auteurs synthétiques contribuent-ils plus que d'autres au processus de communication scientifique ?

b) Peut-on trouver quelque utilité à identifier les auteurs et jour--naux concourant le mieux au processus de communication scien--tifique ?

De tels développements conduisent à des considérations qualitatives. Il serait en particulier imprudent, dans une telle perspective, d'affirmer que les auteurs synthétiques apportent plus que d'autres à la communauté scientifique.

L'étude de B. R. BOYCE et D. MARTIN (10) apporte, à cet égard, quelques éclaircissements. Suivant la même méthode que SHAW sur un échantillon plus important d'articles, les auteurs obtiennent des résultats comparables : en particulier, le nombre d'auteurs synthétiques demeure une faible fraction de la population.

Une approche succincte de la différence de performances entre les deux catégories d'auteurs en indexation par citation ne permet pas d'établir une corrélation certaine entre qualité "synthétique" et taux de citation. De même, il ne semble pas y avoir de rapport entre le type d'auteurs et le type ou le niveau des revues où ils publient. Les auteurs concluent à la possibilité d'étude des réseaux bibliographiques autres que fondés sur la co-rédaction.

Une autre étude de W. M. SHAW (100) permet à cet égard d'évaluer les caractéristiques d'une mesure de diversité basée sur le lien de citation réciproque, soit entre auteurs, soit plus globalement entre journaux.

Là encore, il est impossible d'affirmer :

a) que les auteurs "synthétiques" au sens de la co-rédaction le sont également au sens de la citation réciproque,
b) que les auteurs "synthétiques" au sens de la co-rédaction écrivent plus dans les journaux "synthétiques" au sens de la citation réciproque.

C. OPTIMISATION D'ACQUISITIONS DOCUMENTAIRES

La thèse de doctorat en philosophie de K. GARLAND ([47]), conduite par SHAW, constitue une extension des travaux de ce dernier.
Le but recherché par GARLAND est une quantification de l'adéquation d'un document à une collection. Ayant constaté qu'un grand nombre de livres achetés dans les bibliothèques universitaires ne sont jamais utilisés, l'auteur cherche à établir un critère de sélection pertinente basé sur la mesure de BRILLOUIN. Pour cela, GARLAND procède en trois étapes, la mesure de diversité n'intervenant que dans la dernière de ces étapes :

1) Agrégation de 416 livres par simple liaison.
La liaison est créée par la présence de mots-clés en commun (co-occurrence). Un seuil établi selon le degré plus ou moins élevé de co-occurrence permet de régler le processus d'agrégation.

2) Détermination de la validité des agrégats.
L'accord entre les agrégats formés et les sous-classes de la clas--sification de la Bibliothèque du Congrès permet de choisir un seuil de liaison optimum.

3) Impact de l'addition d'un document isolé.
La mesure de BRILLOUIN permet de mesurer la diversité de la collection de livres assemblée en espèces formées par les agrégats.

L'acquisition d'un document produit une modification des agrégats : celle-ci se traduit par une modification de diversité qui caractérise le document acquis et que l'on mesure. On constate en particulier que les livres traitant de sujets généraux font baisser I tandis que les livres traitant de sujets spécifiques font augmenter I.
Comme dans le cas de la co-rédaction, il est difficile d'étendre le raisonnement sur le plan qualitatif : l'étude ne fait pas la part de l'apport de redondance qui peut être inutile pour l'utilisateur, de l'approfondissement d'une spécialisation qui peut être fort utile, et de la diversification du fonds qui peut être un atout à moyen et long termes.

D. DIVERSITE DE CITATION RECIPROQUE

La thèse de doctorat en philosophie de W. M. M. DA ROCHA PARAN--HOS ([28]), dirigée elle aussi par SHAW, aborde le problème de l'éva--luation quantitative des journaux scientifiques.
On distingue dans ce domaine deux grands types d'évaluation :

1) L'évaluation "externe" faite sur la base des activités d'une unité documentaire par des mesures de nombre d'emprunts, de photo--copies, de citations par les utilisateurs de l'unité...

2) L'évaluation "interne" basée sur les propriétés objectives de la littérature scientifique elle-même.

L'étude de DA ROCHA PARANHOS fait appel à cette seconde appro--che, utilisant pour cela cinq mesures :

1) La productivité : nombre d'articles publiés dans un journal pendant une période donnée.

2) Les citations reçues : fréquence avec laquelle les articles d'un journal ont été cités dans la littérature pendant une période donnée.

3) Le facteur d'impact : relativisation de la mesure précédente en fonction de la taille du journal cité, à savoir le rapport $\frac{c}{a}$, avec :
c = nombre de citations dans l'année A des articles des années A - 2 et A - 1 ; a = nombre d'articles publiés dans les années A - 2 et A - 1.

4) La mesure d'influence : pondération des citations reçues en fonction de l'importance du journal citant.

5) La mesure de BRILLOUIN.

Partant du Journal Citation Report publié par l'I.S.I. sur un échan-tillon de 856 journaux biomédicaux, la démarche de l'auteur comprend trois grandes étapes :

1) Agrégation des journaux par simple liaison.
La liaison est créée par une relation de citation réciproque repré-sentée par le minimum du degré de liaison du citant au cité, chaque journal de la paire étant tour à tour citant et cité :
degré de liaison = $\frac{\text{nombre de citations que le citant donne au cité}}{\text{nombre de citations que le citant donne au total}}$.
Un seuil est établi pour 8 valeurs de ce degré de liaison.

2) Mesure de diversité.
La contribution de chaque journal au processus de communication scientifique est mesurée selon le même principe que dans les expé-riences de diversité de co-rédaction. Il en ressort un lot de 184 journaux "synthétiques" à tous les seuils.

3) Etude comparée d'évaluation.
Partant de ce lot de journaux synthétiques, la mesure de BRIL-LOUIN est comparée aux quatre mesures précédentes (productivité, citations reçues, facteur d'impact, mesure d'influence).

L'étude, intéressante par la variété des sujets abordés et par l'impor-tance des moyens informatiques mis en oeuvre, rejoint - avec plus de nuance - celle de BOYCE et MARTIN.

Plusieurs insuffisances sont mises en lumière, dont :

1) La grande sensibilité de la distribution en agrégats par rapport au degré de liaison, sans que le découpage revête une signification concrète très nette.

2) La très faible corrélation, dans beaucoup de cas, entre I et les autres mesures. En fait, les qualités des mesures "citations reçues" et "facteur d'impact" sont mises en évidence.

VI. STOCKAGE DES DONNEES EN ORDINATEUR

La nécessité de loger le plus possible de données en ordinateur sous le plus faible volume a conduit les informaticiens à mettre au point diver--ses techniques de compactage. La "génération de variété" de M. F. LYNCH et al. (25) désigne une démarche débouchant principalement sur la compression de données et la recherche de texte en machine. L'expres--sion utilisée est destinée à rendre compte de la liberté de choix dans la sélection des composants d'un texte. Ainsi, un texte peut être considéré comme formé d'un ensemble de symboles de taille arbitraire allant du caractère alphabétique au mot.
Partant d'un texte en langage naturel caractérisé par une distribution hyperbolique fréquence / rang des lettres simples, on aboutit à une distribution rectangulaire de n-grammes à fréquence à peu près cons--tante. L'équiprobabilité de ces fragments accroît l'entropie - donc dimi--nue la redondance - du texte et reste aisément codable.

A. COMPRESSION DE TEXTE

Les signes d'un texte sont représentés, pour traitement informatique, par un code binaire occupant un espace de longueur fixe (multiplet) composé communément de 6 ou 8 chiffres binaires ("bits"). Un multi--plet court (6 chiffres) ne permet de transcrire sous forme codée qu'une collection restreinte de 2^6 = 64 signes, tandis qu'un multiplet plus long (8 chiffres) se prête au codage d'un registre plus étendu de 2^8 = 256 signes comprenant majuscules et minuscules par exemple. On peut aussi utiliser différemment la collection de signes disponi--bles : si on considère une combinaison de lettres comme un symbole unique, la transcription permet d'emmagasiner plus de texte pour le même nombre de chiffres binaires, compte tenu d'un choix judicieux des symboles. La génération de variété suppose la sélection d'unités de texte (n-grammes) codables par le même nombre de chiffres binaires. La taille de la collection de ces unités est choisie de façon

à s'adapter à la longueur de l'unité de stockage : si les multiplets contiennent 8 chiffres (octets), on pourra coder 256 symboles distincts. Le codage est réalisé au moyen d'une table emmagasinée en mémoire interne.

1) Création des symboles.

Un ensemble de 256 symboles est constitué de chaînes de caractères alphanumériques de longueur variable numérotés de 0 à 255, de telle façon qu'un texte quelconque est représentable par une suite de nombres occupant chacun un octet. On remarquera toutefois l'absence des lettres de bas de casse.

Exemple :

Collection de 256 symboles :

-0-	-O-	-SE-	-GHT -	-MA-	-PRO-	-TH -
-1-	-P-	-BO-	-GH-	-ME -	-PR-	-THE-
-2-	-Q-	-BUT -	-GO-	-ME-	-QU-	-TH-
-3-	-R-	-BY -	-H -	-MI-	-R -	-TION -
-4-	-S-	-CA-	-HAS -	-MO-	-R. -	-TION-
-5-	-T-	-CE -	-HA-	-MRS. -	-RA-	-TI-
-6-	-U-	-CE-	-HER -	-N -	-RE -	-TO THE -
-7-	-V-	-CH -	-HE -	-N, -	-RE-	-TO -
-8-	-W-	-CH-	-HE-	-NA-	-RI-	-TO-
-9-	-X-	-CI-	-HIS -	-NC-	-RO-	-TR-
-=-	-Y-	-CK-	-HI-	-ND-	-RS -	-TT-
-?-	-Z-	-COM-	-HO-	-NE-	-RS-	-UL-
- -	-\-	-CON-	-IC-	-NG -	-RT-	-UN-
-"-	-'S -	-CO-	-IL-	-NG-	-RY -	-UR-
-$-	-, AND -	-CT-	-IN THE -	-NI-	-S -	-US-
-'-	-, -	-D -	-ING -	-NOT -	-S, -	-UT-
-(-	-. THE -	-DA-	-IN -	-NO-	-S. -	-VE -
-)-	-. -	-DE-	-ING-	-NT -	-SAI-	-VER-
-+-	-A -	-DI-	-IN-	-NT-	-SA-	-VE-
-,-	-AC-	-E -	-IO-	-O -	-SE-	-VI-
-*-	-AD-	-EA-	-IS -	-OF THE -	-SH-	-W -
-.-	-AG-	-ED -	-IS-	-OF -	-SI-	-WAS -
-/-	-AI-	-EN -	-IT -	-ON THE -	-SO-	-WA-
-A-	-AL -	-EN-	-IT-	-ON -	-SS -	-WERE -
-B-	-AL-	-ER -	-KE-	-ON-	-SS-	-WE-
-C-	-AM-	-ER-	-L -	-OR -	-ST -	-WHO -
-D-	-AND -	-ES -	-LA-	-OR-	-ST-	-WH-
-E-	-AN -	-ES-	-LD -	-OT-	-SU-	-WILL -
-F-	-AN-	-EV-	-LE -	-OUT -	-T -	-WITH -
-G-	-AR-	-EX-	-LE-	-OU-	-TA-	-WI-
-H-	-AS -	-FE-	-LI-	-OW-	-TED -	-Y -
-I-	-AS-	-FF-	-LL -	-P -	-TER -	-Y, -
-J-	-ATIO-	-FI-	-LL-	-PA-	-TER-	-Y. -
-K-	-AT -	-FOR -	-LO-	-PE-	-TE-	-YEA-
-L-	-AT-	-FOR-	-LU-	-PLA-	-THAT -	
-M-	-BA-	-FROM -	-LY -	-PO-	-THE -	
-N-	-BE -	-GE-	-M -	-PRES-	-THER-	

Extrait d'un texte à emmagasiner :

THE FULTON COUNTY GRAND JURY SAID FRIDAY AN INVESTIGATION OF ATLANTA'S RECENT PRIMARY ELECTION PRODUCED "NO EVIDENCE" THAT ANY IRREGULARITIES TOOK PLACE. THE JURY FURTHER SAID IN TERM-END PRESENTMENTS THAT THE CITY EXECUTIVE COMMITTEE, WHICH HAD OVER-ALL CHARGE OF THE ELECTION, "DESERVES THE PRAISE AND THANKS OF THE CITY OF ATLANTA" FOR THE MANNER IN WHICH THE ELECTION WAS CONDUCTED. THE SEPTEMBER-OCTOBER TERM JURY HAD BEEN CHARGED BY FULTON SUPERIOR COURT JUDGE DURWOOD PYE TO INVESTIGATE REPORTS OF POSSIBLE "IRREGULARITIES" IN THE HARD-FOUGHT PRIMARY WHICH WAS WON BY MAYOR-NOMINATE IVAN ALLEN JR. "ONLY A RELATIVE HANDFUL OF SUCH REPORTS WAS RECEIVED", THE JURY SAID, "CONSIDERING THE WIDESPREAD INTEREST IN THE ELECTION, THE NUMBER OF VOTERS AND THE SIZE OF THIS CITY".

Le même texte décomposé en unités :

( ) (THE) (F) (UL) (TO) (N) (CO) (UN) (T) (Y) (G) (RA) (ND) ( )
(J) (UR) (Y) (SAI) (D) (F) (RI) (DA) (Y) (AN) (IN) (VE) (ST) (I)
(G) (ATIO) (N) (OF) (AT) (LA) (NT) (A) ('S) (RE) (CE) (NT) (PR)
(I) (MA) (RY) (E) (LE) (CT) (IO) (N) (PRO) (D) (U) (CE) (D) (")
(NO) ( ) (EV) (I) (DE) (NC) (E) (") ( ) (THAT) (AN) (Y) (I) (R)
(RE) (G) (UL) (AR) (IT) (I) (ES) (TO) (O) (K) ( ) (PLA) (CE)
(. THE) (J) (UR) (Y) (F) (UR) (THER) ( ) (SAI) (D) (IN) (TER) (M)
(-) (EN) (D) (PRES) (EN) (T) (ME) (NT) (S) (THAT) (THE) (CI) (T)
(Y) (EX) (E) (C) (UT) (I) (VE) (COM) (MI) (TT) (E) (E) (,) (WH)
(IC) (H) (HA) (D) (O) (VER) (-) (AL) (L) (CH) (AR) (GE) ( )
(OF THE) (E) (LE) (CT) (IO) (N,) (") (DE) (SE) (R) (VE) (S) (THE)
(PR) (AI) (SE) ( ) (AND) (TH) (AN) (K) (S) (OF THE) (CI) (T) (Y)
(OF) (AT) (LA) (NT) (A) (") ( ) (FOR) (THE) (MA) (N) (NE) (R)
(IN) (WH) (IC) (H) (THE) (E) (LE) (CT) (IO) (N) (WAS) (CON) (D)
(U) (CT) (E) (D) (. THE) (SE) (P) (TE) (M) (BE) (R) (-) (O) (CT) (O)
(BE) (R) (TER) (M) (J) (UR) (Y) (HA) (D) (BE) (EN) (CH) (AR)
(GE) (D) (BY) (F) (UL) (TO) (N) (SU) (PE) (RI) (OR) (CO) (UR)
(T) (J) (U) (D) (GE) ( ) (D) (UR) (W) (O) (O) (D) (P) (Y) (E)
(TO) (IN) (VE) (ST) (I) (G) (AT) (E) (RE) (PO) (RT) (S) (OF) (PO)
(SS) (I) (B) (LE) (") (I) (R) (RE) (G) (UL) (AR) (IT) (I) (ES) (")
( ) (IN THE) (HA) (R) (D) (-) (F) (OU) (GHT) (PR) (I) (MA) (RY)
(WH) (IC) (H) (WAS) (W) (ON) (BY) (MA) (Y) (OR) (-) (NO) (MI)
(NA) (TE) ( ) (I) (V) (AN) (AL) (LE) (N) (J) (R) (.) (") (ON)
(LY) (A) (RE) (LA) (TI) (VE) (HA) (ND) (F) (UL) ( ) (OF) (SU)
(CH) (RE) (PO) (RT) (S) (WAS) (RE) (CE) (I) (VE) (D) (") (,)
(THE) (J) (UR) (Y) (SAI) (D) (,) (") (CON) (SI) (DE) (RI) (NG)
(THE) (WI) (DE) (S) (PR) (EA) (D) (IN) (TER) (ES) (T) (IN THE)
(E) (LE) (CT) (IO) (N,) (THE) (N) (U) (M) (BE) (R) (OF) (V) (OT)
(ER) (S) (AND) (THE) (SI) (Z) (E) (OF) (TH) (IS) (CI) (T) (Y)
(") (.)

La création du jeu de symboles répond au souci d'approcher l'équi--probabilité.

La méthode de création des symboles est basée sur des décomptes statistiques opérés sur un échantillon de texte tiré du Brown Corpus (recueil standard d'anglais-américain contemporain). Un algorithme permet de comptabiliser et de ranger par ordre décroissant les fréquences d'apparition des caractères alphanumériques isolés. Les symboles ainsi pris en compte sont loin d'être équifréquents. Les extrémités du tableau se présentent ainsi :

Caractère		E	I	T	O	.....	/	%	;	=	£
Fréquence	11003	7497	6117	5865	5608	.....	6	3	2	1	1

Le calcul est repris pour non plus des caractères isolés, mais des digrammes (suites de deux caractères), ce qui permet de compléter le tableau précédent en s'arrêtant à une valeur-plancher de fré--quence. Le même processus est poursuivi pour les trigrammes et un certain nombre de n-grammes jusqu'à ce qu'on obtienne un jeu de 256 symboles à distribution de fréquence homogène.

2) Résultats.

L'équipe de l'Université de Sheffield obtient un taux de compression de l'ordre de 51 à 52 % sur divers échantillons de textes tirés du Brown Corpus.

Des travaux analogues menés à l'Université de Bradford par YAN--NAKOUDAKIS et al. (110) sur différents jeux de symboles (64, 128, 256, 512, 1024) aboutissent à des valeurs inférieures à celles obtenues par LYNCH. Ces valeurs sont croissantes de 64 à 1024 symboles, de l'ordre respectivement de 25 %, 36 %, 44 %, 49 %, 54 %.

Une telle démarche est particulièrement utile si on désire stocker des données en mémoire interne et en accès direct, ou les trans--mettre dans une voie de télécommunication nécessitant une forte capacité de transmission. Le développement actuel des bases

et banques de données en ligne et des structures en réseau ne peut qu'appeler ce type de traitement de données.
Il s'agit là d'une méthode de compactage parmi d'autres : LOUIS-GAVET (67) en distingue au moins 8. Chacune a ses avantages et ses inconvénients et doit faire l'objet d'un bilan global tenant compte non seulement du taux de compression, mais aussi du temps de traitement et des risques d'erreurs.

B. RECHERCHE DE TEXTE

La démarche de LYNCH consiste à rechercher en machine les mots des notices d'une base de données à partir de la comparaison entre le terme de la recherche et la suite des enregistrements sans passer par un fichier inversé.
Pour cela, le texte est fractionné de plusieurs façons redondantes en n-grammes survenant le plus équifréquemment possible dans la base, et utilisés comme entrées d'une matrice décrivant chaque document sous la forme d'un vecteur à valeurs binaires 1 ou 0 indi--quant la présence ou l'absence du symbole.

Exemple :
Partie de matrice (de "E" à "FOR ") décrivant les phrases (i) et (ii) :

	(i)	(ii)
E	1	1
EV	1	0
EVA	0	0
EVO	0	1
EVS	1	0
EA	0	0
EC	0	0
ECT	1	0
ECTR	0	0
EDV	0	0
EL	0	0
ELE	0	0
EM	0	0
EN	0	1
ENE	0	0
ENT	0	0
ENTV	0	0
ER	0	0

.../...

ERV	1	0
ERA	0	0
ERI	0	0
ES	0	0
ESV	0	0
ET	0	0
EX	0	0
F	1	1
FV	0	0
FVTHEV	0	0
FE	1	0
FF	1	0
FL	0	0
FORV	0	0

(i) THE FREE CARRIER EFFECT IN N-TYPE SILICON

(ii) FREE OXYGEN ATOMS IN ORGANIC CRYSTALS

Les termes de la recherche sont eux aussi fragmentés dans le même jeu de symboles et comparés grâce à la matrice à la collection de fragments appartenant aux enregistrements. Les documents contenant les symboles par lesquels la question est représentée deviennent "candidats pertinents", la recherche ne faisant pas intervenir l'ordre de succession des symboles.
Une comparaison caractère par caractère effectuée ensuite sur les candidats permet d'éliminer les documents dans lesquels les symboles ne sont pas enchaînés dans le bon ordre.
Un tel système offre un certain nombre d'avantages :

1) Il évite la création d'un fichier inversé.

2) Il permet certaines facilités d'interrogation : en particulier, les termes ou les titres étant traités comme des chaînes continues de caractères (dont l'espace), la recherche en ligne peut faire appel à la troncature à gauche, à droite, ou les deux.

Dans l'état actuel des travaux, toutes les possibilités d'un tel système n'ont pas encore été explorées. Il est possible qu'une fragmentation des données et des termes d'interrogation puisse déboucher sur la solution de certains problèmes linguistiques auxquels se heurte l'inter-rogation des bases de données.

VII. ETUDES DE DOMAINES CONNEXES A LA T.M.C.

A. INFORMATION SEMANTIQUE ET RESSEMBLANCE FLOUE

Utilisant les bases de la théorie de l'information sémantique de R. CARNAP et Y. BAR-HILLEL (17), P. PIETILAINEN (84) propose une méthode de classement de publications par ordre décroissant de res--semblance avec une question.

Dans le plus simple des cas - mots d'un texte se succédant sans lien syntaxique - la quantité d'information de chaque terme t est définie par CARNAP et BAR-HILLEL comme l'information spécifique locale en t, non pondérée par la probabilité d'apparition de t :

$inf(t) = - \log_2 p(t)$.

La comparaison d'un texte d (document) avec une question q (suite de termes) est opérée par une fonction de ressemblance floue R :

$$R = \frac{\sum_{t \in q \cap d} inf(t)}{\sum_{t \in q} inf(t)}$$

La somme du numérateur se fait à partir du comptage des termes en commun entre q et d, celle du dénominateur à partir du comptage des termes de la question.

inf(t) proprement dite est calculée sur la base de la spécificité rela--tive du terme dans la question, en fonction de sa fréquence dans une base de données. Cette fréquence est accessible lors d'une sélec--tion de terme au sein de la base de données : elle comptabilise le nombre de documents qui comprennent le terme t :

$$p(t) = \frac{f(t)}{\sum_{t \in q} f(t)} .$$

Il ne s'agit donc pas d'une véritable probabilité d'occurrence, mais plutôt d'un indice de spécificité relative du terme au sein de la question.

L'expérimentation, pratiquée sur 4 bases de données différentes, per-
-met la comparaison - après élimination des mots vides - entre une question composée en langage naturel sous forme d'une phrase, et les titres d'un certain nombre de publications à ressaisir. Une liste de publications classées par ordre décroissant de ressemblance est obtenue par calcul de la fonction R. Cette méthode, proche dans son esprit de celle de J.-B. CRAMPES (27) comporte deux inconvénients :

1) Elle privilégie les documents les plus longs pour deux raisons :
   a) Ceux-ci apportent au numérateur de la fonction R un grand nombre de t à comparer aux termes de q ;
   b) Le dénominateur ne permet une relativisation de l'information apportée par les termes en commun que par rapport à la question seule.

2) Elle ne prend en compte que la coïncidence stricte de termes entre q et d. Si tous les termes de d sont des synonymes de q, R = 0. La relation floue reste donc tributaire d'une comparaison du type "tout ou rien" - présence ou absence de terme dans q et d - comme dans toute démarche où les recherches d'association sont basées sur un comptage des descripteurs communs. En particulier, ne peuvent compter ni les écarts de terminologie, ni les associations sémantiques.
   L'auteur, sensible à cet inconvénient, parvient à assouplir la relation question / document par une procédure en deux temps rappelant celle de H. E. STILES (106) et comportant une reformulation de la question initiale. Cette méthode, qui donne lieu à une publication ultérieure (85), consiste à utiliser la question initiale non plus pour ressaisie mais comme première étape d'un processus itératif. Les documents à plus forte relation de ressemblance ressaisis à la suite de la question initiale donnent naissance à un ensemble élargi de termes trouvés dans le titre et (ou) la zone descripteurs de plus d'un document. Ces termes sont enfin utilisés à leur tour com-
   -me question définitive. Ainsi, à l'aide de ce qu'on pourrait appeler des circonymes ("searchonyms"), l'auteur parvient à dépasser la coïncidence stricte et, par exemple, à récupérer le singulier de termes-question énoncés sous forme plurielle.

B. INFORMATION HYPERBOLIQUE

1) Principes de base.

La formule de SHANNON se présente comme l'espérance mathé-matique de l'information apportée par un message, les probabilités d'occurrence des symboles étant données a priori. Les travaux de J. KAMPE DE FERIET sur la théorie générale de l'information apportent un déplacement du point de vue. On passe d'une estima-tion a priori à une connaissance a posteriori de l'information four-nie par la réalisation d'un événement. Ce qui entraîne, dans ce dernier cas, une redéfinition de l'information, qui pourra être fonc-tion d'autre chose que d'une probabilité - une mesure par exem-ple - et qui, par conséquent, pourra prendre en considération des aspects sémantiques et subjectifs (61) (62).

Les travaux de F. FOREST (41) (42) (43) posent le problème de l'application de l'information hyperbolique en sciences de l'infor-mation, qui devrait pouvoir se traduire par la quantification de la distance entre deux éléments (descripteurs, documents par exemple). On décrit l'information généralisée en termes non plus de proba-bilité mais de mesure d'un ensemble et, plus particulièrement dans le contexte documentaire, de cardinal d'un ensemble.

Soit un ensemble Ω d'événements élémentaires, C une classe de parties de Ω, et A un événement appartenant à C.

L'information J(A) respecte trois axiomes de base :

A1 : $J : C \rightarrow \bar{\mathbb{R}}^+$

J(A) est un nombre non négatif.

A2 : $B \subset A \Rightarrow J(A) \leq J(B)$

J est monotone par rapport à l'inclusion, avec les valeurs extrêmes $J(\emptyset) = +\infty$ et $J(\Omega) = 0$.

A3 : Si les événements A et B sont indépendants,

$J(A \cap B) = J(A) + J(B)$.

On définit l'information hyperbolique par :

$$J(A) = \frac{1}{\mu(A)} - \frac{1}{\mu(\Omega)} ,$$

avec annulation du second terme si $\mu(\Omega) = +\infty$.

2) Contexte documentaire.

L'information hyperbolique offre des propriétés intéressantes si on mesure les événements par leur cardinal. Sa "monotonicité" par rapport à l'inclusion lui confère la capacité de favoriser la rareté ou la spécificité.

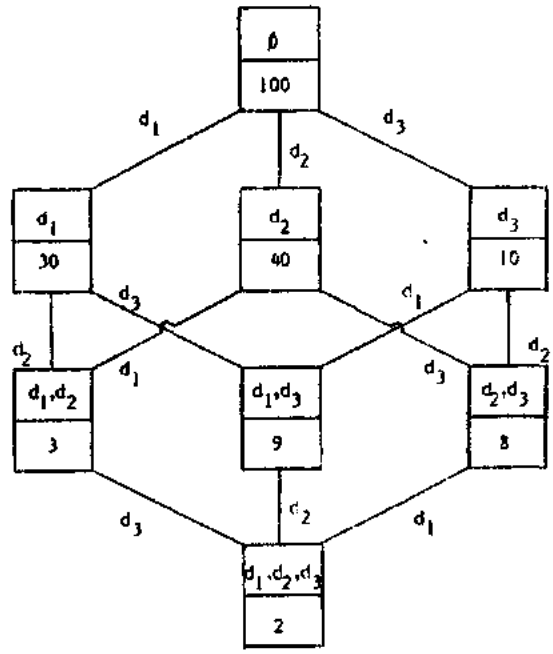
D'autre part, l'indépendance de deux A et B se traduit par : $J(A \cap B) = J(A) + J(B)$.

Outre l'information hyperbolique, FOREST utilise la théorie des graphes et des questionnaires pour proposer une formalisation théorique, très rigoureuse du point de vue mathématique, des recherches documentaires.

Une recherche documentaire est, dans cette perspective, symbolisée par le cheminement sur un graphe dont les sommets sont des états $(\{d_i\},m_i)$ et les arcs des interventions d_i, ce qui signifie :

Le sommet $(\{d_i\},m_i)$ correspond à un ensemble de m_i documents indexés par le(s) descripteur(s) $\{d_i\}$, l'arc d_i correspond à l'intro--duction du descripteur d_i dans la recherche.

Exemple : On recherche les documents (au nombre de 2) caracté--risés par la relation $d_1 \cap d_2 \cap d_3$ au sein d'une collection de 100 documents. Les diverses façons de réaliser la recherche sont repré--sentées par les divers chemins $[\emptyset,\{d_1,d_2,d_3\}]$ du graphe ci-dessous :



On pourra, à partir de certains critères (choix des descripteurs, ensemble des documents demeurés pertinents à chaque étape,...) évaluer quel est le meilleur chemin pour aboutir au résultat. Pour cela l'auteur fait intervenir une information de cheminement déduite de l'information I(x) des sommets franchis et de l'information $I(x,y) = \mu(x)\,[I(y) - I(x)]$ des arcs parcourus.

Les exemples donnés par FOREST n'étant pas tirés de recherches documentaires expérimentales réelles, ne peuvent être soumis à évaluation.

Il est à noter que l'auteur évoque la possibilité d'appliquer l'information hyperbolique à la définition d'une distance entre deux descripteurs, voire entre deux documents.

DIFFICULTES INHERENTES A L'UTILISA-TION DE LA THEORIE MATHEMATIQUE DE LA COMMUNICATION

I. LIMITES DES FONCTIONS DE SHANNON

Invité en 1956 à un congrès de documentation, le très célèbre théoricien du codage R. M. FANO affirmait : "Le corpus de connaissance théorique présentement disponible sous le nom de théorie de l'information ne pro-cure aucune solution aux difficiles problèmes que vous affrontez en essayant d'exploiter les possibilités surprenantes des ordinateurs numéri-ques. En outre, permettez-moi de dissiper l'idée que la présence du mot "information" dans "théorie de l'information" implique que cette théorie est nécessairement appropriée à la "mécanisation des bibliothè-ques" " (39). E. GARFIELD fait lui aussi remarquer quelque vingt ans plus tard - et quelques mois après S. E. ROBERTSON (93) - le peu d'appli-cabilité à son avis de la T.M.C. à la ressaisie de l'information. Affirmant dans un "Current comment" plaisamment intitulé "Information theory and all that jazz" (45) que "la science de l'information et la théorie de l'infor-mation sont deux domaines tout à fait distincts", GARFIELD laisse peut-être involontairement la porte ouverte : on peut en effet, d'une manière générale, résoudre une question grâce à une théorie développée dans un autre domaine de la connaissance. D'ailleurs, dans le pire des cas, FANO concède que, en dehors de toute application directe, "la théorie de l'information devrait pouvoir suggérer de nouveaux points de vue à partir desquels on pourrait penser les problèmes documentaires".

Plus près de nous, M.-P. SCHÜTZENBERGER fait remarquer, dans un exposé corrosif mais somme toute encourageant, que les tentatives d'uti-lisation de la T.M.C. dans diverses disciplines se heurtent à des écueils inhérents à la difficulté de mesurer précisément les phénomènes en pré-sence. Le développement d'emplois judicieux de la notion d'information ne serait donc qu'une question de temps (96).

Il n'en reste pas moins que divers obstacles s'opposent actuellement à l'application de la T.M.C. dans le domaine des sciences et techniques documentaires. On peut en citer trois, d'importance inégale :

- difficulté de définir les objets de l'expérience,
- difficulté d'application du concept de codage,
- difficulté d'aborder les problèmes de signification.

A. DIFFICULTE DE DEFINIR LES OBJETS DE L'EXPERIENCE

Il s'agit de déterminer les événements sur lesquels appliquer un processus de communication ou une mesure de complexions.
On rejoint ainsi le souci de FAIRTHORNE de dresser un inventaire des objets auxquels peut s'appliquer directement ou indirectement la théorie de l'information.
Les axes de recherche passés en revue précédemment correspondent en fait à un certain nombre de points de vue différents sur les ob--jets et phénomènes à prendre en compte dans un processus de communication ou d'observation. On peut classer ces objets et phé--nomènes en sept catégories :

1) Mot.

On peut envisager de diverses façons l'utilisation des mots :

a) Considérer les mots-clés comme des variables dotées d'une probabilité d'occurrence. C'est ainsi que GARFIELD considère le mot-clé (descripteur) comme événement d'un message don--nant lieu à communication par l'intermédiaire d'une carte perforée. Le point de vue de MARON est voisin, les probabi--lités s'appliquant également à des chapitres (catégories) de classification. L'étude de PIETILAINEN entre également dans ce cadre, bien que la probabilité d'occurrence soit remplacée par une spécificité relative liée à la question.

b) Considérer des groupes de mots-clés possédant des propriétés communes. L'étude de ZUNDE et SLAMECKA s'appliquant à des groupes de mots-clés renvoyant à un même nombre de documents ne peut malheureusement être retenue, le problème étant mal posé.

c) Développer une analogie arbitraire avec des matériaux d'ana--lyse linguistique. L'analyse linguistique conduisant BRINER à une valeur d'indexation d'un mot-clé ou d'un terme-clé fait

intervenir la capacité de voie des composants structurels du texte écrit que sont les mots et groupes de mots. L'analogie reste cependant très artificielle, d'autant plus que la formule de la capacité d'une voie traversée par des messages du type continu est employée dans le contexte par nature discontinu du message écrit.

Une telle analogie, prise au pied de la lettre, place le point de vue de BRINER à l'opposé de celui de GARFIELD qui considère le mot comme un élément de message à transmettre et non pas comme la voie dans laquelle il est transmis.

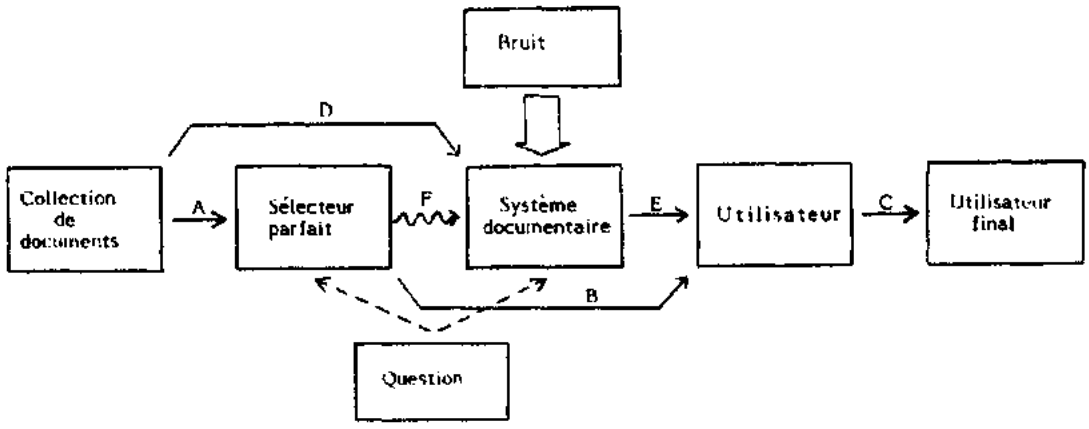
2) Caractère alphabétique et n-gramme.

On retrouve dans les travaux de LYNCH le souci de GARFIELD de rentabiliser l'outil informatique. Compte tenu à la fois des propriétés du texte écrit et des contraintes liées à l'emploi de l'ordinateur, le message reçu par l'ordinateur est un jeu étendu de caractères alphabétiques et de n-grammes dotés de probabi--lités d'occurrence à peu près égales.

3) Pertinence.

HAYES, MEETHAM, BELZER, CAWKELL et GUAZZO considèrent la pertinence comme une variable prenant la valeur 0 ou 1. Les études des quatre derniers auteurs portent sur la transmission de signaux entre un sélecteur parfait et un utilisateur (les deux pouvant se confondre). L'axe de recherche ainsi délimité apparaît comme le plus minutieusement suivi jusqu'à maintenant dans le domaine des applications de la T.M.C..

Le processus peut être décrit selon le schéma de la page suivante, très proche de celui de SHANNON.



Le schéma peut être explicité comme suit :

a) Parcours de l'information lors d'une ressaisie par le sélecteur parfait : A - B - C.
b) Parcours de l'information lors d'une ressaisie par le système documentaire : D - E - C.
c) Parcours virtuel de l'information lors d'une ressaisie par le système documentaire : A - F - E - C.

L'analogie avec un processus de télécommunication est ici poussée assez loin, d'autant plus que le sélecteur parfait transcrit le message (l'ensemble des évaluations de pertinence) en signaux formés de symboles binaires 1 et 0.
L'objet donnant lieu à communication est ici un message constitué d'une suite d'événements que sont la pertinence et la non perti--nence de chaque document par rapport à la question.

4) Flux d'information.
La modélisation qualitative et graphique de FAIRTHORNE permet de visualiser les diverses tâches assurées par le praticien des bibliothèques et centres de documentation. Les objets étudiés

sont six concepts généraux de la T.M.C. (message - voie - code et source - destinataire - désignation) du point de vue de leurs relations dans un contexte documentaire.

5) Diversité d'une population.

L'évaluation de la diversité d'une population bibliographique à partir de la mesure du nombre de complexions possibles de N objets est basée sur l'analyse combinatoire. Les objets sont cons--titués par des publications regroupées par classes de co-rédaction (SHAW), de co-occurrence de mots-clés (GARLAND) ou de cita--tion réciproque (DA ROCHA PARANHOS). Il n'y a pas à propre--ment parler de communication, mais simplement mesure d'une population réelle et comparaison avec une population restreinte. D'une façon générale, bien que l'observation d'une population puisse apparaître comme un processus de communication objet - observateur, on touche ici plus au domaine de la théorie de l'information que de celle de la communication.

6) Nombre de documents.

L'application conjointe de la théorie des graphes et de l'infor--mation hyperbolique à la formalisation de processus de ressaisie de l'information conduit FOREST à prendre en compte dans les calculs d'information de cheminement le cardinal des sous-ensembles d'une collection documentaire estimée pertinents à chaque étape de la ressaisie. Ce point de vue est en fait voisin du premier, prenant en compte les mots, dans la mesure où le nombre de documents pertinents est conditionné par l'intervention d'un descripteur.

7) Fréquence / rang.

L'unification des lois empiriques du type fréquence = f(rang) envisagée par ZUNDE est une question fort complexe, à laquelle on ne peut être assuré de trouver une solution. Au-delà d'une

forme hyperbolique commune, les phénomènes étudiés sont très divers et la détermination des événements auxquels on attribue une fréquence et un rang est spécifique de chaque loi empirique.

B. DIFFICULTE D'APPLICATION DU CONCEPT DE CODAGE

Le principal souci de SHANNON est l'optimisation de la transmis--sion, ce qui le conduit à souhaiter le codage le plus performant. La difficulté pour le scientifique de l'information est de tirer le maximum d'applications de la théorie de la communication.
Jusqu'à maintenant, la notion de codage n'a donné lieu qu'aux appli--cations énumérées plus haut et ne faisant pas intervenir tous les raffinements de la T.M.C. : applications de GARFIELD, de LYNCH et du groupe d'auteurs s'intéressant à la performance des systèmes documentaires.
Peut-on aller plus loin ? La question en entraîne trois autres :
- Qu'y a-t-il de codable ?
- Peut-on faire apparaître des concepts analogues à ceux de capacité de voie ?
- L'abus du raisonnement analogique ne risque-t-il pas de conduire à des spéculations stériles ? (7)

La notion de codage en T.M.C. est liée à celle d'économie de la transmission : il s'agit d'utiliser le moins possible de symboles binaires et de réduire le bruit dans la limite permise par la capacité de la voie.
Le scientifique de l'information est certes habitué à pratiquer une forme particulière de codage qui, partant de la correspondance entre le contenu d'un document et une partie bien déterminée d'une table de classification, consiste à assigner au document un code de classification alphanumérique symbolisant cette correspondance. Les codes de classification ont pour but la symbolisation la plus simple et la plus claire possible : il est plus commode de classer

une publication en 110.C.02.F.03.B que sous la séquence hiérarchique de PASCAL formulée en clair :

Analyse numérique. Informatique. Automatique. Statistique et pro-
-babilités. Recherche opérationnelle. Gestion. Economie. [110]

. Informatique. [110.C]

.. Informatique théorique. [110.C.02]

... Théorie de l'information. [110.C.02.F]

.... Codage. [110.C.02.F.03]

..... Théorie du codage et du décodage. [110.C.02.F.03.B]

La transmission du signal codé, essentielle en T.M.C., ne pose quant à elle que des problèmes mineurs (fautes de frappe sur une fiche, griffonnage d'une cote sur un formulaire de prêt, par exemple). Le coeur du problème réside ailleurs : le souci de transmission éco-
-nomique de l'information est sous-jacent mais secondaire si on consi-
-dère le système publication - classification - lecteur. Le signal 110.C.02.F.03.B permet au message "Analyse numérique. Informati-
-que. Automatique. [...] Théorie du codage et du décodage" de passer plus facilement. La notion de bruit au sens de la T.M.C. passe cepen-
-dant au second plan car les imperfections d'une classification sont principalement d'ordre sémantique : il ne suffit pas que la symboli-
-sation soit simple et élégante, encore faut-il que l'architecture de la table soit complète, logique et judicieuse. C'est le message même qui est en cause. Plus largement, les imperfections du système publi-
-cation - classification - lecteur sont surtout d'ordres sémantique et pragmatique (problèmes 2 et 3 de WEAVER).

Dans ces conditions, et compte tenu du maillon "lecteur" de la chaîne de communication, il semble difficile de dépasser le stade des tests "stimulus codé / information transmise" pratiqués par les psychologues et permettant par exemple de quantifier la "capacité de voie" d'un sujet (80).

C. DIFFICULTE D'ABORDER LES PROBLEMES DE SIGNIFICATION

Il s'agit là d'une question aussi vieille que la T.M.C. elle-même, et qui concerne au plus haut point les sciences de l'information.

Certes, les messages à transmettre ont généralement un sens, mais il n'y a aucun rapport entre ce sens et le problème technique de télécommunication que pose la transmission des messages.

La mesure de l'information est essentiellement quantitative : elle repose sur la probabilité d'apparition des événements qui constituent le message, que ces événements pris séparément ou globalement aient un sens ou non. Les possibilités d'extension de la T.M.C. aux ques--tions touchant la signification ont été développées par BAR-HILLEL (7) et MACKAY (69) principalement, dont les recherches assez générales pourront nous servir de point de départ.

Les scientifiques de l'information, comme les chercheurs en commu--nication humaine et les psychologues de la perception, ont été sensibles depuis longtemps à cette limitation (57), ce qui a conduit certains auteurs à rechercher une mesure de l'information totalement différente de celle de SHANNON. En plus de l'information généra--lisée, notons en particulier la mesure de M. C. YOVITS (111) liée à la notion de prise de décision et basée sur la probabilité, pour un décideur, de choisir une ligne d'action suite à la réception d'une information*.

* *On trouvera dans N. J. BELKIN (8) une étude générale des concepts d'information.*

II. NOTIONS UTILES

Malgré les difficultés d'établissement des bases d'une expérience, le sous-emploi du codage et l'abandon peut-être pas irrémédiable des ques--tions de signification, un certain nombre de notions créées par la théo--rie de SHANNON, ou utilisées par elle, aident déjà ou peuvent aider les scientifiques de l'information.

On peut citer, en particulier :

- le modèle général de la communication,
- le calcul des probabilités,
- les grandeurs caractérisant une voie avec bruit,
- la redondance.

A. MODELE GENERAL DE LA COMMUNICATION

Le schéma classique d'un processus de communication, plus ou moins enrichi selon les exemples particuliers d'expériences, sert maintenant de support visuel à un très grand nombre d'études et de recherches. Il permet de visualiser un mécanisme d'interrelations et de déplace--ment des flux d'informations, sensibilisant ainsi élèves, praticiens et chercheurs à un mode de pensée synthétique pouvant stimuler l'imagination par des analogies fécondes si elles restent contrôlées. Il sensibilise également à des types de raisonnement appartenant au domaine de la cybernétique et de la théorie des systèmes, et peut ainsi familiariser le chercheur en sciences de l'information avec de nouveaux outils et le conduire, par ricochets, à une vision pluridisci--plinaire de certains problèmes perçus jusqu'alors par lui comme spécifiques des sciences de l'information.

B. CALCUL DES PROBABILITES

La T.M.C. s'appuie sur la notion de fréquence d'un événement, notion menant à celle de probabilité.
Une des difficultés d'application de la T.M.C. aux sciences de l'infor--mation réside dans la définition de la source d'information et de la voie (44). Il en découle tout naturellement qu'une fois la source choisie, il faut associer des probabilités aux événements composant les messages qui en sont issus.
Chaque fois donc qu'une étude de sciences de l'information se voudra quantitative, il sera nécessaire de pouvoir doter les objets ou phéno--mènes étudiés de probabilités, ce qui conduira éventuellement à une réflexion fructueuse sur la nature de ces objets ou phénomènes.

C. GRANDEURS CARACTERISANT UNE VOIE AVEC BRUIT

S'il semble difficile d'étendre la notion de codage à l'étude d'un certain nombre d'objets et de phénomènes documentaires, il reste la possibilité d'assigner à ces derniers des probabilités et, par consé--quent de déduire des mesures numériques de la quantité d'informa--tion :

1) du message d'entrée choisi,
2) du message de sortie choisi,
3) du système message d'entrée - message de sortie,
4) transmise dans la voie,
5) du message de sortie quand le message d'entrée est connu (ambiguïté),
6) du message d'entrée quand le message de sortie est connu (équivocation).

Ces trois dernières fonctions étroitement liées au bruit n'ont jusqu'ici guère été utilisées en sciences de l'information - BELZER et CAW--KELL mis à part - et on peut le regretter : la notion de bruit per-

-met en effet de rendre compte de phénomènes complexes d'inter--dépendance et conduit même les biologistes à envisager le bruit comme un facteur d'ordre et un générateur d'information (4) (5) (6). D'une façon schématique, le bruit est partiellement interprété par les biologistes de la façon suivante :

- Une transmission sans bruit se traduit par une réplication pure et simple du message émis, sans évolution.
- Une transmission totalement couverte par le bruit se traduit par une absence de lien entre message émis et message reçu.
- Une transmission avec bruit se traduit par une réplication impar--faite : la déformation partielle du message émis est un gage d'évolution.

E. REDONDANCE

Cette notion est déjà fort familière aux scientifiques de l'information et de la communication. Elle gagne à être quantifiée et permet de relativiser les mesures informatives puisqu'elle se présente comme un quotient. Certaines précautions s'imposent néanmoins. Le fait que nous ayons une connaissance intuitive de la notion de redondance ne doit pas nous faire perdre de vue son caractère complexe. La redondance peut se manifester de diverses façons : ajout de symboles, dépendance de symboles successifs, remplacement de symboles par des supersymboles.

De même que "information transmise", "ambiguïté" et "équivocation", le terme de redondance ne correspondra pas nécessairement au sens habituel du mot dans le langage courant.

APPLICATIONS DE LA THEORIE MATHE-MATIQUE DE LA COMMUNICATION DANS LE DOMAINE DES BASES DE DONNEES

I. LE CONTEXTE DES BASES DE DONNEES

L'application de la T.M.C. au domaine des bases de données documen--taires est demeurée jusqu'à maintenant un thème largement inexploré. On peut l'expliquer en partie par l'insuffisance de données quantitatives sur le contenu lexical des bases, ainsi que par le coût élevé de toute expérimentation en grandeur réelle. Nous nous sommes efforcés d'aborder la question sous un angle essentiellement pratique à partir des données dont nous disposions. Les mesures ici proposées seront basées sur l'expé--rience concrète d'interrogation en mode dialogué des bases de données. Celles-ci permettent en effet d'associer un nombre d'occurrence à tout terme faisant l'objet d'une question : on est ainsi tout naturellement conduit à utiliser les fréquences qui en découlent dans les équations fon--damentales de la T.M.C.. Ces fréquences seront obtenues en effectuant le quotient du nombre d'occurrence des notices comportant l'objet de la question par le nombre total des notices présentes en mémoire.
La base de données se présente en fait comme une importante collection d'objets "sélectionnables" et "combinables" dont chacun est caractérisé par le nombre d'occurrence des notices le comprenant.

II. QUANTITE D'INFORMATION D'UNE NOTICE

A. PRISE EN COMPTE DES MOTS INFORMATIFS

Prenons une base de données comprenant en tout N références. Chacune de ces références apporte à la base de données un certain nombre de mots informatifs * présents dans les champs "titre", "adresse" et "descripteurs" (cas de la base PASCAL) et fait apparaître la base comme un vaste corpus de mots informatifs. Chacun de ces mots informatifs est caractérisé au sein du corpus par le nombre

** On retiendra seulement ce que R. ESCARPIT appelle les mots notionnels ou mots informatifs, à l'exception des mots-outils jouant un rôle unique--ment syntaxique ou opératoire ([35]). L'ordre de succession des mots n'in--tervient pas et ceux-ci pourront être utilisés sous leur forme tronquée. Bien qu'une étude sur la T.M.C. privilégie par nature le point de vue statistique, il est difficile d'échapper tout à fait à un certain nombre de questions linguistiques dont certaines sont exposées plus loin en D.*

f(i) de références comprenant le mot choisi. Ce nombre d'occurrence est concrètement indiqué lors d'une sélection simple.

Exemple pris dans la base PASCAL contenant environ 4,5 millions de notices au début de l'année 1983 :

S ELECTROLUMINESCEN?

1 1140

Le nombre d'occurrence associé à ELECTROLUMINESCEN? sous sa forme tronquée est

$$fr(1) = \frac{f(1)}{N} = \frac{1140}{4500000} = 253\ 10^{-6}.$$

Une notice-document constituée de n mots informatifs possédant chacun un nombre d'occurrence relatif

$$fr(i) = \frac{f(i)}{N},$$

que nous noterons p(i) par la suite, peut être considérée comme un message selon la correspondance d'ensemble :

Base de données : Source
Document : Message
Mot informatif : Symbole
Fréquence : Probabilité

Prenons l'hypothèse simplificatrice où les n mots du document - du titre en l'occurrence - constituent des événements aux probabilités de survenue statistiquement indépendantes.

On peut définir la quantité d'information - au sens de la T.M.C. - du document par une formule analogue à celle de SHANNON. Cepen--dant, l'application pratique d'une telle formule suppose qu'on prenne certaines libertés avec la T.M.C..

Les principales différences avec les conditions habituelles d'appli--cation de la théorie sont, en effet :

1) Objet des fréquences.

Les fréquences ne sont pas à proprement parler celles des symboles eux-mêmes au sein de la source, mais celle des notices possédant les symboles. On peut cependant estimer qu'il y a concordance

entre les deux types de phénomènes, l'apparition fréquente d'un symbole se traduisant par l'apparition fréquente de notices le comprenant.

2) Particularité des messages.

Alors que la formule de SHANNON doit représenter la quantité d'information moyenne par symbole dans l'ensemble des messages utilisant un jeu donné de symboles, on ne se propose de l'appliquer en fait qu'à un message isolé. Si on choisit un document D dans la base de données, on ne peut en effet le considérer que comme un micro-message peu représentatif de la source. Ce micro-mes--sage ne résulte que de l'émission de quelques mots par rapport à une centaine de millions, en admettant que chacune des 4,5 millions de notices de PASCAL contient une vingtaine de mots informatifs en moyenne, en comptant titre, adresse et descripteurs. En fait, aucun document présent dans la base n'est assez long pour apparaître comme un reflet du contenu de cette base.

3) Application des probabilités.

Les valeurs des nombres d'occurrence sont généralement très fai--bles, ce qui peut rendre délicate l'identification des fr(i) et des p(i). Afin de souffrir le moins possible des distorsions dues à ces faibles valeurs, on aura toujours intérêt à travailler dans de très grosses bases de données et à éviter les fichiers "échantillons" expérimentaux.

4) Normalisation de la fonction entropique.

On ne prend en compte dans les calculs que la somme des quan--tités d'information spécifiques pondérées de chaque mot effec--tivement présent dans le document, de telle sorte que la somme des fréquences n'est plus égale à 1 mais généralement bien in--férieure *. Ce fait rend la fonction entropique très sensible au

* *Si on considère non plus un message isolé mais l'ensemble de la base de données, on constate que la sommation à 1 des p(i) ne peut de toute façon être observée du fait des recouvrements entre mots présents à la fois dans plusieurs notices.*

nombre de symboles et rend moins significative la notion de quan--tité d'information par symbole.

Dans un tel cas, il convient de normaliser la fonction de SHANNON en la divisant par la somme des probabilités des symboles compo--sant le message (90) :

$$H(D) = - \frac{\sum_i p(i) \log_2 p(i)}{\sum_i p(i)} .$$

Notons qu'on retrouve bien la formule habituelle quand $\sum_i p(i) = 1$.

La formule H(D) ainsi définie permet de caractériser chaque document appartenant à une base de données en fonction de la fréquence d'apparition associée aux termes qui le composent et d'effectuer sur le document un certain nombre de calculs appli--cables aux messages discontinus.

B. PRISE EN COMPTE D'AUTRES CHAMPS INTERROGEABLES

Toute sélection d'un terme i sur un champ interrogeable donne nais--sance à un ensemble comprenant un certain nombre de notices pou--vant servir de nombre d'occurrence f(i).

On peut disposer, par conséquent, de tout un jeu de quantités d'infor--mation différentes définies par rapport à chacun de ces champs et obtenues par des quotients $\frac{f(i)}{N}$.

Exemple de champs interrogeables : base PASCAL dans le système QUEST (Agence spatiale européenne) :

1) auteurs : préfixe AU=
2) affiliations : suffixe /CS
3) mots des titres selon la langue :
   a) suffixe /TI pour tous titres,
   b) suffixe /ET pour titres anglais,
   c) suffixe /FT pour titres français,
   d) suffixe /GT pour titres allemands,
   e) suffixe /OT pour autres langues,

4) codes de classification : préfixe CC=
5) descripteurs : suffixe /CT
6) langues : préfixe LA=
7) types de document : préfixe DT=
8) sources : préfixe JN=

Le champ "codes de classification" pourra faire l'objet d'applications particulièrement intéressantes, en complément de celles découlant de la sélection des mots informatifs.

C. AFFINEMENT DE LA MESURE DE H(D)

Par commodité de calcul, nous avons retenu l'hypothèse simplificatrice de l'indépendance des symboles composant le message. Il est cependant difficile de considérer les mots d'un document, par exemple, comme des éléments statistiquement indépendants.
Il est sûr, par exemple, qu'une relation de dépendance unit les mots suivants susceptibles d'apparaître simultanément dans un même document :
CELLULE , ADN ;
CRISTAL , RESEAU ;
COMPLEXE , LIGAND ; etc.
Il s'agit là d'exemples flagrants. En fait, les n mots d'un titre ou d'une zone de descripteurs sont liés plus ou moins fermement les uns aux autres selon que leurs champs sémantiques s'interpénètrent plus ou moins *.
Une formule plus générale de H pourrait découler du principe que la quantité d'information du document est la somme des quantités d'information conditionnelles du document déterminées par la présence de chaque terme et pondérées par la probabilité d'apparition de ce terme :

* *Il conviendrait de tenir compte de l'ordre de succession des mots dans un titre, dont la suite - comme tout texte intelligible - constitue un processus markovien.*

$$H_r(D) = \frac{\sum_i p(i)\, H(D|i)}{\sum_i p(i)} \qquad \text{avec} \quad H(D|i) = -\frac{\sum_j p(j|i) \log_2 p(j|i)}{\sum_j p(j|i)} \; ;$$

d'où éventuellement une formulation d'une grandeur apparentée à la redondance mais difficile à justifier si on s'éloigne trop des conditions d'application des équations générales de la T.M.C. :

$$R = 1 - \frac{H_r(D)}{H(D)} .$$

D. PROBLEMES LINGUISTIQUES

1) Traitement des mots.

L'analogie entre une base de données et un corpus de mots infor--matifs fait apparaître deux questions particulières : l'élimination des mots vides et la normalisation des mots présents sous des formes différentes.

a) Elimination des mots vides.

On peut distinguer deux grandes catégories de mots vides : les mots-outils et les mots informatifs de faible poids séman--tique (70) (27).

- Les mots-outils sont principalement les articles ; les adjectifs démonstratifs, possessifs, interrogatifs, indéfinis ; les pro--noms ; les prépositions ; les conjonctions. Ces mots jouent, la plupart du temps, un rôle syntaxique.

  Les mots ordinaires peuvent apparaître comme une catégorie particulière de mots-outils jouant un rôle opératoire : ce sont les auxiliaires (être, avoir et toutes leurs formes conju--guées) et certains adverbes.

  Les mots informatifs sont tous les autres mots (noms com--muns, noms propres, verbes, adjectifs, ...).
- Les mots informatifs de faible poids sémantique sont ceux qui apparaissent dans un très grand nombre de références et qui, de ce fait, ont un contenu sémantique faible. Si ces derniers jouent un rôle neutre en première analyse, leur élimi-

-nation n'est pas forcément utile et peut même conduire à une baisse du taux de rappel (102).

D'une façon générale, l'élimination des mots vides est à présent couramment résolue par traitement informatique pouvant asso--cier critères grammaticaux (liste de catégories grammaticales vides) et morphologiques (liste de mots vides) (40).

b) Normalisation des mots.

La normalisation des formes différentes d'un même mot se traduit par la troncature et le traitement de certains mots à orthographes multiples.

- La troncature des mots de même radical permet de faire intervenir sous une forme commune tous les dérivés à rôle sémantique identique d'un même radical, certains de ces dérivés étant fréquemment employés, d'autres plus rarement. On pourra ainsi éviter d'éventuelles aberrations dues à une trop faible fréquence attachée à ces derniers dérivés quand ils se trouvent isolés de leur famille, et obtenir des résultats d'occurrence et de co-occurrence plus significatifs.

  Exemple : analyse, analyses, analysée, analytique, ... réduits en ANALY?.

  Comme dans le cas de mots vides, la coupure automatique des désinences ou la réduction à une forme-mère est couram--ment utilisée dans des expériences d'indexation automatique (40) (95) ou d'analyse statistique de texte (105).

- Un tel traitement laisse de côté les mots à orthographe dif--férente non réductible à un tronc.

  Exemple : Un même mot sous forme de sigle et sous forme développée : PAC et pompe à chaleur.

  Un même mot composé sous forme liée ou sous forme con--tractée : semi-conducteur et semiconducteur.

  On peut, soit utiliser les mots tels qu'ils se présentent - l'uniformité de la distribution statistique d'une forme à l'autre pouvant ne pas trop affecter les résultats - soit les rassembler en paquet par union logique OU pour la même raison que pré--cédemment, soit adopter la forme normalisée par les lexiques PASCAL.

2) Jugement de l'approche statistique.

Le principe de l'utilisation des statistiques en analyse de texte n'a pas toujours été bien accueilli. Il apparaît, d'après L. B. DOY--LE (32) trois grandes causes de défiance envers l'approche statis--tique :

a) Il a été difficile pendant longtemps de disposer de textes vrai--ment longs lisibles en machine et de méthodes de traitement informatisé.

b) La rédaction automatique de résumés analytiques, un des pre--miers secteurs d'application de l'approche statistique, n'a pas d'emblée été couronnée de succès. Malheureusement, beaucoup ont blâmé plutôt l'utilisation des statistiques que des attentes irréalistes.

c) Le traitement statistique apparaît comme une méthode indirecte et non naturelle d'analyse du langage. L'analyse syntaxique, d'autre part, est perçue par beaucoup de linguistes comme une méthode directe et donc naturelle.

Ajoutons qu'un autre point de faiblesse du traitement statistique réside dans les incertitudes théoriques justifiant son application. Si on se limite à l'exemple particulier de l'étude des fréquences de mots dans un document, on pourra trouver dans la littérature professionnelle trois avis discordants (93) : certains estiment que les termes porteurs d'information pertinente au document sont ceux qui ont la plus forte occurrence dans ce document, d'autres pensent au contraire que les termes rares peuvent être plus impor--tants, d'autres encore que les termes de fréquence moyenne ont le plus grand "pouvoir de résolution". En fait, on retrouve parfois dans ces opinions divergentes trois attitudes traditionnelles en analyse des systèmes documentaires : privilégier respectivement les forts taux de rappel, les forts taux de précision et atteindre un juste milieu. Dans cette mesure, la discussion devrait porter autant sur les objectifs des études présentées que sur l'utilisation des statistiques.

Bien que MARON estime que les statistiques sont au coeur du problème de l'indexation et de la ressaisie (73), il ne s'agit pas non plus d'affirmer que les statistiques sont la seule clé de la recherche dans ce domaine. On peut cependant insister sur deux points :

a) L'affirmation que sans exploitation statistique, certaines pro--priétés du langage naturel, surtout sous forme de corpus de mots, restent inexpliquées. Les travaux de B. MANDELBROT en apportent la preuve.

b) Les mots isolés ("mots-signaux") étiquetés par des fréquences d'occurrence et de co-occurrence permettent de définir un problème scientifique, voire à l'échelle du texte d'une publi--cation, de baliser le déroulement du texte considéré comme un dispositif de canalisation des intérêts. Une telle théorie excluant l'analyse syntaxico-sémantique sous-tend la démarche méthodologique de l'équipe du Centre de Sociologie de l'Inno--vation dans ses travaux de cartographie des sciences et tech--niques (16).
La priorité donnée par l'équipe au phénomène de co-occurrence correspond au souci de rendre compte de l'aspect réticulaire de la production scientifique et technique, et rejoint la pers--pective de la T.M.C., le mot mettant en relation des contextes et, dans le même mouvement, se mettant en relation avec d'autres mots et d'autres contextes.

Il est donc justifié, sur de telles bases, d'accepter en première analyse le bien-fondé d'une démarche statistique nuancée, un trai--tement syntaxique plus ou moins poussé ne pouvant que faciliter l'harmonisation, l'identification et la connaissance de l'enchaîne--ment des mots-signaux.
On trouvera par exemple dans la thèse de F. DEBILI (29) consa--crée à l'analyse syntaxico-sémantique une perspective d'appli--cation en documentation faisant appel à la fois aux relations lexicales-sémantiques et aux fréquences dans l'étude particulière

des distances phrase - phrase.

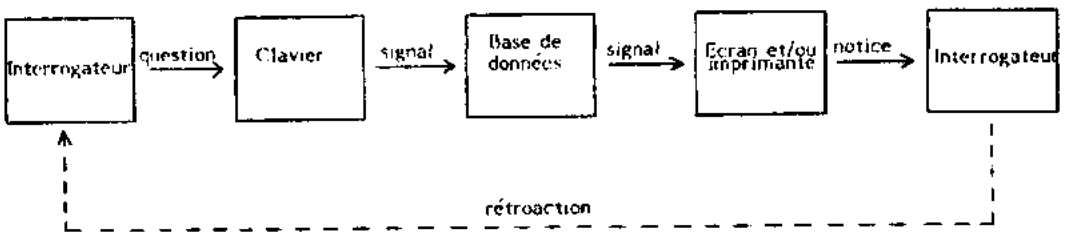
Les démarches proprement linguistique et statistique apparaissent en fait complémentaires et seule la complexité du problème empê--che un dépassement rapide de l'opposition quelque peu artifi--cielle des approches.

III. INTERROGATION D'UNE BASE DE DONNEES

Ayant défini une mesure arbitraire et simplifiée de l'information conte-
-nue dans une notice, il devient possible de tenter d'appliquer la notion d'information à la ressaisie, voire à l'indexation.
L'application ici proposée est abordée sous l'angle du rapport question - réponse dans un système de ressaisie du type probabiliste.

A. RAPPORT MOT-CLE - DOCUMENT

L'interrogation d'une base de données est considérée comme un processus de communication circulaire avec rétroaction, l'ensemble interrogateur - matériel - logiciel fonctionnant comme un système cybernétique chargé de convoyer le maximum d'information selon le schéma ci-dessous :



Le message émis se présente sous la forme d'une question, le message reçu se présente sous la forme de notices bibliographiques (docu-
-ments) comprenant **ou non** les termes de la question, mais suffi-
-samment proches sémantiquement de la question pour être jugées pertinentes.
Soit, par exemple, une question simple constituée d'un seul mot-clé
$Q = \{q\}$

et un document

$D = \{m_1, \ldots, m_j, \ldots, m_n\}$

constitué des n mots informatifs de son titre.

La transmission du message Q par le système de ressaisie est sym--bolisée par un tableau de contingence :

				Emis Question q
				p(q)
Reçu	Document D	m_1	p(1)	p(q,1)
		m_2	p(2)	p(q,2)
		...	....	.....
		m_j	p(j)	p(q,j)
		...	....	.....
		m_{n-1}	p(n-1)	p(q,n-1)
		m_n	p(n)	p(q,n)

Les valeurs figurant dans la colonne sont celles des probabilités con--jointes (quotient par N des nombres d'occurrence) attachées à q et m_j, obtenues par intersection selon l'opérateur logique ET des ensem--bles correspondant aux sélections de q et m_j. Elles mesurent la frac--tion de documents de la base comprenant à la fois les mots q et m_j. Supposons que $q = m_2$. Une transmission parfaite suppose la ressaisie d'un ou de plusieurs documents se composant d'un seul mot informatif - q en l'occurrence - avec p(q,2) = p(q). La survenue de m_1, m_3, ... , m_n est donc considérée comme due au bruit, sans que l'on puisse affirmer a priori que le bruit réduise la pertinence de la réponse. Le modèle reste le même si la question se présente sous forme d'une séquence de m mots-clés

$Q = \{q_1, \ldots, q_i, \ldots, q_m\}$.

Dans ce cas, la transmission est symbolisée par un tableau de contin--gence à n lignes et m colonnes :

			Question Q				
			q_1	...	q_i	...	q_m
			p(1.)	...	p(i.)	...	p(m.)
Document D	m_1	p(.1)	p(1,1)	...	p(i,1)	...	p(m,1)
	...	...	...	...	...	...	...
	m_j	p(.j)	p(1,j)	...	p(i,j)	...	p(m,j)
	...	...	...	...	...	...	...
	m_n	p(.n)	p(1,n)	...	p(i,n)	...	p(m,n)

Le problème étant de convoyer au travers de la voie "base de don--nées" le maximum d'information, la prochaine étape consiste à proposer diverses mesures de couplage question / document tirées de la T.M.C. ou de ses prolongements, que l'on pourra comparer à d'autres mesures d'affinité utilisées en statistique.

B. FONCTIONS DE COUPLAGE

1) Caractéristiques générales.

Les mesures de couplage sont choisies de façon à rendre compte de la proximité entre termes-question

$Q = \{q_1, \ldots, q_i, \ldots, q_m\}$

et termes-réponse

$D = \{m_1, \ldots, m_j, \ldots, m_n\}$

avec les caractéristiques suivantes :

a) La quantification de proximité souhaitée doit apporter une souplesse dans la comparaison question / réponse, souplesse permise par une approche probabiliste qui s'oppose à l'identi-fication habituelle stricte "terme appartenant à la question / terme appartenant au document". Cette dernière procédure, du type tout-ou-rien, correspond en fait à une conception vec-torielle binaire des données : chaque terme de la question ou de la réponse est représenté par une valeur 1 ou 0 dans l'espace des t termes différents répertoriés dans la base de données, et la pertinence d'un document à une question est symbolisée par le partage d'un ou de plusieurs termes identiques.
Les mesures de couplage les plus efficaces devront au contraire permettre une évaluation de la pertinence d'un document vis-à-vis d'une question **même sans identité entre un terme de la question et un quelconque des mots du document.**

b) On attache une importance particulière à la détermination de l'écart entre l'indépendance statistique a priori des messages question et réponse, et leur dépendance statistique observée au sein de la base de données. Une telle démarche s'appuie sur l'hypothèse que l'écart positif par rapport à l'indépendance croît en fonction de la pertinence de la réponse. Il s'agit donc d'introduire dans les mesures, sous une forme plus ou moins directe, la différence entre les probabilités conjointes a priori attachées aux termes $q_i \in Q$ et $m_j \in D$, soit $p(q_i)\ p(m_j)$, et les probabilités conjointes observées, soit $p(i,j)$.
Les mesures seront un reflet de cette différence, compte tenu de la situation particulière suivante, fréquemment rencontrée : au-delà de la stricte indépendance statistique, on pourra observer le cas extrême où $p(i,j) = 0$. Ce cas correspond à une incompa-tibilité de présence conjointe de q_i et m_j au sein d'une notice.

c) Les mesures de couplage doivent être affectées le moins possible par des caractéristiques quantitatives peu ou pas liées à la pertinence : taille des messages en présence, quantité d'infor-mation - dans l'absolu - de ces messages.
A moins de ne comprendre qu'un seul mot informatif, un titre

se présente d'ordinaire comme un message redondant. Il en résulte, par exemple, que la multiplication du nombre de sym--boles ne doit pas nécessairement modifier la force de couplage avec une question. On doit éviter, en particulier, que les titres les plus longs, à pertinence égale, soient systématiquement privilégiés et supposés à tort plus proches de la question. De même, l'association de plusieurs termes à fréquences différentes formant une question ne sera significative que si chacun de ces termes, quelle que soit sa fréquence, intervient autant que possible à égalité dans la comparaison avec la réponse. Dans le cas contraire, une question complexe comportant plusieurs termes se ramènerait à quelques nuances près à la question simple formée par le terme dominant et perdrait de ce fait beaucoup de son intérêt. La résolution d'une telle contrainte suppose la relativisation de la mesure de couplage en fonction de caractéristiques propres aux messages en présence, comme leur quantité d'information (30).

Remarque :

On peut établir un parallèle entre le principe du couplage question / document et la théorie "épidémiologique" de W. GOFFMAN et V. A. NEWILL (48). Dans cette théorie, la ressaisie de l'informa--tion est déterminée par une mesure probabiliste ξ de contact effectif entre une question Q (agent contaminable) et l'ensemble des documents D (agents contaminants).

La réponse optimale à la question est obtenue pour une valeur de ξ supérieure à un seuil ξ_0. La formule de la mesure n'est d'ailleurs pas précisée, les auteurs se cantonnant à la formali--sation d'un modèle théorique général.

2) Fonctions dérivées de T.

a) Quantité d'information transmise relativisée.

Si on normalise les fonctions entropiques, l'information appor-tée par la comparaison entre question et document est cal-culée de la façon suivante :

$T(Q;D) = H(Q) + H(D) - H(Q,D)$, avec :

$$H(Q) = - \frac{\sum_i p(i) \log_2 p(i)}{\sum_i p(i)} ,$$

$$H(D) = - \frac{\sum_j p(j) \log_2 p(j)}{\sum_j p(j)} ,$$

$$H(Q,D) = - \frac{\sum_{i,j} p(i,j) \log_2 p(i,j)}{\sum_{i,j} p(i,j)}$$

La sommation s'effectue sur les mots informatifs q_i compo-sant la question Q et les mots informatifs m_j composant le titre du document D, les mots-outils étant négligés.

La relativisation de T conduit à diverses formules proposées dans des expériences de décomposition de systèmes complexes (30) et faisant intervenir les quantités d'information de Q et de D (M_1 et M_3), ou bien celle conjointe de Q et D (M_2), ou encore celle unique de Q (M_4) ;

$M_1 = \frac{T(Q;D)}{\sqrt{H(Q)\,H(D)}}$, coefficient de RICHETIN (91),

$M_2 = \frac{T(Q;D)}{H(Q,D)}$, coefficient de DUSSAUCHOY (34),

$\frac{T(Q;D)}{H(Q) + H(D)}$, coefficient de DUFOUR (33),

$M_4 = \frac{T(Q;D)}{H(Q)}$, coefficient de CONANT (23).

M_1, M_2 et M_4 prennent, dans les conditions normales d'utili-sation de la T.M.C., leurs valeurs dans l'intervalle [0,1], tandis que le coefficient de DUFOUR est compris entre 0 et 1/2. Le doublement de ce dernier permet d'obtenir un coefficient M_3 prenant également ses valeurs dans [0,1], dans les mêmes conditions d'utilisation :

$$M_3 = \frac{2\ T(Q;D)}{H(Q) + H(D)} .$$

Dans les conditions normales d'application de la T.M.C., on obtient exactement le même résultat si on calcule T par la différence de l'information à la sortie et de l'ambiguïté :

$T = H(D) - H(D|Q)$, avec :

$$H(D|Q) = - \frac{\sum_{i,j} p(j|i) \log_2 p(j|i)}{\sum_{i,j} p(j|i)} \quad , q_i \text{ étant donné.}$$

T peut être relativisée selon les 4 formules précédentes :

$$M'_1 = \frac{T}{\sqrt{H(Q)\ H(D)}}$$

$$M'_2 = \frac{T}{H(Q,D)}$$

$$M'_3 = \frac{2\ T}{H(Q) + H(D)}$$

$$M'_4 = \frac{T}{H(Q)}$$

qui apportent des résultats différents des précédents.

b) Mesure d'information mutuelle.

Une fonction d'association entre variables couramment préco-nisée en classification automatique est dérivée de concepts informationnels (55) (108) : la mesure d'information mutuelle I(i;j) permet de quantifier la dépendance statistique de deux objets :

$$I(i;j) = \log_2 \frac{p(i,j)}{p(i)\ p(j)}$$

Si q_i et m_j sont statistiquement indépendants, $p(i)\ p(j) = p(i,j)$, d'où $I(i;j) = 0$. $I(i;j)$ est d'autant plus élevé que q_i et m_j sont statistiquement liés. La "mesure d'information mutuelle" revient à l'évaluation de l'information statistique contenue en i sur j, ou vice versa, donc de l'écart par rapport à l'indé-

-pendance statistique. Appliquée au cas de la liaison question / document, la mesure d'information mutuelle s'écrit :

$I(Q;D) = \sum_{i,j} p(i,j)\ I(i;j)$, soit :

$$I(Q;D) = \sum_{i,j} p(i,j) \log_2 \frac{p(i,j)}{p(i)\ p(j)} .$$

Dans les conditions classiques d'utilisation de la T.M.C., I et T sont identiques.

3) Autres fonctions.

Il existe dans la littérature un certain nombre de fonctions de la fréquence de co-occurrence de deux symboles i et j pouvant exprimer une affinité. Nous en choisirons trois particulièrement répandues, en conservant présente à l'esprit la restriction évo-quée en page 71 sur l'objet des fréquences.

a) Fonction cosinus.

Cette fonction a été notamment employée dans l'hypothèse vectorielle binaire où un document apparaît comme un vecteur dans l'espace à d dimensions des d descripteurs (108).

Transposée dans un modèle probabiliste, elle s'écrit (94) :

$$C(i;j) = \frac{p(i,j)}{\sqrt{p(i)\ p(j)}} .$$

Avec, dans le cas général :

$$C(Q;D) = \sum_{i,j} \frac{p(i,j)}{\sqrt{p(i)\ p(j)}}$$

b) Fonction de JACQUARD.

Cette fonction, qui correspond au quotient

$$\frac{p(i \text{ ET } j)}{p(i \text{ OU } j)} ,$$

a été utilisée par différents auteurs (32) (59) (94) (92) dans des expériences de classification automatique :

$$J(i;j) = \frac{p(i,j)}{p(i) + p(j) - p(i,j)} .$$

D'où

$$J(Q;D) = \sum_{i,j} \frac{p(i,j)}{p(i) + p(j) - p(i,j)} .$$

c) Coefficient de corrélation.

GUIAŞU et THEODORESCU (51) proposent le coefficient de corrélation R comme mesure d'affinité entre deux sym--boles :

$$R(i;j) = \frac{p(i,j) - p(i)\,p(j)}{\sqrt{p(i)\,p(i^C)\,p(j)\,p(j^C)}}$$

Ce coefficient faisant intervenir les complémentaires de i et j peut être simplifié. Du fait de la faiblesse des fréquences rencontrées, on posera $p(i^C) = p(j^C) \# 1$ et on aboutira à la formule de la "contingence effective" (78). Dans le cadre de l'affinité question / document, ce coefficient peut s'écri--re :

$$R(Q;D) = \sum_{i,j} \frac{p(i,j) - p(i)\,p(j)}{\sqrt{p(i)\,p(j)}} \quad .$$

d) Formule du khi-deux.

Mentionnons pour mémoire les formules du type khi-deux (106) :

$$K(Q;D) = \sum_{i,j} \frac{[p(i,j) - p(i)\,p(j)]^2}{p(i)\,p(j)} \quad .$$

4) Information mutuelle en information généralisée.

Les travaux de J. LOSFELD (66), inspirés comme ceux de F. FOREST du formalisme de KAMPE DE FERIET, proposent une mesure de l'information construite à partir des probabilités. Cette mesure respecte les trois axiomes énoncés plus haut en page 54 (paragraphe sur l'information hyperbolique) ainsi que deux hypothèses relatives à la mesure de l'information et non admises par la T.M.C. :

H1 : $I(i) = \sum_h p(h)\, I(i|h)$, h représentant les différents paramètres d'observation de l'événement i, que la T.M.C. ne prend pas en compte, du moins aussi directement.

H2 : $I(i,j) = f[p(i),p(j),p(i,j)]$, alors que la T.M.C. ne fait dépendre l'information conjointe que de la probabilité conjointe p(i,j).

Dans ce cadre et au terme d'un développement très rigoureux, LOSFELD constate que seule une information du type

$$I(i) \geq K \left[\frac{1}{p(i)} - 1 \right]$$

vérifie les hypothèses 1 et 2. Il aboutit à une mesure $\mathscr{L}(i;j)$ de l'information mutuelle de deux événements de probabilités p(i) et p(j), et de probabilité conjointe p(i,j) :

$$\mathscr{L}(i;j) = \frac{p(i,j)}{p(i)\ p(j)} - 1 \ , \quad \text{avec} \quad p(i) > 0,\ p(j) > 0,\ p(i,j) \geq 0.$$

Si on poursuit le raisonnement sur les mêmes bases, l'information mutuelle entre les ensembles

$Q = \{q_1, \ldots, q_i, \ldots, q_m\}$ et

$D = \{m_1, \ldots, m_j, \ldots, m_n\}$

se présente comme la quantité moyenne des $\mathscr{L}(i;j)$ obtenue par pondération :

$$\mathscr{L}(Q;D) = \sum_{i,j} p(i,j) \left[\frac{p(i,j)}{p(i)\ p(j)} - 1 \right].$$

L'application de cette formule au cas de la liaison question / document au sein d'une base de données fait toutefois apparaître un double obstacle. En effet, une telle formulation privilégie les probabilités conjointes p(i,j) fortes dans l'absolu par rapport aux faibles p(i,j), alors que dans le contexte des bases de données la présence conjointe de mots n'a aucune valeur dans l'absolu : Si les termes CHIMI? (p = 387548 / 4500000) et SYSTEM? (p = 342948 / 4500000) apparaissent 29663 fois ensemble (p(i,j) = 29663 / 4500000), cette co-occurrence n'a pas forcément plus d'importance que celle de BASE?(W)DONNEES (p = 59 / 4500000) et BIBLIOTHE? (p = 11610 / 4500000) qui ne se produit que 2 fois (p(i,j) = 2 / 4500000). En documentation, la fréquence de co-occurrence résultant de l'intersection de deux "événements" à faible fréquence peut avoir d'autant plus d'importance qu'on a affaire à des mots-clés très spécifiques et à fort pouvoir sémantique. De plus, on aurait pu songer à relativiser $\mathscr{L}$ comme cela a été fait pour T afin d'obtenir différents coefficients. Une telle relativisation par les quantités d'information ne ferait qu'accentuer le déséquilibre au profit des fortes p(i,j), pour la raison suivante : ces dernières sont souvent issues de p(i) et p(j) fortes dans l'absolu. Hors, si on admet une valeur

$$I(i) = \frac{1}{p(i)} - 1$$

pour l'information, I(i) décroît quand p(i) croît. Si on considère l'ordre de grandeur des fréquences rencontrées dans les bases de données, la mesure de l'information est proche de $\frac{1}{p(i)}$. Il en découle que le quotient par I(i) revient en gros à un produit par p(i).

En fait, la pondération des $\mathscr{L}(i;j)$ par p(i,j), bien que s'inscrivant dans la logique rigoureuse des hypothèses de base développées par LOSFELD, ne peut - dans le cas particulier d'application que nous avons retenu - que limiter la qualité de la fonction de couplage*.

C'est pourquoi nous placerons à égalité tous les couples de mots i et j pour ne retenir qu'une forme cumulée des $\mathscr{L}(i;j)$ sans pondération :

$$L(Q;D) = \sum_{i,j} \frac{p(i,j) - p(i)\,p(j)}{p(i)\,p(j)} \quad .$$

C. DIVERSES FORMES DE QUESTIONS

Rien ne distingue dans le modèle probabiliste et sur le plan formel la question du document. Tous deux se présentent comme des séquences de symboles. C'est pourquoi le couplage question / document - qui est en fait un couplage message / message - peut s'appliquer à une question présentée de diverses façons.

1) Rapport code de classification / document.

La transmission pourra ainsi être symbolisée par un tableau de contingence reliant un code de classification C à un docu--ment D et excluant par définition toute possibilité de trans-

* *On pourrait d'ailleurs faire la même remarque au sujet de la fonction I(Q;D).*

-mission sans bruit.

			Emis Question C
			p(c)
Reçu Document D	m_1	p(1)	p(c,1)
	...	...	...
	m_i	p(1)	p(c,i)
	...	...	...
	m_n	p(n)	p(c,n)

De la même façon que précédemment, les valeurs de probabilités conjointes de c et i sont obtenues par sélection de c, puis sélec-tion de m_i, puis intersection des ensembles ainsi obtenus. Le choix d'une mesure de couplage efficace devrait pouvoir conduire à des expériences de classification automatique selon un mode opératoire inspiré de celui de MARON ([72]). Une telle procédure "par attribution" apparaît d'ailleurs davantage assi-milable à une assistance à la classification qu'à une solution définitive et auto-suffisante, car les codes (catégories) sont figés, ainsi qu'en partie les relations statistiques codes / mots. D'où la nécessité d'une intervention manuelle permettant de refléter l'évolution des connaissances (entrée de termes nou-veaux, redécoupage du plan de classification).

2) Comparaison de deux documents.

La détermination des liens d'affinité sémantique entre deux documents se présente exactement de la même façon que celle du rapport question / document dans le cas général (matrice à n lignes et m colonnes). Les valeurs obtenues par les mesures de couplage devraient permettre d'assembler les notices par agrégats selon des méthodes de classification automatique. On peut également imaginer une transposition du lien document / document à l'évaluation de la qualité d'un titre : si on consi-

-dère que le résumé est un reflet fidèle du contenu d'une publi--cation, l'affinité entre le titre et le résumé doit être la plus grande possible*. On pourra symboliser la transmission d'infor--mation entre résumé et titre selon un tableau de contingence prenant en compte tous les mots informatifs du résumé et du titre.

Des mesures tout à fait analogues pourraient s'appliquer à un ensemble de descripteurs considéré comme un document, afin d'évaluer leur pertinence par rapport à un titre ou un résumé.

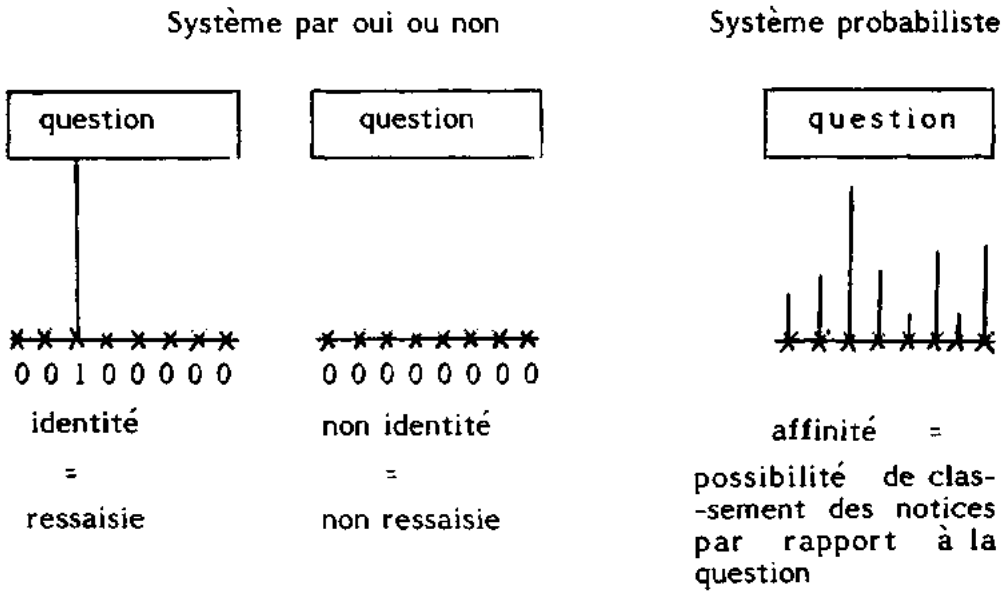
Remarque :

Les index de l'I.S.I., construits sur le principe de l'analyse de citations, ont permis un comptage de co-citations, ainsi que la constitution d'agrégats (46). Il serait intéressant de comparer les résultats obtenus avec ceux que peut apporter la théorie de l'information, notamment dans l'évaluation de l'affinité séman--tique de deux documents. On dispose en effet, dans ce domaine, d'une fréquence de co-citations qui, relativisée, devrait concor--der avec une mesure adéquate de couplage obtenue par les fréquences de mots. De la même façon, un tableau de contin--gence basé sur les fréquences de citations et de co-citations devrait déboucher sur des résultats significatifs qu'il serait intéressant de comparer avec ceux découlant d'un comptage relativisé des co-citations.

* *Du moins dans le domaine scientifique et technique où le titre a une fonction "informative" et utilitaire unanimement reconnue et souhaitée, même si les résultats sont parfois inégaux.*

D. CONCLUSION

La mesure brute de la quantité d'information d'un message ne semble pas apporter par elle-même d'élément significatif. Par contre, les mesures liées au bruit ou, plus généralement, aux phéno--mènes de dépendance statistique doivent pouvoir décrire diverses affinités et interrelations utiles en sciences de l'information, bien que l'on s'éloigne des conditions strictes d'application de la T.M.C., voire de la théorie originelle elle-même au bénéfice de résultats tirés de la théorie généralisée. Un certain nombre d'essais sont détaillés dans la partie expérimentale et permettent une première comparaison des fonctions proposées. On peut en particulier constater que les fonctions directement tirées de la T.M.C., bien qu'exprimant une relation de pertinence, n'apparaissent pas comme les plus efficaces. Dans ce paragraphe sur le contexte des bases de données, l'accent a été mis sur les problèmes de ressaisie de l'information. Une telle façon de concevoir la ressaisie répond à la notion intuitive de spectre ou à celle plus complexe d'hologramme ([86]). Alors qu'une ressaisie classique par oui ou non laisse de côté tout terme différent de la question, la ressaisie de type probabiliste permet de rendre compte de l'affinité sémantique plus ou moins grande entre la question et chaque terme de chaque document, selon la représentation schématique :



Cette démarche pourrait être d'autant plus justifiée que le système documentaire est plus pauvre en mots informatifs (base de données sans descripteurs additionnels par rapport au titre, par exemple). Nous nous sommes limités à quelques hypothèses bien déterminées d'application du couplage question / document. Il est cependant possible d'envisager l'extension du modèle à des "messages" appar--tenant à divers champs interrogeables dans les bases de données, afin d'aborder par exemple des problèmes de structure et de socio--logie de la recherche scientifique, avec au besoin l'intervention du facteur temps.

RELATION AVEC LA SIGNIFICATION

Les limites de la T.M.C. ont été soulignées dès le début par WEAVER : seul le problème technique de l'exactitude de transmission des symboles utilisés ressortit à la théorie de SHANNON. Les questions sémantiques et pragmatiques sont écartées et ne se manifestent que dans la mesure où la structure même du jeu de symboles (les données) reflète les niveaux sémantique et pragmatique du processus de communication.
Ces limites ont tout naturellement entraîné des déceptions. On peut toute--fois remarquer que ces déceptions sont souvent formulées à l'occasion d'études assez générales d'épistémologie de l'information ou de communi--cation humaine (6) (104) qui dépassent largement la problématique con--crète des techniques documentaires. En fait, la question demeure large--ment ouverte dans le domaine des sciences de l'information et il nous semble utile de la réexaminer avec un regard neuf à la lumière d'un certain nombre de démarches employées dans d'autres domaines.

I. CONTEXTE DOCUMENTAIRE

Nous avons jusqu'à présent traité des grandeurs statistiques indépen--damment de leur contenu.
Peut-on aller plus loin et, par le biais de mesures quantitatives, aborder la notion de signification ? Afin d'apporter quelques éléments de ré--ponse, il convient de délimiter notre démarche.
La question de la signification de l'information est bien trop vaste et complexe pour être abordée de front, et le développement qui va suivre n'est possible qu'au prix d'un certain nombre de simplifications et d'une restriction au seul contexte documentaire.

A. DELIMITATION DE LA DEMARCHE A LA NOTION DE PERTINENCE

Les valeurs absolues de quantité d'information ne nous servent pas directement. Les mesures proposées ne sont utiles que lorsqu'elles rendent compte d'une **pertinence** : pertinence entre question et

document, entre document et document, entre code de classification et document, etc.).

La pertinence se traduit par une dépendance statistique entre mes--sage reçu (document) et message émis (question ou tout autre message "témoin"). Cette dépendance statistique amène une valeur élevée des fonctions du type information transmise ou information mutuelle.

B. INFORMATION ET NOTION DE GAIN

Dans un processus de comparaison message émis (témoin) / message reçu (document), on observe le comportement d'une fonction associée à une dépendance statistique. Un document pertinent étant repéré par une valeur élevée de cette fonction, la pertinence revient à la constatation d'un gain.

Cette idée de gain permet de rejoindre notre expérience intuitive de la notion d'information. C'est pourquoi on retiendra comme définition générale que l'information est un processus de communi--cation de la connaissance (79). Il convient, afin de ne pas restrein--dre le champ d'application de cette définition, de prendre le terme communication dans un sens très large qui déborde la notion de transmission. Cette notion, liée à un modèle directionnel du type source - écoulement - récipient, ne doit pas occulter l'aspect symé--trique du couplage témoin / document : la communication peut être vue en terme plus large de structure statistique d'un système témoin / document.

C. CONNAISSANCE ET SIGNIFICATION

Une communication de la connaissance suppose :

- un contenu : la signification,
- un processus : l'établissement d'une relation.

Nous avons vu que la théorie de l'information en général peut tra--duire quantitativement une dépendance statistique apparaissant dans le cas d'une pertinence, ce qui revient à la manifestation d'une relation entre un message et un autre message.
La signification peut-elle de même se manifester d'une façon ou d'une autre ? Pour cela, examinons trois points :

- sous quelle forme simplifiée ce contenu peut-il apparaître ?
- à qui s'adresse-t-il ?
- a-t-il une valeur absolue ou relative ?

D. VEHICULE SIMPLIFIE DE LA SIGNIFICATION

1) Nature du message.

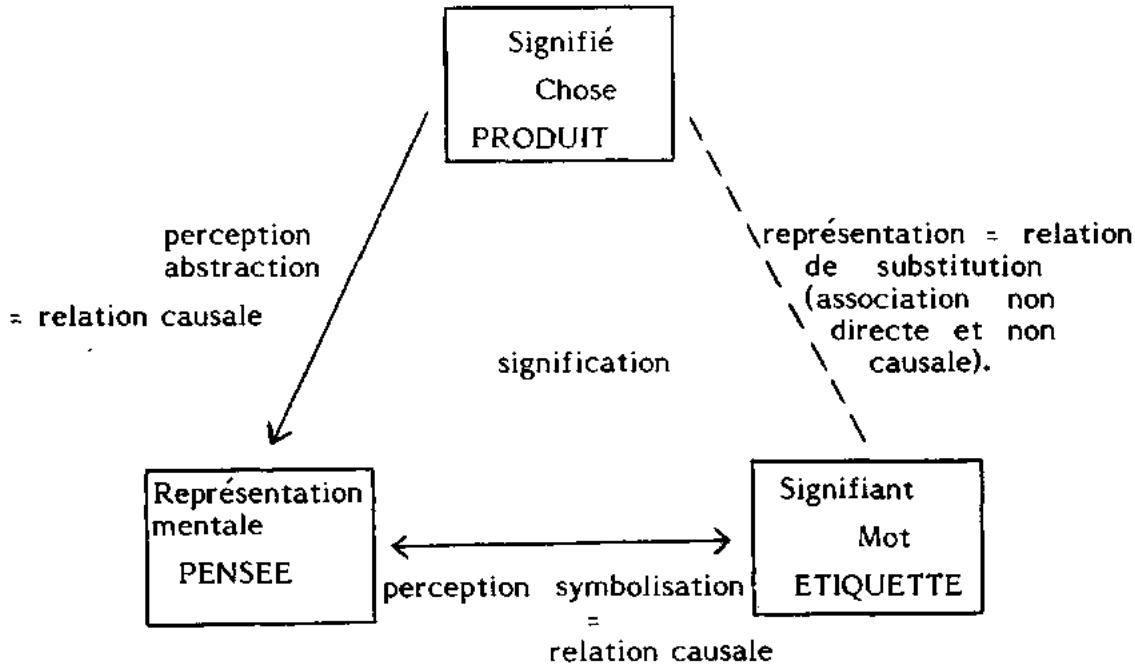
Comme l'a souligné R. ESCARPIT ([35]) il est difficile de consi--dérer la pensée autrement que comme une grandeur continue. Cependant, afin de la prendre en compte dans une analyse infor--mationnelle, on est obligé de la transformer - et malheureuse--ment de la déformer - en une grandeur discrète, c'est-à-dire composée d'unités distinctes.
Or, les seules unités discrètes dont nous disposons pour quantifier l'information contenue dans la pensée sont les unités constitutives du langage. De plus, les sciences de l'information proposent au praticien des langages documentaires facilement quantifiables (mots-clés, codes de classification par exemple). On dispose donc, pour véhiculer la signification, d'unités distinctes relativement simples.

Remarque :

L'utilisation de grandeurs discrètes pour symboliser la pensée est généralement jugée valide dans le cas de l'information de type rationnel et utilitaire, matière première principale des sci--ences de l'information. Par opposition, la pensée de type émo--tionnel apparaît comme une grandeur analogique donc continue (56) (68) (97). Dans ce dernier cas, un modèle plus proche de la réalité devrait vraisemblablement faire appel à une fonction entropique continue si toutefois on estime encore justifié l'emploi de la T.M.C. .

2) Signification.

Les rapports entre unités du langage et signification ont été schématisés par le diagramme triangulaire de OGDEN et RI--CHARDS, diagramme commenté en particulier par C. CHERRY (20) et J. COSNIER (26).



Ce diagramme met en évidence le caractère conventionnel du rapport entre le signifié et le signifiant. C'est pourquoi, par

analogie, nous avons adopté les termes "produit" et "étiquette" : le rapport d'association entre les deux, bien que parfois faussé par des appellations abusives, n'en est pas moins acceptable en première analyse et particulièrement dans une perspective docu--mentaire. Il faut cependant remarquer qu'un tel schéma ne visualise pas directement l'influence du contexte des messages (6) (86), essentiel dans le domaine documentaire. En effet, la signification n'est pas une propriété intrinsèque des messages : non seulement elle dépend du contexte des messages mais encore elle peut être perçue dans certaines conditions comme formée par le jeu même des relations contexte / message.
En fonction de ces remarques, le schéma pourra être partielle--ment adapté à chacun des destinataires de la signification.

E. DESTINATAIRES DE LA SIGNIFICATION

La question posée est la suivante : signification pour qui ?

- Le scientifique de la discipline.
- Le scientifique de l'information.
- Le système bibliographique.

1) Le scientifique de la discipline : l'assimilation.
Son savoir, élaboré au cours d'années d'études et de recherche, intègre toute nouvelle connaissance dans une structure logique comparable à celle d'un volumineux ouvrage de synthèse en per--pétuelle refonte. La signification du contenu d'une connaissance correspond à tout un réseau complexe de liens avec la structure logique. Une quantification apparaît difficile. Supposons que nous puissions accepter le diagramme de OGDEN et RICHARDS : dans ce cas, la signification fait intervenir le résultat de la synthèse d'un très grand nombre de triangles dans le plan de la structure logique.

2) Le scientifique de l'information : le rangement intelligent.

Le savoir, en particulier celui du scientifique de la discipline, est compris par le scientifique de l'information comme un assem--blage de symboles (étiquettes) compatibles avec les langages documentaires. Ce savoir pourrait être représenté par une géo--graphie personnelle de champs sémantiques symbolisés par des clés où tout savoir de toute discipline peut trouver une place, **une fois reconditionné** par traduction dans un ou plusieurs langages documentaires. Les clés de ces langages sont autant d'unités distinctes véhiculant la signification.

Supposons que l'on admette, là encore, le diagramme de OGDEN et RICHARDS. L'application est plus simple que précédemment dans la mesure où on accorde une priorité au sommet "étiquette" du triangle et aux liaisons inter-étiquettes entre divers triangles.

3) Le système bibliographique : le rangement commandé.

Ce qu'on pourrait considérer, pour la facilité de l'exposé, comme son savoir est symbolisé par des clés de langages documentaires assignées par le scientifique de l'information. De plus, apparaît une collection d'objets documentaires appartenant à divers champs interrogeables, répartis selon des fréquences facilement acces--sibles.

Bien qu'il soit incapable d'assimiler une quelconque connaissance, le système bibliographique est intéressant dans la mesure où il se présente comme un reflet non pensant mais apte à quantifi--cation du scientifique de l'information.

F. RAPPORT ENTRE LES DESTINATAIRES

La question posée est la suivante : signification par rapport à quoi ? La compréhension et l'acquisition de connaissance se ramènent dans le cas du scientifique de l'information et du système biblio--graphique à l'établissement d'une correspondance étiquettes-sym--boles / étiquettes-clés.
Cela suppose nécessairement la compatibilité des trois destinataires de la signification.

1) Le scientifique de l'information et le système bibliographique doivent vivre en symbiose, le second servant de mémoire, d'ins--trument de tri et de mise en rapport, ainsi que de comptable, au premier.

2) De plus, le scientifique de la discipline et le scientifique de l'information gagnent à connaître chacun la règle de jeu de l'au--tre. En particulier, le premier doit reconnaître que seul un savoir atomisé en unités et repérable dans un système de référence nécessairement imparfait est pour le moment compatible avec le stockage et la ressaisie de l'information.
La signification véhiculée par les unités constitutives du langage - les mots - est en effet exprimée par la relation :

a) mot / structure logique pour le scientifique de la discipline ;
b) mot / clé de langage documentaire pour le scientifique de l'information.

Le problème revient à l'acceptation par les deux parties de l'équi--valence structure logique - langage documentaire.

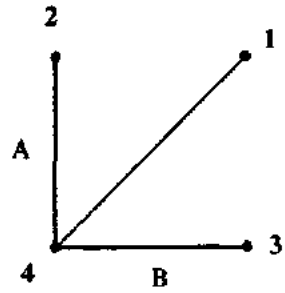
G. CONCLUSION

Nous avons posé qu'un transfert de connaissance suppose un contenu - la signification - et un processus - l'établissement d'une relation. Il ressort de notre analyse que ces deux conditions, **dans le contexte documentaire,** n'en font qu'une : l'établissement d'une relation entre un message et une clé de langage documentaire considérée comme un autre message, ce qui rejoint la remarque de I. J. GOOD : "il existe une étroite analogie entre signification de la signification et mesures de pertinence" (49).

II. APPROCHE PAR LA REPRESENTATION

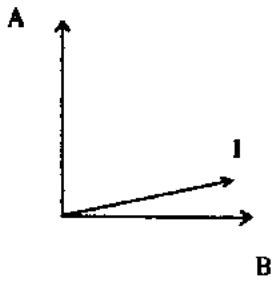
A. REPRESENTATION D'UN CONCEPT

D. M. MACKAY (69) associe étroitement communication, signifi--cation et représentation.
Il est ainsi amené à représenter un concept par la position du som--met d'un vecteur dans un espace de propositions, visualisant ainsi en partie les relations qui unissent le concept à son contexte.
Exemple :
Cas le plus simple : propositions par oui ou non.
Soient deux propositions indépendantes A et B :



En 1 : à la fois A et B,
En 2 : A mais pas B,
En 3 : B mais pas A,
En 4 : ni A ni B.

Cas plus complexe : participation partielle de caractéristiques.



En I : plus de B que de A.

L'attribution d'une mesure à chaque proposition permet une repré--sentation de la signification d'une affirmation, à partir des éti--quettes, pour un observateur donné. Il faut ajouter que chaque observateur a son propre jeu de vecteurs de base (référentiel).

B. INFORMATION ET FORME

Le concept de représentation est étroitement lié aux mécanismes cognitifs du récepteur et entre dans une définition opérationnelle de l'information : MACKAY appelle information ce qui provoque ou valide une activité représentationnelle.

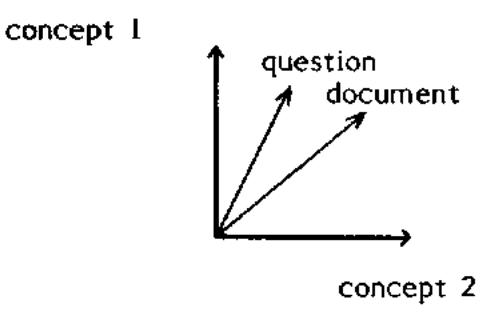
Cette définition associe le concept d'information à celui d'activité adaptative interne se manifestant par la sélection probabiliste d'une représentation. Elle permet d'introduire la notion de pertinence, vue comme une mesure de recouvrement entre une représentation-patron provoquée par une question et une représentation-message provoquée par une information.

Sous cet angle dynamique, informer revient au sens propre à donner une forme (82) et évaluer la pertinence revient à comparer des formes. Ainsi envisagée, la théorie de la forme permet une approche des phénomènes de perception et d'identification compatible avec la théorie de l'information, notamment par l'utilisation du concept

anglais de patron ("pattern").

Exemple :

On peut symboliser l'identification entre une question et un docu--ment de deux mots dans un espace formé par deux concepts du document proches des deux mots :



La question est formée par une proportion de 1 et de 2 en faveur du 1 et le docu--ment est formé par une pro--portion plus équilibrée de 1 et 2.

Remarque :

On pourrait également se placer du point de vue moins subjectif et plus abstrait de A. MOLES (81) et considérer la forme comme la manifestation d'une cohérence interne résultant d'une redondance entre symboles. Dans ce cas, il faut considérer l'ensemble question / réponse comme un message unique redondant et non plus comme un couple message émis / message reçu.

C. INFORMATION - ACTION

Les raisonnements de MACKAY, axés sur les problèmes de percep--tion, ne peuvent être transposés intégralement en sciences de l'information. Ils peuvent cependant déboucher sur une meilleure compréhension du problème information - action, que nous ne ferons ici qu'effleurer, abordé sous l'angle de l'établissement d'une matrice des probabilités de transition à des réponses adaptatives.

1) Effets opérationnels de l'information.

Pour MACKAY, le gain d'information implique une modification des perspectives personnelles et donc une aptitude à réagir dif--féremment. Ce qui est en jeu quand nous recevons de l'infor--mation, c'est notre préparation conditionnelle à conduire le passage du modèle des événements à percevoir vers le modèle de notre propre réaction interne ou externe. L'état total de préparation d'un individu peut être symbolisé par une matrice de probabilités de transition décrivant statistiquement le modèle adaptatif total (interne ou externe) pour chaque configuration possible de stimuli, interne ou externe.

Une telle perspective suppose la maîtrise de problèmes fort com--plexes relevant principalement de la psychologie quantitative ([15]), comme par exemple :

a) l'inventaire des événements à percevoir, y compris les stimuli internes ;
b) l'inventaire des réactions internes et externes possibles ;
c) la confection d'une matrice de fréquences relatives de tran--sition prises dans un ensemble d'organismes identiques.

Il s'agit en fait pour MACKAY de dépasser le problème technique de la transmission et de décrire un processus de communication où intervient la signification de l'information. Son hypothèse est la suivante : en situant l'étude sur l'effet de l'information sur la matrice de probabilités de transition, nous pouvons trouver une place pour tous les concepts relevant de l'information aussi bien pour la T.M.C. que pour la sémantique - au moins en terme de métalangage d'un observateur.

2) Action - résultat.

Les études de YOVITS ([111]) peuvent servir de base à une illustra--tion analogique concrète de l'état total de préparation de MAC--KAY. Les perspectives sont cependant différentes, quoique voi--sines, YOVITS axant ses études sur la décision plus que sur

la perception. L'information, telle que la définit l'auteur, revêt un sens utilitaire très concret puisqu'elle apparaît comme "une donnée utile dans un processus de décision". Cette démarche est conforme aux conceptions des chercheurs opérationnels et économistes américains - comme R. L. ACKOFF (1) et J. MARSHAK (74) - utilisant l'information comme une notion quantifiable en théorie de la décision.

YOVITS considère deux matrices de passage de m "lignes d'action" vers n résultats : l'une est formée des probabilités subjectives w_{ij} de transition de l'action i au résultat j ; l'autre des valeurs subjectives v_{ij} accordées au passage de l'action i au résultat j. Partant de ces matrices, YOVITS procède en trois grandes étapes :

a) Définition d'une valeur attendue de l'action i au temps t :

$$EV_i(t) = \sum_{j=1}^{n} w_{ij}\, v_{ij} \ .$$

b) Définition de la probabilité de sélection de l'action i :

$$P(a_i) = \frac{(EV_i)^c}{\sum_{k=1}^{m} (EV_k)^c} \quad ,$$

c étant un paramètre traduisant la confiance du décideur en sa connaissance de la situation.

c) Définition d'une mesure d'information au temps t exprimant à ce moment l'incertitude du choix d'une ligne d'action appropriée :

$$I(t) = m \sum_{i=1}^{m} P(a_i)^2 - 1$$

Cette mesure représente bien d'une certaine façon une incertitude, dans le sens de "dispersion", puisqu'elle est dérivée de la variance de P(a) :

$$I(t) = \frac{\text{Var } P(a)}{\overline{P(a)}^2}$$

Un point est cependant notable : la variable aléatoire P(a) est elle-même une probabilité et non un événement au sens où

nous avons employé ce terme jusqu'à maintenant.

Partant d'un état initial au temps t_o où la mesure d'information est $I(t_o)$, la réception au temps t_1 de données D fait passer la mesure à $I(t_1)$.

La quantité d'information QI exprime la modification de la mesu-re d'information du décideur à la suite de la réception de D :

$QI(D,t_1) = I(t_1) - I(t_o)$.

YOVITS poursuit son analyse par la quantification de divers para-mètres pragmatiques tels que l'efficacité du décideur, la valeur de l'information et l'efficacité de l'information. La notion de coût de l'information, particulièrement importante dans le contexte de l'action, n'est cependant pas abordée.

3) Complémentarité des démarches.

Les études de YOVITS, qui permettent une modélisation de l'étape possibilité d'action - résultat, peuvent être complétées en amont par une modélisation de l'étape précédente événement - possibilité d'action interne ou externe. On aborde ainsi l'étude des trois problèmes énumérés par WEAVER : technique, sémantique, d'effi-cacité. Cette prise en considération n'est rendue possible qu'au prix d'une analyse comportementale de l'utilisateur avec tous les inconvénients qui peuvent en découler, en particulier :

- complication du modèle qui, cependant, ne résulte que d'une simplification frustrante de la réalité ;
- découpage arbitraire des données ;
- difficulté de l'expérimentation ;
- subjectivité des valeurs attribuées aux matrices et à certains paramètres ;
- insuffisance générale des modèles comportementaux ;

  Remarque : Les modèles comportementaux sont dénoncés par K. POPPER pour qui, contrairement à l'impression première, nous pouvons apprendre davantage sur le comportement de production en étudiant les produits eux-mêmes que nous ne pouvons apprendre sur les produits en étudiant le comportement de production (80). Une telle insuffisance apparaît d'ailleurs

chez le chercheur en économie R. N. LANGLOIS (64) qui associe la signification au processus stimulus - [boîte noire*] - réponse, avec stimulus = information et réponse = signification. Si cette association peut se justifier dans la science de l'action qu'est l'économie, elle n'est transposable que partiellement en sciences de l'information. Nous référant de nouveau à POPPER et à l'exégèse de B. C. BROOKES (15), nous pouvons placer une grande partie du domaine d'étude des sciences de l'information dans le "Troisième Monde" de la connaissance objective.

Même si nous hésitons à considérer, comme le fait POPPER, que ce troisième monde a une existence autonome, nous sommes amenés à admettre en son sein certaines lois soit ne dépendant plus entièrement de facteurs subjectifs, soit en dépendant d'une façon beaucoup plus complexe que ne l'indique le modèle causal stimulus - [boîte noire] - réponse.

On peut rattacher à certaines de ces limitations la grande prudence de MACKAY : ses hypothèses ne sont accompagnées d'aucun modèle expérimental précis d'exploitation d'une grille information - pos--sibilité d'action, ce qui d'ailleurs ne remet pas en cause la qualité de son analyse.

Si YOVITS propose un mode d'exploitation d'une grille action - résultat bien défini, il faut replacer son étude dans le contexte qui lui a donné naissance : la théorie de la décision économique appliquée dans le cadre d'une équipe homogène de dirigeants. L'exemple qu'il choisit pour illustrer sa démarche s'inscrit d'ailleurs dans la ligne de l'analyse économique (décision pour un cultivateur d'engager sa saison sur telle ou telle récolte).

* *Une boîte noire désigne un système dont l'organisation est décrite au moyen d'une fonction de transfert et non par observation directe de sa structure interne* (107).

III. APPROCHE PAR REFERENTIEL STRUCTURE

Cette seconde approche s'attache d'une façon différente aux propriétés du texte, délaissant en première analyse le comportement de l'utilisa--teur. Elle tient compte néanmoins de l'intervention d'utilisateurs inter--médiaires se manifestant indirectement par l'élaboration de langages documentaires structurés.

Partons d'un exemple concret :

Afin d'illustrer son affirmation que le concept d'information sémantique n'a aucun rapport avec la T.M.C., BAR HILLEL ([7]) développe un exem--ple martial : "Il est tout à fait sensé d'affirmer, par exemple, qu'un rapport

"L'ennemi a attaqué à l'aube"

porte moins d'information que

"L'ennemi a attaqué à l'échelle d'un bataillon à 5 h 30",

et il est également parfaitement clair que la seconde affirmation est plus précise que la première. Il est pour cela judicieux de demander si on ne peut affiner l'évaluation comparative en évaluation quantitative et dire **combien d'information en plus** est portée par le second rapport". BAR HILLEL estime qu'on ne peut aller au-delà de l'affirmation quali--tative que la seconde proposition est plus précise que la première. C'est pourquoi il semble préférable d'examiner la question non plus sous l'angle du langage courant, mais sous celui d'un langage documen--taire.

A. INFORMATION DES MICRO-MESSAGES

Ramenons les deux rapports à leur plus simple expression :

$X = \{A = \text{ennemi}, B = \text{attaquer}, C = \text{aube}\}$

$Y = \{A = \text{ennemi}, B = \text{attaquer}, D = \text{bataillon}, E = \text{heure } 5{:}30\}$.

Afin d'évaluer ce qui détermine la quantité d'information d'un micro-message, reportons-nous aux axiomes de base de la théorie de l'infor--mation généralisée (61) (62) :

1) L'information associée à un message (proposition) est un nombre non négatif.

2) L'information croît quand l'événement observable décroît, c'est-à-dire au fur et à mesure que l'on localise mieux les événements du message dans l'espace des événements possibles (espace des phases) :
$M \subset N \Rightarrow J(M) \geqq J(N)$.

3) L'information est de forme additive : si M et N sont indépendants, $J(M \cap N) = J(M) + J(N)$.

Le second axiome, essentiel dans le cas présent, pose le problème en deux termes :
- en terme de nombre d'événements dans un message,
- en terme d'inclusion d'événements les uns par rapport aux autres.
Ces deux termes font apparaître en fait deux critères voisins de comparaison des messages.

B. COMPARAISON DES MICRO-MESSAGES

Comparons X et Y élément par élément :

1) Les éléments A et B ne posent aucun problème : ils se retrouvent dans les deux messages. Ils apportent autant d'information à X qu'à Y.

2) L'élément D n'a pas d'équivalent en X. Il s'agit d'un événement supplémentaire en faveur du message Y apportant une meilleure localisation de Y par rapport à X dans l'espace des événements possibles. Si on caractérise un rapport par les séquences successives
"l'ENNEMI",
"l'ENNEMI a ATTAQUE",
"l'ENNEMI a ATTAQUE à l'échelle d'un BATAILLON",
on accumule des éléments permettant de mieux préciser le message et d'accroître l'information qu'il apporte.
Cependant, ce qui apparaît dans le langage courant comme une accumulation de détails correspond au choix descendant, à partir d'un ensemble plus vaste (les hommes ou les soldats, par exemple), de trois sous-ensembles (66) :

A = sous-ensemble des ENNEMIS,
AB = sous-ensemble des ENNEMIS ayant ATTAQUE,
ABD = sous-ensemble des ENNEMIS ayant ATTAQUE à l'échelle d'un BATAILLON.

Ces sous-ensembles sont inclus de la façon suivante :
$ABD \subset AB \subset A$.
Ces inclusions successives se traduisent, selon le second axiome de KAMPE DE FERIET, par la propriété suivante qui confirme l'expérience intuitive :
$J(ABD) \geqq J(AB) \geqq J(A)$.
A cette étape, on rend compte de l'intervention de D dans le message Y par la comparaison de B et BD, avec
$J(BD) \geqq J(B)$.

3) Les éléments C et E décrivent une même réalité - le temps - de façon respectivement imprécise et précise. La correspondance entre ces éléments peut, de la même façon, être décrite en terme d'inclusion.
L'exemple du temps est caractéristique. La précision attachée à l'indication du temps découle typiquement d'une hiérarchie selon,

par exemple, l'échelonnement simplifié suivant que l'on pourrait rencontrer dans un langage documentaire :

. Siècle
. . Année
. . . Mois
. . . . Jour
. . . . . Heure
. . . . . Minute
. . . . . Seconde

Dans une telle hiérarchie, l'élément C trouve sa place entre Jour et Heure, et l'élément E à Minute. Toute ambiguïté étant par ailleurs levée quant à l'appartenance au Jour, au Mois, à l'Année, au Siècle, on peut comme précédemment comparer C et E selon le critère
$E \subset C \Rightarrow J(E) \geqq J(C)$.

C. SPECIFICITE ET HIERARCHIE

Affirmer qu'une proposition est plus précise qu'une autre, donc con--tient plus d'information, revient à raisonner en terme de partition. Le langage courant se prête mal au découpage structuré d'une col--lection de concepts de référence. Par contre certains langages documentaires (classifications, thésaurus) permettent d'accorder la précision d'un concept à sa localisation hiérarchique.
Afin d'offrir une base aussi ferme que possible à la comparaison de deux ou plusieurs messages, il convient de préciser les caractéris--tiques d'utilisation d'un langage hiérarchisé.

1) Spécificité.

La localisation d'un terme est associée à la spécificité du concept qu'il représente. Un terme peut décrire plus ou moins finement

une réalité et donc présenter plus ou moins de poids informatif. La spécificité est une propriété intellectuelle. Il convient de dis-tinguer la spécificité de la rareté, propriété statistique, même si les deux propriétés vont souvent de pair. La spécificité provient de partitions successives, alors que la rareté est observée par simple comptage de fréquence de terme, sans examen du contexte dans lequel ce terme est rencontré. Il ne peut être question de confondre les deux propriétés car cela reviendrait à établir un rapport sans fondement entre inclusion de sous-ensembles et dif-férences des fréquences associées à ces sous-ensembles.

2) Profondeur.

Il ne faut cependant pas négliger les indications apportées par les mesures de fréquences, car elles peuvent permettre de déceler une inadaptation du langage au domaine qu'il décrit.

On constate en particulier assez souvent l'apparition trop fréquente de termes au sein de descriptions bibliographiques. Dans de tels cas, on peut se demander si ce défaut n'est pas dû à un manque de profondeur de la hiérarchie : les concepts traduits par l'indexeur n'ont pu être associés à des termes placés à un niveau assez bas dans le langage hiérarchisé *.

C'est pourquoi deux messages ne sont vraiment comparables que

a) si tous les concepts peuvent être traduits dans le même langage documentaire hiérarchisé,
b) si le langage hiérarchisé comprend suffisamment de niveaux pour éviter une surpopulation à certains niveaux de termes abusivement considérés comme égaux,
c) si tous les termes sont placés au plus bas dans le référentiel.

Remarque:

On a constaté dans l'exemple précédent que l'accumulation de détails revient à une succession d'inclusions. Il faut cependant admettre la difficulté de conception d'un langage hiérarchisé comprenant un large éventail d'inclusions successives, parfois difficiles à prévoir.

* *On peut citer comme exemple la pauvreté du thésaurus de la NASA dans le domaine des aéronefs plus légers que l'air.*

D. COMPLEMENTARITE AVEC L'APPROCHE PROBABILISTE

Nous constatons que cette approche par structuration du référentiel est moins directement liée à la notion de pertinence que l'approche probabiliste. En effet, si l'inclusion permet de comparer la précision de plusieurs messages voisins, elle ne permet pas d'établir directe--ment et rapidement l'affinité sémantique de ces messages. Les deux démarches sont cependant complémentaires si on désire affiner la relation de pertinence.

Dans un premier temps, il s'agit de mesurer une relation d'affinité entre messages par des mesures informatives utilisant des probabilités (fréquences).

Dans un deuxième temps, on peut envisager de comparer la précision (spécificité) de deux messages suffisamment voisins sémantiquement pour que la comparaison soit utile. L'approche par inclusion permet cette seconde opération si on dispose d'un langage hiérarchisé suffi--samment riche. Ce dernier point est capital : le langage documen--taire servant de référentiel doit être adapté à l'état des connais--sances de façon à pouvoir accepter et localiser tous les termes rencontrés dans la littérature.

Une telle contrainte conduit à la mise au point d'un lexique de compa--tibilité très riche permettant le passage du vocabulaire libre au vocabulaire de référence.

Le principal intérêt de cette seconde approche est d'asseoir la per--tinence non seulement sur une affinité statistique mais aussi sur une identité aussi grande que possible des niveaux de spécificité. Il est parfois utile, sur le plan de la pratique documentaire, de ré--pondre à une question par une liste de références se situant au même niveau de spécificité que la question (général, ponctuel).

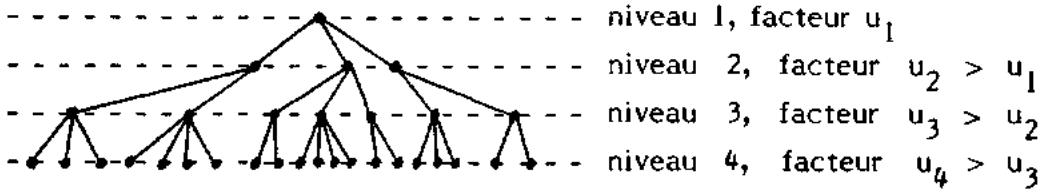
Il peut y avoir là une direction pour de futures recherches basées sur la pondération de l'affinité statistique en fonction de la proxi--mité des niveaux de spécificité de la question et du document. La notion de facteur d'utilité est un point de départ possible pour de telles recherches. Il s'agirait d'affecter à chaque terme d'un message un facteur croissant avec le niveau de spécificité de ce terme au sein du référentiel structuré. Ce facteur d'utilité, proposé par exemple par GUIAŞU et THEODORESCU (51), permet de pondérer

la quantité d'information d'un message en fonction d'une particularité assignée à chaque symbole du message (sa spécificité par exemple). Il devrait de même permettre de pondérer l'information transmise ou mutuelle entre messages en fonction de la proximité des niveaux de spécificité.

Dans le cas de l'information de SHANNON, la fonction entropique devient* :

$$H(X) = -\sum_i u(i)\, p(i) \log_2 p(i) ,$$

u(i) dépendant du niveau hiérarchique :



Cette double approche, adaptée aux équations utilisées dans la simple approche probabiliste, devrait favoriser les affinités sémantiques entre messages de niveaux de précision voisins et donc affiner l'iden-tification d'un message avec les clés d'un langage documentaire.

* *On trouvera dans le rapport de HAYES (53) une bibliographie commentée sur la pondération de l'entropie.*

CONCLUSION

A. CHONEZ a insisté récemment sur la lenteur de l'évolution des tech--niques documentaires liées à la ressaisie : "les méthodes d'indexation et de recherche documentaire couramment utilisées n'ont fait aucun progrès notable depuis dix ans et elles restent très insuffisantes ; seule la technique d'accès a progressé. Il faudra se tourner vers d'autres voies pour avoir quelque chance de maîtriser intelligemment et efficacement une masse documentaire croissante..." (21).

La théorie de l'information peut-elle contribuer à faire avancer les cho--ses ? C'est possible : les résultats du présent travail montrent que cette dis--cipline semble avoir plus de rapport avec les sciences de l'information qu'il n'y paraît à première vue, notamment dans l'analyse des bases de données. Il convient cependant de poser le problème sur des bases saines. En parti--culier, il est utile :

- de dépasser la transposition immédiate du domaine des télécommunications à celui des sciences de l'information,
- de distinguer la théorie de SHANNON des développements auxquels elle a donné naissance,
- de compléter des modèles purement probabilistes.

Concernant le premier point, on peut remarquer qu'il existe deux façons de considérer la théorie de SHANNON :

Une qui considère la transmission de l'information. Elle est bien adaptée aux problèmes de télécommunications et utilise principalement la notion de codage.

Une qui considère la structure statistique des "messages". Elle a été par--ticulièrement employée par les psychologues et les biologistes et utilise principalement la notion d'écart de deux variables aléatoires par rapport à l'indépendance statistique. Elle a été reprise et étendue récemment par les analystes de systèmes aux fins de décomposition de systèmes complexes en sous-systèmes faiblement couplés.

Cette dernière façon de considérer la théorie de SHANNON n'a plus grand chose à voir avec la transmission d'une quelconque information sauf si on admet que la théorie de l'information est tout à la fois une théorie de l'observation et une théorie de la complexité. Dans ce cas, l'observation revient à une transmission d'information entre l'objet étudié et un obser--vateur.

Concernant le deuxième point, il faut admettre que la T.M.C. est un point de départ quasi-obligatoire mais n'est pas toute la théorie de l'in--formation. Comme le montre l'illustration expérimentale décrite en annexe, il est à cet égard utile de se tourner vers les développements récents de la théorie de l'information, particulièrement ceux issus des travaux de KAMPE DE FERIET et son équipe sur la théorie de l'information généra--lisée. Ces travaux menés principalement en France et en Italie sont moins connus que ceux de l'école anglo-saxonne et ont donc plus de mal à filtrer dans les milieux de la recherche en sciences de l'information.
La théorie de l'information généralisée semble proposer, par les démarches intellectuelles mises en jeu dès le départ, une problématique adaptée aux besoins des scientifiques de l'information : isoler un message, utiliser des probabilités a priori et non a posteriori, faire intervenir des "observateurs", abandonner au besoin l'utilisation des probabilités...

Enfin, il apparaît que l'approche probabiliste, bien qu'indispensable, gagnera à être complétée par une analyse linguistique permettant d'affiner l'identification de messages et cela d'autant plus que les messages en pré--sence seront courts et donc difficiles à situer dans un contexte sûr.
L'examen des rapports information - signification est particulièrement révélateur de la complexité du problème.
D'une part, si les mécanismes d'information - action font intervenir les probabilités, elles le font dans le cadre de la théorie de la décision, bien différent de celui de la théorie de l'information classique ;
d'autre part, la notion de spécificité d'un événement fait intervenir des relations d'inclusion entre sous-ensembles, inaccessibles par les seules probabilités.
Là encore, les probabilités forment une sorte de noyau méthodologique indispensable qu'il convient d'adapter ou de compléter.

PARTIE EXPERIMENTALE

I. BUT DE L'EXPERIMENTATION

Le but de la présente expérimentation est modeste. Il s'agit d'apporter une **illustration** à l'intérêt que peut présenter la théorie de l'information et non pas de construire un système. Le parti choisi a été d'écarter les fonctions se révélant inadaptées à l'échelle d'un échantillon restreint et, à plus forte raison, inadaptées au contexte des bases de données. Les fonctions manifestement inadéquates étant écartées, reste à prouver l'efficacité dans les conditions les plus rudes des fonctions restantes supposées dignes d'intérêt. Une telle preuve requiert une expérimentation à grande échelle exigeant de très importants moyens financiers et hu-mains.

II. ECHELLE DE L'EXPERIMENTATION

Dans un article du Journal of Documentation (103), SPARCK JONES et VAN RIJSBERGEN recommandent de pratiquer les expériences de traite-ment et de ressaisie de l'information sur des fichiers d'au moins 1000 notices. Cependant, on constate que nombre d'expériences - et non des moindres - sont faites sur des fichiers beaucoup plus restreints, pour des raisons évidentes de coût et de personnel.

En fait, le nombre brut de notices employées n'entre pas seul en ligne de compte. Si on veut évaluer la qualité d'un échantillon et la pertinence des résultats obtenus, il faut tenir compte d'un certain nombre de para-mètres. Du point de vue quantitatif seul, le nombre de notices n'est par lui-même pas suffisamment significatif : il faut par exemple savoir à combien de mots-clés correspond la collection de notices. Pour citer un exemple célèbre*, MARON retient 260 documents, mais n'utilise que 90 mots-clés et restreint sa grille de probabilités à cet échantillon fort réduit (72). Car, en plus, il faut savoir si la collection se suffit à elle-

* *Cf. page 19.*

même et donne seule naissance aux données traitées dans les calculs, ou bien si les propriétés de l'échantillon sont quantifiées sur une base plus large. Dans le premier cas, on aboutit à des tableaux mot-clé / mot-clé ou mot-clé / code de classification remplis de 0, donc peu signi--ficatifs, même pour un échantillon de quelques milliers de notices. Par contre, si dans le second cas le fichier-test est la partie observée d'un énorme fichier dans lequel on puise les données traitées dans les calculs, on travaille en grandeur réelle avec des tableaux bien remplis et donc une fiabilité certaine par rapport au cas précédent : l'intersection de deux mots-clés, par exemple, correspond à une réalité stable et non pas au hasard d'un choix de documents.
C'est dans ces conditions favorables que nous avons travaillé.

III. METHODE DE L'EXPERIMENTATION

Le principe de l'expérimentation est le suivant :
Constituons une base de données fictive limitée à un certain nombre de notices. Ces notices décrivent des publications dont le sujet est bien défini et connu. Effectuons au sein de cette base de données un certain nombre de recherches documentaires. Cela revient à poser des questions dont on doit pouvoir évaluer la réponse : le sujet de chaque notice étant bien défini et connu, et une question étant posée, un observateur peut affirmer que telle notice est ou n'est pas pertinente.
Si une mesure de couplage tirée de la théorie de l'information permet de **trier** les notices de la base de données selon un **ordre** décroissant de pertinence à la question
ou d'**affecter** préférentiellement ces mêmes notices à un code de clas--sification,
on pourra s'interroger sur la qualité de cette mesure.
Plus l'ordre de tri obtenu ou bien l'affectation à un code de classifi--cation retenue correspondront à l'évaluation la plus objective possible d'un observateur, plus la mesure sera jugée performante.

Les possibilités d'expérimentation - c'est-à-dire très concrètement le temps d'interrogation - dont nous disposions nous ont permis de sélec--tionner un jeu de 88 références bibliographiques provenant de la base de données PASCAL. Les titres de 4 de ces références étant traduits de l'anglais en français, l'échantillon se compose en tout de 92 notices correspondant à 88 publications.

Ces notices sont extraites des collections du Bulletin signalétique - version "papier" de PASCAL - conservées à la Bibliothèque universitaire de Reims, section des Sciences et Techniques. Cette bibliothèque ne possède qu'une collection fragmentaire du Bulletin signalétique, ce qui explique la disparité dans le temps du jeu de notices : la plus ancienne est tirée d'un fascicule de 1974, les plus récentes sont tirées de fasci--cules de 1982.

Les références se répartissent de la façon suivante :

25 dans la Section 101 du Bulletin signalétique (Sciences de l'informa--tion, Documentation) au sous-paragraphe "Accès au stock documentaire et mode d'exploitation" : choix systématique de la 3ème notice du sous-paragraphe ;

18 dans la Section 110 (Analyse numérique, Informatique, Automatique, Recherche opérationnelle, Gestion, Economie) au paragraphe "Théorie des graphes" : choix systématique de la 3ème notice du paragraphe et 20 aux paragraphes "Robotique" en 110.D.06.C et 110.E.02.A : choix systématique de la 1ère notice de chacun de ces paragraphes ;

25 dans 25 autres sections afin d'élargir l'éventail des sujets abordés et des termes employés : choix au hasard.

L'échantillon résultant de ce choix est détaillé dans les pages suivantes, dans l'ordre des sections du Bulletin signalétique, avec sous-ordre chronologique :

Liste des notices (numéro d'ordre B.S. et titre) :

74-101-1762
RALF : a new sofware package for the whole complex of punched card oriented documentation.

74-101-2293
Intérêt de la visualisation conversationnelle en documentation automa-tique.

74-101-2709
Accélération de la recherche dans des systèmes de recherche documen-taire à descripteurs avec une organisation directe du fonds de recher-che.

74-101-3340
La CDU et les équipements de recherche.

76-101-319
PANDORA control system and retrieval language.

76-101-682
SERLINE. On-line serials bibliographic and locator retrieval system.

76-101-1963
Dispositif pour le tri et la sélection de cartes à encoches latérales supports d'information.

76-101-2270
Système de sécurité pour mémoire informatique.

77-101-620
SCORPIO, a subject content oriented retriever for processing infor-mation on-line.

77-101-1124
Commission information et documentation. Journée annuelle - Lyon, 3 juin 1976. Introduction de la journée.

77-101-1903
Un modèle d'accès standard dans les systèmes de gestion de bases de données.

77-101-3372
Le système d'information juridique JURIS.

77-101-4002
La base de données de la bibliothèque du VTI.

80-101-2027
Utilisation des bases de données Chemical Abstracts Service (CAS). Structure du fichier.

80-101-2702
Applications of viewdata.

80-101-3496
Sleeping beauty : MERLIN, a state of the art report.

81-101-874
Generalization of the graph center concept, and derived topological centric indexes.

81-101-1512
A computer data base system for indexing research papers.

81-101-3056
Organisation et utilisation d'un canal de transmission de données pour un système de recherche de l'information en conversationnel.

81-101-3409
Searching in academia. Nearly 50 libraries tell what they are doing.

81-101-3964
Online in industrial and research libraries.

81-101-4830
Systematic information retrieval and directional data analysis of oligopeptide units in protein data bank.

82-101-536
Subject specialists searching Chemical Abstracts on SDC.

82-101-1951
Planning online search service in a state university.

82-101-2294
Fast, parallel relaxation screening for chemical patent data-base search.

75-110-10889
Décomposition des polytopes.

75-110-12656
On some properties of n-tournaments : a note.

77-110-10448
Détermination du nombre structural d'un graphe comportant des blocs par la méthode des sections.

80-110-2481
The enumeration of bipartite graphs.

80-110-4661
A characterization of Robert's inequality for boxicity.

80-110-8431
The book thickness of a graph.

81-110-15931
Randomly k-axial graphs.

81-110-17957
Recent results in partition (Ramsey) theory for finite lattices.

82-110-224
Polyhedra related to a lattice.

82-110-1325
Ramsey numbers involving graphs with long suspended paths.

82-110-2411
A depth first search algorithm to generate the family of maximal independent sets of a graph lexicographically.

82-110-3701
A construction of geodetic blocks.

82-110-4832
Cycles in strong oriented graphs.

82-110-5947
On weak persistency of Petri nets.

82-110-7156
On coverings of random graphs.

82-110-8996
Finding a minimum equivalent graph of a digraph.

82-110-11126
Thermodynamic bond graphs and the problem of thermal inertance.

82-110-13533
Exposants de longueur pour des familles de graphes polytopes.

80-110-9201
Construction analytique des systèmes de commande des robots indus--triels.

81-110-16641
SIGLA. Olivetti robot programming language.

81-110-16683
A knowledge-based interactive robot-vision system.

81-110-19145
An adaptive trajectory control of manipulators.

81-110-19215
Use of optical reflectance sensors in robotics applications.

82-110-500
Control of force distribution in robotic mechanisms containing closed kinematic chains.

82-110-500 (trad.)
Commande de la distribution des forces dans les mécanismes de robo-tique comportant des chaînes cinématiques fermées.

82-110-1637
A perspective on robotics research and this issue.

82-110-2811
Sense-controlled flexible robot behavior.

82-110-4168
Some critical areas in robotics research.

82-110-4185
Processus d'apprentissage programmé comme aide à la conception.

82-110-5212
Commande du mouvement des robots manipulateurs sur la base des algorithmes cinématiques du second ordre.

82-110-6383
Robots avec des capteurs de force et de moment.

82-110-7852
Artificial intelligence, automatic control and development.

82-110-7876
Laser electro-optic system for rapid three-dimensional (3-D) topogra-phic mapping of surfaces.

82-110-9712
On the equivalence of Lagrangian and Newton-Euler dynamics for manipulators.

82-110-9797
Un système pour l'expression et la résolution de problèmes orienté vers un contrôle de robots.

82-110-11932
Optimisation dynamique de ressources et reprogrammation dynamique en robotique.

82-110-11975
A microcomputer based artificial intelligence laboratory.

82-110-14152
The inverse kinematic problem for anthropomorphic manipulator arms.

82-110-14200
Robotique et intelligence artificielle.

82-120-7178
Test for a richness-dependent component in the systemic redshifts of galaxy clusters.

82-130-10892
Effect of suspended particulates on the measurement of gas velocity using a Pitot-static tube.

82-140-1736
Estimation de l'efficacité des régleurs de tension en charge à thyristor.

82-145-4820
Oscillateurs microondes stables intégrés à transistors et résonateurs diélectriques.

82-161-1305
Vacancy trapping in plastically deformed metals studied by hyperfine interactions.

82-161-1305 (trad.)
Piégeage des lacunes dans les métaux déformés plastiquement. Etudes au moyen des interactions hyperfines.

82-173-10041
Isomérisation des ions alcools éthyléniques phényles en phase gazeuse.

82-221-616
Base metal deposits in sedimentary rocks : some approaches.

82-221-616 (trad.)
Gîtes de métaux de base dans les roches sédimentaires : quelques approches.

82-224-2267
Contribution à l'étude du Trias carbonaté des Pyrénées occidentales et centrales.

82-310-501
Essai comparatif d'un nouveau capteur épidural pour la mesure et la surveillance de la pression intracrânienne.

82-320-512
Dosage des protéines par la méthode de Bradford au bleu de Coomassie G250. Problème des interférences.

82-330-6967
Action de la thiobiline dans le traitement des troubles fonctionnels digestifs.

82-340-8664
Influence des sources d'azote sur le métabolisme du glycogène chez Saccharomyces carlsbergensis.

82-361-8020
Activation of progesterone receptor by ATP.

82-362-1974
Etude ultrastructurale, immunocytochimique et radioimmunologique d'un glucagonome pancréatique humain.

82-363-3815
Fréquence, homologie et clonage des plasmides cryptiques du Bacillus thuringiensis.

82-365-10949
Le polymorphisme chez les Crénilabres méditerranéens du genre Sym--phodus.

82-370-4608
La photosynthèse du tournesol : recherches sur le mode de fixation du CO_2.

82-390-37
Cognitive factors in subjective stabilization of the visual world.

82-730-9773
La gazéification souterraine profonde du charbon.

82-740-5805
Amélioration de la résistance à la corrosion par grenaillage de pré--contrainte.

82-745-1487
The automatic control of electron beam welding equipment.

82-745-1487 (trad.)
Commande automatique du matériel de soudage par faisceau d'électrons.

82-780-589
Effet d'une radiolyse à basse température sur la résistance de fibres de polyéthylène.

82-880-375
Sequential control of continuous distillation.

82-885-3996
Incinération des résidus urbains : une source de dioxines ?

82-892-4223
Comportement à l'humidité des éléments de construction : application aux toitures.

Afin de composer des tableaux de contingence significatifs, les termes sont systématiquement tronqués selon le tronc le plus long trouvé dans les lexiques PASCAL (1982) pour les mots français et le thésaurus de la NASA (1976) pour les mots anglais. La troncature est limitée à une lettre pour les mots de moins de 5 lettres afin d'éviter les dépassements de capacité de disque. Les noms de systèmes, sigles et verbes ne sont pas tronqués. On aboutit ainsi à une collection de 401 mots après élimination des mots-outils.

Les fréquences d'occurrence et de co-occurrence sont obtenues par interrogation en ligne de la base de données PASCAL dans le système QUEST de l'Agence spatiale européenne. L'interrogation de QUEST permet l'accès au fichier unifié n° 14 rassemblant les 4500000 notices PASCAL recueillies sur une période de 10 ans (1973-1982).

Les résultats des sélections (nombres d'occurrence) et des combi--naisons "ET" (nombres de co-occurrence) sont rassemblés dans des tableaux de co-occurrence mot-clé question / mot-clé document et code de classification / mot-clé document.

Exemple : mots-clés question / mots-clés documents :

		BASE?(W)DONNEE?	BIBLIOGRAPHI?	CHIMI?	GRAPHE?	MANIPULATEUR?	ROBOT?
		59	16005	387548	7808	719	2237
EQUIPEMENT?	29971	0	99	796	17	17	71
EQUIPMENT?	18426	0	100	574	4	13	142
EQUIVALEN?	5997	0	5	589	88	1	4
ESSAI?	72993	2	277	7408	9	3	17
ESTIMAT?	29571	1	61	1900	88	1	8
ETHYLEN?	52842	0	260	9761	25	1	1
ETUDE?	527510	4	1982	60607	350	34	84

.../...

Exemple : codes de classification / mots-clés document :

		101	110	120	130	140	145	161	221	224	310	320	330
		47464	163038	166202	69409	47640	113705	82521	48640	37886	46990	131543	238746
EQUIPEMENT?	29971	807	3278	97	430	1215	2921	759	23	8	912	13	85
EQUIPMENT?	18426	1165	1206	555	872	999	2430	149	21	3	221	14	27
EQUIVALEN?	5992	30	1020	441	314	770	802	37	12	38	170	121	70
ESSAI?	72993	97	911	833	3482	3592	2497	191	207	222	360	86	7370

.../...

.../...

Les fonctions de couplage sont calculées dans un deuxième temps, notice par notice, à partir des tableaux de contingence particuliers à une notice pour chaque question. Les valeurs nulles entrant dans les fonctions logarithmiques sont remplacées par 10^{-6}.

Exemple : question "BASE?(W)DONNEES - BIBLIOGRAPHI? - CHIMI?" appliquée à la notice 81-101-3056 (avec mention des nombres d'occurrence $f(q_i)$, $f(m_j)$ et de co-occurrence $f(i,j)$ définis page 70) * :

	BASE?(W) DONNEES 59	BIBLIOGRAPHI? 16005	CHIMI? 387548
ORGANISAT? 18088	2	88	304
UTILISAT? 83700	8	799	9062
CANAL? 21008	0	34	723
TRANSMISSION? 70168	1	81	3381
DONNEE? 50010	59	501	2728
SYSTEM? 342948	14	743	29663
RECHERCHE? 52936	6	1158	3993
INFORMATION? 39732	22	831	888
CONVERSATIONNEL? 3107	2	116	79

Exemple : co-occurrence de 25 codes de classification avec les mots-clés de la même notice : cf. page suivante.

* *Nous ne rappellerons pas dans le présent travail les particularités du logiciel QUEST. On pourra se reporter pour plus de détails aux pages VI-24 à VI-48 du "Manuel d'utilisation PASCAL" publié par le C.D.S.T..*

		101	110	120	130	140	145	161	221	224	310	320	330
		37464	163938	166202	195409	47660	113725	82521	40640	37886	46940	131543	238356
ORGANISAT?	18088	1833	2739	96	44	317	124	13	65	51	502	487	89
UTILISAT?	83709	5026	1369	2150	1824	1156	1205	149	1240	167	1711	417	2379
CANAL?	21008	9	842	267	1867	187	1564	682	20	31	132	120	394
TRANSMISSION?	70168	172	3731	698	2564	1164	8320	1022	[illegible]	5	787	506	1751
DONNEE?	50010	3158	9935	9315	1652	314	1094	469	766	1419	927	181	541
SYSTEM?	342948	5537	44298	5788	16317	4196	8505	4656	599	707	4736	5831	25539
RECHERCHE?	52936	5166	4417	2454	818	668	693	180	1030	611	545	544	2516
INFORMATION?	39732	13750	11519	1086	1594	955	2962	81	134	287	1717	103	154
CONVERSATIONNEL?	3107	1307	1585	62	30	31	62	2	1	7	47	1	0

	340	361	362	363	365	370	390	730	740	745	780	800	885
	284087	182269	46514	78313	367397	105993	140256	169037	154281	343306	152691	96325	76694
ORGANISAT?	882	309	30	1004	959	214	2632	353	704	331	85	180	305
UTILISAT?	1790	696	150	380	638	499	645	15304	6496	1696	13915	11104	2140
CANAL?	330	1475	70	48	1215	94	10	2704	890	511	630	207	653
TRANSMISSION?	13725	5357	951	2008	9654	491	120	600	721	66	360	263	442
DONNEE?	526	428	136	208	701	589	547	1652	1855	173	255	394	665
SYSTEM?	20902	16985	3428	9034	33185	14608	5232	9049	6605	693	3705	6126	2177
RECHERCHE?	3706	1189	280	734	1360	1095	1773	3930	1787	1030	1. 2	1268	1435
INFORMATION?	170	98	10	220	545	77	3361	474	181	72	109	148	207
CONVERSATIONNEL?	0	0	0	1	0	1	5	12	9	4		5	5

IV. RESULTATS DE L'EXPERIMENTATION

Afin de ne pas alourdir l'exposé, les résultats obtenus par les diffé--rentes fonctions de couplage ne sont pas décrits en détail. On ne s'attache en fait qu'à la fonction donnant les meilleurs résultats.

A. COUPLAGE MOTS-CLES QUESTION / MOTS-CLES DOCUMENT

Pour un lot de N notices pertinentes à la question, les performances des fonctions de couplage sont arbitrairement mesurées par un facteur P = r p, avec

r = nombre de notices pertinentes présentes jusqu'au N^e rang / N

p = nombre de notices pertinentes présentes jusqu'au n^e rang / N

n étant le rang de la première notice non pertinente de rang inférieur ou égal à N.

Un couplage idéal place les N notices pertinentes du 1^{er} au N^e rang. Dans ce cas, r = p = P = 1.

1) Question "base?(w)données".

En admettant grossièrement que les 25 notices de la section 101 sont pertinentes, les résultats obtenus par les fonctions définies pages 85 à 90 sont les suivants :

M_2	P = 0,016	M'_4	P = 0,163
M_3	0,016	I	0,179
M_1	0,018	C	0,202
M_4	0,077	R	0,213
M'_2	0,136	K	0,224
M'_1	0,144	J	0,320
M'_3	0,144	L	0,504

La fonction L est la plus performante. L'ordre de succession est détaillé ci-dessous, les notices non pertinentes apparaissant en caractères soulignés :

L =

1118,8 Fast, parallel relaxation screening for chemical patent data-base search.

373,3 Systematic information retrieval and directional data analysis of oligopeptide units in protein data bank.

242,7 A computer data base system for indexing research papers.

222,6 Planning online search service in a state university.

200,8 Organisation et utilisation d'un canal de transmission de données pour un système de recherche de l'information en conversationnel.

192,6 Online in industrial and research libraries.

190,2 Un modèle d'accès standard dans les systèmes de gestion de bases de données.

168,6 La base de données de la bibliothèque du VTI.

166,4 Utilisation des bases de données Chemical Abstracts Service (CAS). Structure du fichier.

118,3 Accélération de la recherche dans des systèmes de recherche documentaire à descripteurs avec une organisation directe du fonds de recherche.

102,8 Intérêt de la visualisation conversationnelle en documentation automatique.

91,0 SERLINE. On-line serials bibliographic and locator retrieval system.

89,5 Commission information et documentation. Journée annuelle - Lyon, 3 juin 1976. Introduction de la journée.

79,0 SCORPIO, a subject content oriented retriever for processing information on-line.

70,7 RALF : a new software package for the whole complex of punched card oriented documentation.

62,1 Commande du mouvement des robots manipulateurs sur la base des algorithmes cinématiques du second ordre.

57,9 PANDORA control system and retrieval language.

48,5 Dispositif pour le tri et la sélection de cartes à encoches latérales supports d'information.

43,1 Searching in academia. Nearly 50 libraries tell what they are doing.

41,3 Le système d'information juridique JURIS.

35,5 Détermination du nombre structural d'un graphe comportant des blocs par la méthode des sections.

33,9 Optimisation dynamique de ressources et reprogrammation dynamique en robotique.

32,3 Subject specialists searching Chemical Abstracts on SDC.

26,3 A depth first search algorithm to generate the family of maximal independent sets of a graph lexicographically.

19,2 Generalization of the graph center concept, and derived topological centric indexes.

- -

16,3 Processus d'apprentissage programmé comme aide à la conception.

6,1 Système de sécurité pour mémoire informatique.

5,6 La CDU et les équipements de recherche.

.....

0,7 Applications of viewdata.

.....

-4,1 Sleeping beauty : MERLIN, a state of the art report.

.....

-9 Effect of suspended particulates on the measurement of gas velocity using a Pitot-static tube.

On pourra remarquer que certaines notices sont relativement défavorisées, comme par exemple "Subject specialists searching Chemical Abstracts on SDC". L'examen du tableau de contingence de la notice permet d'expliquer cette anomalie :

	BASE?(W)DONNEES
	59
SUBJECT? 8629	0
SPECIAL? 21548	0
SEARCH? 4085	2
CHEMICAL 78187	1
ABSTRACTS 2219	0
SDC 25	0

a) Le terme CHEMICAL n'est associé qu'une fois à BASE?(W) DONNEES, alors que Chemical Abstracts est la base de données bibliographiques la plus importante au monde.

b) Le terme ABSTRACTS n'est tout simplement jamais associé à BASE?(W)DONNEES.

c) Même remarque pour le sigle SDC alors que cette corpora--tion commercialise des bases de données.

d) D'une façon générale, le terme BASE?(W)DONNEES est mani--festement sous-employé dans PASCAL puisque, durant une période de 10 ans, on ne le trouve que 59 fois, alors que DATA(W)BASE? est présent 3994 fois . Il en résulte une grande sensibilité des mesures de couplage à toute présence ou absence de terme, source d'aberrations.

2) Question "bibliographi?".

En retenant le même critère de pertinence que précédemment, c'est-à-dire l'appartenance aux 25 notices de la section 101, on obtient les résultats ci-dessous :

M'_2	P = 0,019	M_4	P = 0,134
M'_4	0,022	I	0,448
M'_1	0,024	K	0,448
M'_3	0,024	J	0,512
M_1	0,096	R	0,538
M_2	0,096	C	0,544
M_3	0,096	L	0,598

La fonction L est encore la plus performante. On obtient l'ordre suivant :

L =

368,3 SERLINE. On-line serials bibliographic and locator retrieval system.

50,5 Accélération de la recherche dans des systèmes de recherche documentaire à descripteurs avec une orga--nisation directe du fonds de recherche.

48,9 Online in industrial and research libraries.

37,2 Planning online search service in a state university.

32,1 Le système d'information juridique JURIS.

32,0 Commission information et documentation. Journée annuelle - Lyon, 3 juin 1976. Introduction de la journée.

31,4 Intérêt de la visualisation conversationnelle en docu--mentation automatique.

31,0 Fast, parallel relaxation screening for chemical patent data-base search.

28,6 RALF : a new software package for the whole complex of punched card oriented documentation.

25,1 Systematic information retrieval and directional data analysis of oligopeptide units in protein data bank.

23,4 A computer data base system for indexing research papers.

22,8 Utilisation des bases de données Chemical Abstracts Service (CAS). Structure du fichier.

21,7 Organisation et utilisation d'un canal de transmission de données pour un système de recherche de l'infor--mation en conversationnel.

20,4 SCORPIO, a subject content oriented retriever for processing information on-line.

20,0 Searching in academia. Nearly 50 libraries tell what they are doing.

19,9 Subject specialists searching Chemical Abstracts on SDC.

18,9 La base de données de la bibliothèque du VTI.

18,2 The book thickness of a graph.

17,9 La CDU et les équipements de recherche.

15,7 Sleeping beauty : MERLIN, a state of the art report.

14,9 PANDORA control system and retrieval language.

13,5 Generalization of the graph center concept, and derived topological centric indexes.

9,0 Un modèle d'accès standard dans les systèmes de gestion de bases de données.

8,4 Gîtes de métaux de base dans les roches sédimentaires : quelques approches.

7,0 Base metal deposits in sedimentary rocks : some approaches.

- -

6,9 A perspective on robotics research and this issue.

.....

5,2 Dispositif pour le tri et la sélection de cartes à encoches latérales supports d'information.

.....

0,3 Système de sécurité pour mémoire informatique.

.....

-0,6 Applications of viewdata.

.....

-5,5 Fréquence, homologie et clonage des plasmides cryptiques du Bacillus thuringiensis.

Là encore, on pourra regretter le rang assez médiocre de "Subject specialists searching Chemical Abstracts on SDC" et l'expliquer en grande partie par la co-occurrence nulle de BIBLIOGRAPHI? et de SDC.

3) Question "graphe?".

En admettant que les 18 notices du paragraphe "Théorie des graphes" de la section 110, ainsi que la notice 81-101-874 sont pertinentes, on obtient les résultats suivants :

M'_2	P = 0,008	M_1	P = 0,504
M'_4	0,183	M_2	0,504
M'_1	0,468	M_3	0,504
M'_3	0,468	M_4	0,504
C	0,468	R	0,540
J	0,468	K	0,540
I	0,468	L	0,753

La fonction L obtient le meilleur résultat. Elle permet de ranger les notices selon l'ordre suivant :

L =

801,2 Finding a minimum equivalent graph of a digraph.

781,0 Randomly k-axial graphs.

753,3 The enumeration of bipartite graphs.

578,8 On some properties of n-tournament ; a note.

440,0 Ramsey numbers involving graphs with long suspended paths.

409,5 Exposants de longueur pour des familles de graphes polytopes.

339,9 A depth first search algorithm to generate the family of maximal independent sets of a graph lexicographi--cally.

281,0 Generalization of the graph center concept, and derived topological centric indexes.

251,5 On covering of random graphs.

228,4 Détermination du nombre structural d'un graphe com--portant des blocs par la méthode des sections.

221,5 Cycles in strong oriented graphs.

213,9 Décomposition des polytopes.

209,1 Recent results in partition (Ramsey) theory for finite lattices.

206,5 Thermodynamic bond graphs and the problem of thermal inertance.

200,4 The book thickness of a graph.

80,5 On weak persistency of Petri nets.

40,5 Commande du mouvement des robots manipulateurs sur la base des algorithmes cinématiques du second ordre.

23,5 Un système pour l'expression et la résolution de pro-blèmes orienté vers un contrôle de robots.

20,1 Polyhedra related to a lattice.

- -

20,0 Fast, parallel relaxation screening for chemical patent data-base search.

14,4 A construction of geodetic blocks.

.....

3,6 A characterization of Robert's inequality for boxicity.

.....

-8,2 Essai comparatif d'un nouveau capteur épidural pour la mesure et la surveillance de la pression intracrâ-nienne.

On peut remarquer le rang assez élevé de "Commande du mouve-ment des robots manipulateurs sur la base des algorithmes ciné-matiques du second ordre" et de "Un système pour l'expression et la résolution de problèmes orienté vers un contrôle de ro-bots" : la présence de ALGORITHMES et ORIENTE, termes souvent utilisés en théorie des graphes, permet de l'expliquer. La mauvaise place de "A characterization of Robert's inequality for boxicity" est due principalement à la fréquence nulle d'occur-rence de BOXICITY. Ce terme n'étant vraisemblablement présent qu'une fois dans la base PASCAL ne peut être sélec-tionné.

4) Question "manipulateur?".

On considère comme pertinentes les 21 notices des paragraphes "Robotique" de la section 110. Les résultats sont les suivants :

M'_2	P = 0,061	M_4	P = 0,462
M'_4	0,100	C	0,612
M'_3	0,272	R	0,612
M'_1	0,424	K	0,617
M_1	0,462	I	0,617
M_2	0,462	L	0,732
M_3	0,462	J	0,734

La fonction de JACQUARD obtient la plus grande valeur de P, suivie immédiatement par L. La liste des notices selon les valeurs décroissantes de L est la suivante :

L =

5911,8 On the equivalence of Lagrangian and Newton-Euler dynamics for manipulators.

3120,1 Commande du mouvement des robots manipulateurs sur la base des algorithmes cinématiques du second ordre.

1875,4 The inverse kinematic problem for anthropomorphic manipulator arms.

1759,0 An adaptive trajectory control of manipulators.

1383,9 Commande de la distribution des forces dans les méca-nismes de robotique comportant des chaînes cinéma-tiques fermées.

1291,5 Control of force distribution in robotic mechanisms containing closed kinematic chains.

1273,7 Construction analytique des systèmes de commande des robots industriels.

1256,7 Robotique et intelligence artificielle.

1254,2 Robots avec des capteurs de force et de moment.

1246,5 Sense-controlled flexible robot behavior.

1232,9 SIGLA. Olivetti robot programming language.

1222,0 Use of optical reflectance sensors in robotics appli-
-cations.

1221,6 Un système pour l'expression et la résolution de pro-
-blèmes orienté vers un contrôle de robots.

1217,6 Optimisation dynamique de ressources et reprogramma-
-tion dynamique en robotique.

1217,5 A perspective on robotics research and this issue.

1211,3 Some critical areas in robotics research.

875,6 A knowledge-based interactive robot-vision system.

364,4 SCORPIO, a subject content oriented retriever for processing information on-line.

86,7 A microcomputer based artificial intelligence labora-
-tory.

86,4 Artificial intelligence, automatic control and develop-
-ment.

67,8 Commande automatique du matériel de soudage par faisceau d'électrons.

- -

24,2 Processus d'apprentissage programmé comme aide à la conception.

.....

-2,0 Laser electro-optic system for rapid three-dimensional (3-D) topographic mapping of surfaces.

.....

-6,6 Influence des sources d'azote sur le métabolisme du glycogène chez Saccharomyces carlsbergensis.

5) Question "robot?".

Comme pour la question précédente, on considère comme pertinentes les 21 notices des paragraphes "Robotique". Les résultats sont les suivants :

M'_2	P = 0,054	M_1	P = 0,408
M'_4	0,075	I	0,501
M'_3	0,204	R	0,689
M'_1	0,340	K	0,694
M_4	0,367	C	0,732
M_2	0,385	J	0,735
M_3	0,385	L	0,818

La fonction L donne le meilleur résultat, selon l'ordre suivant :

L =

4047,5 Sense-controlled flexible robot behavior.

3532,2 SIGLA. Olivetti robot programming language.

2444,6 Commande du mouvement des robots manipulateurs sur la base des algorithmes cinématiques du second ordre.

2416,5 Optimisation dynamique de ressources et reprogramma-tion dynamique en robotique.

2252,7 A knowledge-based interactive robot-vision system.

2110,1 Robotique et intelligence artificielle.

2085,1 Commande de la distribution des forces dans les méca-nismes de robotique comportant des chaînes cinéma-tiques fermées.

2067,5 Robots avec des capteurs de force et de moment.

2067,3 Construction analytique des systèmes de commande des robots industriels.

2046,0 Control of force distribution in robotic mechanisms containing closed kinematic chains.

2027,4 Use of optical reflectance sensors in robotics appli-cations.

2022,7 Un système pour l'expression et la résolution de pro-blèmes orienté vers un contrôle de robots.

2015,1 A perspective on robotics research and this issue.

2009,1 Some critical areas in robotics research.

1698,8 On the equivalence of Lagrangian and Newton-Euler dynamics for manipulators.

427,9 The inverse kinematic problem for anthropomorphic manipulator arms.

377,1 An adaptive trajectory control of manipulators.

165,3 A microcomputer based artificial intelligence labora--tory.

154,2 Artificial intelligence, automatic control and develop--ment.

90,6 Commande automatique du matériel de soudage par faisceau d'électrons.

88,6 The automatic control of electron beam welding equip--ment.

- -

60,8 Utilisation des bases de données Chemical Abstracts Service (CAS). Structure du fichier.

.....

30,8 Processus d'apprentissage programmé comme aide à la conception.

.....

25,5 Laser electro-optic system for rapid three-dimensional (3-D) topographic mapping of surfaces.

.....

-6,7 Fréquence, homologie et clonage des plasmides cryp--tiques du Bacillus thuringiensis.

6) Question "base?(w)données, bibliographi?".

Comme pour les termes isolés vus en 1) et 2), supposons que les 25 notices de la section 101 sont pertinentes. On obtient les résultats suivants :

M'_2	P = 0,019	M_4	P = 0,104
M'_4	0,022	C	0,470
M'_1	0,024	I	0,480
M'_3	0,026	R	0,493
M_2	0,104	K	0,512
M_3	0,096	J	0,544
M_1	0,104	L	0,604

Pour cette question composée de deux termes, la fonction L est encore une fois la plus performante. L'ordre de succession est le suivant :

L =

1149,8 Fast, parallel relaxation screening for chemical patent data-base search.

459,3 SERLINE. On-line serials bibliographic and locator retrieval system.

398,4 Systematic information retrieval and directional data analysis of oligopeptide in protein data bank.

266,1 A computer data base system for indexing research papers.

259,8 Planning online search service in a state university.

241,5 Online in industrial and research libraries.

222,5 Organisation et utilisation d'un canal de transmission de données pour un système de recherche de l'informa-tion en conversationnel.

199,2 Un modèle d'accès standard dans les systèmes de ges-tion de bases de données.

189,2 Utilisation des bases de données Chemical Abstracts Service (CAS). Structure du fichier.

187,5 La base de données de la bibliothèque du VTI.

168,8 Accélération de la recherche dans des systèmes de recherche documentaire à descripteurs avec une orga-nisation directe du fonds de recherche.

134,2 Intérêt de la visualisation conversationnelle en docu-mentation automatique.

121,5 Commission information et documentation. Journée annuelle - Lyon, 3 juin 1976. Introduction de la journée.

99,4 SCORPIO, a subject content oriented retriever for processing information on-line.

99,3 RALF : a new software package for the whole complex of punched card oriented documentation.

73,3 Le système d'information juridique JURIS.

72,8 PANDORA control system and retrieval language.

63,1 Searching in academia. Nearly 50 libraries tell what they are doing.

60,7 Commande du mouvement des robots manipulateurs sur la base des algorithmes cinématiques du second ordre.

53,7 Dispositif pour le tri et la sélection de cartes à enco--ches latérales supports d'information.

52,2 Subject specialists searching Chemical Abstracts on SDC.

33,3 Détermination du nombre structural d'un graphe compor--tant des blocs par la méthode des sections.

32,7 Generalization of the graph center concept, and derived topological centric indexes.

32,6 Optimisation dynamique de ressources et reprogramma--tion dynamique en robotique.

29,9 A depth first search algorithm to generate the family of maximal independent sets of a graph lexicographi--cally.

- -

23,5 La CDU et les équipements de recherche.

.....

11,6 Sleeping beauty : MERLIN, a state of the art report.

6,4 Système de sécurité pour mémoire informatique.

.....

0,1 Applications of viewdata.

.....

-13,7 Effect of suspended particulates on the measurement of gas velocity using a Pitot-static tube.

En posant une question à deux termes, on améliore légèrement la performance de ressaisie par rapport aux questions simples sémantiquement proches BASE?(W)DONNEES et BIBLIOGRAPHI?. L'introduction d'une redondance dans la question permet de faire reculer les notices non pertinentes comme "Commande du mouvement des robots manipulateurs sur la base des algorithmes cinématiques du second ordre".

La succession des notices est fortement influencée par la compo--sante BASE?(W)DONNEES de la question dont le pouvoir discriminant est plus élevé que celui de la composante BIBLIO--GRAPHI?. Cette caractéristique, liée à la rareté du premier terme, se manifeste en réponse aux termes isolés, par un écart important entre début et fin de liste : 1118,8 à -9 pour le premier terme contre 368,3 à -5,5 pour le second terme.

7) Question "base?(w)données, bibliographi?, chimi?".

Il est difficile dans ce cas d'établir un critère de qualité de la réponse à la question. Les deux premiers termes imposent un rangement en tête de liste des notices de la section 101. La présence du troisième terme doit se traduire par un classement privilégié des 3 notices suivantes :

"Fast, parallel relaxation screening for chemical patent data-base search",

"Subject specialists searching Chemical Abstracts on SDC",

"Utilisation des bases de données Chemical Abstracts Service (CAS). Structure du fichier".

Une 4ème notice se rapporte aux banques de données biologiques et peut apparaître assez proche du domaine chimique :

"Systematic information retrieval and directional data analysis of oligopeptide units in protein data bank".

Cependant, la position seule de ces 4 notices dans la liste ne suffit pas à rendre compte de la qualité de l'ordre de classe--ment. Il faut d'une part que ces notices soient en bonne posi--tion, d'autre part qu'elles soient environnées de notices appar--tenant à la section 101. La liste obtenue par la fonction L est la suivante :

L =

1155,9 Fast, parallel relaxation screening for chemical patent data-base search.

457,9 SERLINE. On-line serials bibliographic and locator retrieval system.

397,8 Systematic information retrieval and directional data analysis of oligopeptide units in protein data bank.

265,3 A computer data base system for indexing research papers.

257,7 Planning online search service in a state university.

240,5 Online in industrial and research libraries.

219,1 Organisation et utilisation d'un canal de transmission de données pour un système de recherche de l'information en conversationnel.

197,7 Un modèle d'accès standard dans les systèmes de gestion de bases de données.

197,6 Utilisation des bases de données Chemical Abstracts Service (CAS). Structure du fichier.

186,1 La base de données de la bibliothèque du VTI.

164,8 Accélération de la recherche dans des systèmes de recherche documentaire à descripteurs avec une organisation directe du fonds de recherche.

132,2 Intérêt de la visualisation conversationnelle en documentation automatique.

118,3 Commission information et documentation. Journée annuelle - Lyon, 3 juin 1976. Introduction de la journée.

108,0 RALF : a new software package for the whole complex of punched card oriented documentation.

98,7 SCORPIO, a subject content oriented retriever for processing information on-line.

71,4 PANDORA control system and retrieval language.

70,7 Le système d'information juridique JURIS.

63,5 Subject specialists searching Chemical Abstracts on SDC.

59,4 Searching in academia. Nearly 50 libraries tell what they are doing.

.....

Le présent classement reste identique jusqu'à la 13ème position au classement obtenu par la question BASE?(W)DONNEES, BIBLIOGRAPHI?. L'intervention de la composante CHIMI? fait seulement avancer la notice "Subject specialists searching Chemical Abstracts on SDC" de 2 places.

La forte valeur prise par L dans le cas de questions simples à faible fréquence avantage considérablement le terme BASE?(W)DONNEES. On peut considérer cet inconvénient comme un moindre mal : une relativisation de la part apportée par chaque terme simple q_i de la question en fonction de la quantité d'information de q_i serait pire car elle avantagerait considérablement les termes à forte fréquence. Avec

$$I(q_i) = \frac{1}{p(i)} - 1$$

et $f(i) << N$, la fonction relativisée est sensiblement proportionnelle à f(i) :

$$\frac{L}{I(q_i)} = \frac{f(i)}{N} L .$$

Afin de rendre l'information mutuelle moins sensible à la fréquence du terme-question, un quotient par une fonction du type log $I(q_i)$ pourrait représenter une solution de moyen terme, bien qu'il soit difficile d'y apporter une justification théorique. En fait, avec la simplification

$$\frac{1}{p} - 1 \simeq \frac{1}{p} ,$$

cela reviendrait à relativiser l'information mutuelle au sens de LOSFELD par l'information spécifique locale en q_i au sens de SHANNON, soit $- \log_2 p(i)$.

B. COUPLAGE CODE DE CLASSIFICATION / MOTS-CLES

L'expérimentation consiste à ranger automatiquement chacune des notices dans une section du plan PASCAL. La valeur de la fonction de couplage entre le titre de la publication et les trois premiers chiffres du code de classification sert de critère de rangement.

Dans l'échantillon de notices utilisé pour cette expérimentation, on retient 90 notices présentes dans 25 sections de PASCAL. Ce chiffre de 25 correspond au nombre de sections restées inchan--gées depuis 1973, c'est-à-dire depuis l'année la plus ancienne du fichier en ligne. On évite ainsi les ambiguïtés qui pourraient ré--sulter de fusions ou scissions de sections.

La sélection des sections, permettant le calcul de fréquence des codes de classement, est opéré par la commande SCC=numéro de section.?. Chaque notice est caractérisée par un ensemble de 25 valeurs de la fonction de couplage, une pour chaque section. Chaque notice est attribuée à la section pour laquelle la fonction de couplage est la plus élevée, ce qui apparente l'opération à une expérience de classification automatique. Certaines notices étant indexées dans piusieurs sections de PASCAL, leur attribution à l'une quelconque de ces sections sera considérée comme perti--nente.

La fonction L, qui s'est révélée la plus performante dans la pre--mière partie de l'expérimentation, est choisie comme fonction de couplage, sous sa forme normalisée

$$I = \frac{L}{I(CC_i)} \qquad \text{avec } I(CC_i) = \frac{1}{p(CC_i)} - 1 \ .$$

L'emploi de la forme normalisée est intéressant, pour plusieurs raisons :

- il ne s'agit plus de mots pouvant perdre leur "pouvoir de réso--lution" avec une forte valeur de p(i) ;
- le problème de l'association de plusieurs termes dans une même question ne se pose plus ;
- les différentes fréquences p(i) sont assez voisines les unes des autres (dans la proportion de 1 à 8 pour les valeurs extrêmes contre 1 à 6568 dans le cas précédent) pour limiter les distor--sions.

Pour chaque notice :

- la valeur la plus forte de I est inscrite en caractères gras ;
- dans la case correspondant à la bonne section, la valeur de I est soulignée.

Le détail des résultats pour les 90 notices est reproduit dans les tableaux suivants. Afin de faciliter la lecture, les valeurs négatives de I sont remplacées par le signe - .

Les 25 sections utilisées sont :

101 : Sciences de l'information. Documentation.

110 : Analyse numérique. Informatique. Automatique. Recherche opérationnelle. Gestion. Economie.

120 : Géophysique externe. Astronomie et astrophysique.

130 : Physique mathématique. Optique. Acoustique. Mécanique. chaleur.

140 : Electrotechnique.

145 : Electronique.

161 : Structure de l'état condensé. Cristallographie.

221 : Gisements métalliques et non métalliques. Economie minière.

224 : Stratigraphie. Géologie régionale et géologie générale.

310 : Génie biomédical. Informatique biomédicale. Physique bio-médicale.

320 : Biochimie. Biophysique moléculaire.

330 : Sciences pharmacologiques. Toxicologie.

340 : Microbiologie. Virologie. Immunologie.

361 : Reproduction. Embryologie. Endocrinologie.

362 : Diabète. Maladies métaboliques.

363 : Génétique.

365 : Zoologie des Vertébrés. Ecologie animale. Physiologie appli-quée humaine.

370 : Biologie et physiologie végétales.

390 : Psychologie. Psychopathologie. Psychiatrie.

730 : Combustibles. Energie.

740 : Métaux. Métallurgie.

745 : Soudage, brasage et techniques connexes.

780 : Polymères. Peintures. Bois. Cuirs.

880 : Génie chimique. Industries chimique et parachimique.

885 : Nuisances.

	101	110	120	130	140	145	161	221	224	310	320	330	340	361	362	363	365	370	390	730	740	741	780	880	884
73-101-1742	**2,230**	0,098	-	-	-	-	0,102	-	0,020	0,070	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
73-101-2253	**1,328**	0,660	-	-	0,008	0,027	-	-	-	0,080	-	-	-	-	-	-	0,158	-	0,115	-	-	0,032	-	-	-
76-101-2709	**1,817**	0,448	0,034	-	0,034	0,119	-	0,013	0,053	0,055	-	-	-	-	-	-	-	-	0,042	0,040	-	0,043	-	-	-
73-101-3340	**1,103**	0,098	-	-	0,011	0,023	-	-	-	0,009	-	-	-	-	-	-	-	-	-	0,286	0,033	0,012	0,006	0,009	0,016
76-101-319	**1,072**	1,008	-	0,064	-	-	-	-	-	0,006	-	-	-	-	-	-	-	0,031	0,227	-	-	-	-	-	-
76-101-682	**2,295**	0,401	-	-	0,019	-	-	-	0,049	0,114	-	-	-	-	-	-	0,034	-	0,100	-	-	-	-	-	-
76-101-1963	0,713	0,772	-	0,203	0,333	0,460	0,203	-	0,199	0,074	-	-	-	-	-	0,139	-	-	0,073	-	-	-	-	-	-
76-101-2270	0,628	**0,637**	-	-	0,055	0,156	-	-	-	0,299	-	-	-	-	-	-	0,039	-	0,528	0,070	-	0,054	-	0,027	-
77-101-620	**1,492**	0,560	0,361	-	0,025	-	0,021	-	-	0,177	-	-	-	-	0,010	-	0,028	-	0,334	-	-	-	-	0,450	-
77-101-1124	**1,263**	0,254	0,011	-	0,036	-	-	0,146	0,070	0,093	-	-	-	-	-	-	-	-	0,011	0,383	0,179	0,130	0,056	0,088	0,120
77-101-1903	**0,384**	**1,025**	0,123	-	0,015	0,120	-	0,109	-	0,071	-	-	-	-	-	-	-	-	-	0,045	0,013	0,011	-	-	-
77-101-3172	**1,253**	0,380	-	-	-	0,034	-	-	-	0,011	-	-	-	-	-	-	-	-	0,017	-	-	-	-	-	-
77-101-4002	**1,064**	0,148	0,066	-	-	0,049	-	-	0,023	-	-	-	-	-	-	-	-	-	-	0,523	0,047	-	-	-	0,005
80-101-2027	**0,621**	0,114	0,062	-	0,003	0,034	-	-	0,055	0,092	-	-	-	-	-	-	-	-	-	0,198	0,097	0,146	0,170	0,187	0,009
80-101-2277	**1,134**	0,212	-	-	0,002	0,021	**0,627**	-	-	0,030	-	-	-	-	-	-	-	-	-	-	-	0,174	0,022	-	-

	101	110	120	130	140	145	161	221	224	310	320	330	340	361	362	463	765	370	390	730	740	745	784	880	RET
80-101-3496	0,254	-	-	-	0,280	-	-	0,037	+	0,030	+	0,014	-	-	-	-	0,230	-	0,590	0,090	-	-	-	0,076	-
81-101-874	0,388	0,701	-	0,117	0,028	0,173	0,018	-	0,005	-	-	-	-	-	-	0,252	-	0,163	0,012	0,057	-	-	-	-	-
81-101-1517	0,418	0,685	0,123	0,061	0,011	0,133	-	+	0,069	0,088	-	-	-	-	-	-	-	-	-	0,162	-	-	0,315	0,060	0,064
81-101-3056	1,067	1,185	0,018	+	0,036	0,186	-	-	-	0,661	-	-	-	-	+	-	0,074	-	0,020	0,155	-	0,072	-	0,218	+
81-101-3405	1,279	0,259	0,114	0,007	-	-	-	+	0,098	-	-	-	-	-	-	-	-	-	0,445	0,139	-	-	-	-	-
81-101-3964	1,015	0,081	-	-	0,027	-	-	0,017	-	0,604	-	-	-	-	-	-	-	-	-	-	-	0,064	0,012	0,081	0,082
81-101-4630	1,376	0,077	0,110	+	0,000	0,076	0,042	+	0,004	0,070	0,015	+	+	+	+	-	0,029	-	0,041	-	-	-	-	-	-
82-101-536	1,100	0,114	0,021	-	-	-	-	-	0,002	0,014	-	-	-	-	0,035	-	-	-	0,253	-	-	-	-	-	0,016
82-101-1951	1,270	0,627	0,075	-	0,019	0,006	-	0,027	0,010	0,021	-	-	-	-	-	-	-	-	-	0,124	-	-	-	0,001	0,005
82-101-2294	0,947	0,834	0,017	0,080	0,030	0,081	-	0,215	-	0,024	-	-	-	-	-	-	-	-	-	0,029	-	0,107	0,401	0,043	-
75-110-10869	-	0,276	-	-	-	-	0,016	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
75-110-12656	-	1,059	-	-	-	+	+	+	0,032	-	-	+	-	+	-	-	-	0,046	-	-	-	-	0,021	-	-
77-110-10448	-	0,592	-	0,210	0,057	0,068	0,168	+	0,007	-	0,007	-	-	+	-	+	-	-	-	0,118	0,091	0,034	0,163	0,034	-
80-110-2241	-	1,624	-	-	-	0,021	-	-	-	-	-	-	0,223	+	-	+	-	-	-	-	-	-	-	-	-
80-110-4463	-	0,134	+	0,210	-	-	-	+	-	-	0,025	-	0,013	-	-	0,010	-	-	-	-	-	-	-	0,146	-

	101	110	120	130	140	145	161	221	224	310	320	330	340	361	362	363	365	370	390	730	740	765	780	900	905
80-110-8633	0,251	0,102	-	0,080	0,001	0,101	0,067	0,049	0,016	-	-	-	-	-	-	-	-	-	-	0,425	0,126	0,054	-	0,000	-
81-110-15931	-	1,628	-	0,057	-	0,027	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0,005	-	-	-	-
81-110-17957	-	10,728	-	0,651	-	-	0,250	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
82-110-224	-	0,046	-	0,056	-	+	0,384	-	-	-	-	-	0,414	-	-	-	0,089	+	0,018	-	0,065	-	-	-	-
82-110-1325	+	0,952	+	0,305	0,025	0,010	+	-	-	+	+	-	-	-	+	0,021	+	-	-	0,016	-	-	-	0,012	0,004
82-110-2411	0,601	2,542	0,108	0,022	0,125	0,154	+	+	-	+	+	-	-	+	+	0,155	+	-	0,414	0,042	-	-	-	-	-
82-110-3701	-	0,017	0,798	-	0,084	-	-	-	0,001	-	-	0,111	-	-	-	-	-	-	-	0,044	-	0,058	0,023	-	-
82-110-6832	-	0,408	0,034	0,022	-	+	0,176	-	0,020	-	-	-	-	-	-	-	-	-	-	0,005	0,011	-	0,010	-	-
82-110-5907	-	0,535	0,085	0,153	-	0,001	-	-	-	-	-	-	0,195	-	-	0,010	-	-	0,013	-	-	-	-	-	0,007
82-110-7156	-	0,814	-	0,084	-	0,143	0,035	-	-	+	+	-	-	-	+	+	+	+	-	0,018	0,029	0,052	0,018	-	-
82-110-8996	0,056	2,042	0,032	-	0,121	0,179	-	-	-	0,001	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
82-110-11126	-	0,473	-	0,269	0,011	-	0,707	-	-	-	-	-	-	+	+	+	-	-	-	1,516	0,098	0,022	0,053	0,011	-
82-110-13533	-	1,319	-	0,242	-	-	0,072	-	-	+	+	-	-	-	-	0,024	-	+	0,030	-	0,007	-	-	-	-
83-110-9201	0,016	0,962	-	-	0,291	0,034	0,036	0,083	-	-	-	-	-	+	-	-	+	-	+	0,319	0,109	0,216	0,030	0,122	0,075
83-110-16641	0,163	2,602	+	+	+	0,019	0,005	-	-	0,034	-	-	-	-	+	+	-	-	0,187	-	0,096	0,097	-	-	-

	101	110	120	130	140	145	161	221	224	310	320	330	340	361	362	363	365	370	390	730	740	745	780	880	885
B1-110-16685	-	1,950	-	-	-	-	-	-	-	0,015	0,056	-	-	-	-	-	-	-	0,044	-	-	-	-	-	-
B1-110-19145	-	0,468	-	0,090	0,005	0,113	-	-	-	0,023	-	-	-	-	-	0,083	0,228	-	0,079	-	-	0,037	-	-	0,014
B1-110-19215	0,047	0,394	0,040	0,484	0,010	0,324	-	-	-	0,017	-	-	-	-	-	-	0,209	-	0,037	0,108	-	0,204	0,032	-	0,139
B2-110-580	-	0,522	0,051	0,321	0,032	-	0,045	-	-	0,010	0,047	-	-	-	-	-	-	-	-	0,112	0,063	0,144	0,110	-	-
B2-110-580 trad.	-	0,794	-	0,263	0,196	0,008	-	-	-	-	-	-	-	-	-	-	-	-	-	0,065	0,135	0,181	0,170	0,014	-
B2-110-1637	0,117	0,487	-	-	-	0,017	-	0,178	0,007	-	-	-	-	-	-	-	-	-	0,306	0,331	-	0,119	-	-	0,026
B2-110-2811	0,001	1,472	-	0,007	-	-	-	-	-	0,000	-	-	-	-	-	-	-	-	0,177	-	-	0,108	0,122	-	-
B2-110-4168	0,067	0,425	0,005	0,033	-	-	-	0,076	0,775	-	-	-	-	-	-	-	-	-	0,005	0,120	0,001	0,116	-	-	0,055
B2-110-4185	0,056	0,692	-	-	0,043	0,061	-	-	-	0,087	-	-	-	-	-	-	0,023	-	0,920	0,254	-	0,035	-	0,010	0,030
B2-110-5112	0,008	1,959	0,062	0,260	0,197	0,034	0,194	-	0,002	0,024	-	-	-	-	-	-	-	-	-	-	0,160	0,117	-	-	-
B2-110-6383	-	0,455	0,035	0,447	0,136	0,072	-	-	-	0,026	-	-	-	-	-	-	-	-	-	0,160	0,097	0,132	-	-	-
B2-110-7852	0,130	0,774	-	-	0,023	-	-	-	-	0,087	-		-	-	-	-	-	-	0,541	0,061	0,080	0,055	-	-	0,023
B2-110-7876	-	0,027	0,436	1,230	-	0,595	0,208	-	0,030	0,039	-	-	0,012	-	-	0,076	-	-	-	-	-	-	-	-	-
B2-110-9712	-	1,268	0,048	0,624	0,103	0,072	-	-	-	0,359	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
B2-110-9797	0,204	1,140	-	0,147	-	0,006	0,071	-	-	-	-	-	0,188	-	-	0,233	-	-	0,096	-	-	0,119	-	-	-

	101	110	120	130	140	145	161	221	224	310	320	330	340	341	342	343	345	370	391	730	740	745	780	880	885
82-110-11912	-	**0,993**	0,014	0,126	0,0600	0,171	-	0,013	-	-	-	-	-	-	-	0,108	-	-	-	0,062	0,014	0,116	-	-	-
82-110-11975	0,054	**1,104**	-	-	0,013	0,017	+	-	-	0,123	+	+	-	-	+	+	-	-	0,459	0,029	-	-	-	-	-
82-110-11952	-	**0,427**	0,148	0,253	-	-	+	-	-	0,079	-	-	-	-	0,044	0,057	0,306	-	0,406	-	-	-	-	-	-
82-110-12200	0,010	**0,843**	0,087	-	-	-	-	-	-	0,054	0,046	-	-	-	-	-	-	-	0,443	-	-	0,107	-	-	-
82-120-7176	-	0,032	**1,170**	-	-	-	0,027	-	-	-	-	-	0,093	-	-	0,004	-	-	0,012	-	-	-	-	-	-
82-130-10892	-	-	0,194	1,050	0,208	0,159	-	-	-	0,026	-	-	-	-	-	-	-	-	-	**1,053**	0,005	0,049	-	0,006	0,401
82-130-11736	-	0,200	-	-	**1,805**	0,493	-	-	-	-	-	0,001	-	-	-	-	-	-	-	-	0,009	0,021	-	0,012	-
82-145-4820	-	0,147	0,092	0,038	0,944	**1,495**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0,000	-	-	-
82-161-1305	-	-	-	0,019	-	-	0,429	-	-	-	0,032	-	-	-	-	-	-	-	-	-	**1,030**	0,021	0,520	-	-
82-161-1305 trac	-	-	-	-	-	-	0,568	0,302	0,093	-	-	-	-	-	-	-	-	-	-	-	**1,091**	0,107	0,459	-	-
82-221-616	0,093	0,025	-	-	-	-	0,156	**0,497**	0,324	-	-	-	-	-	-	-	-	-	-	-	0,197	0,021	-	-	0,018
82-221-616 trac	-	0,340	-	-	-	-	-	**0,844**	0,242	-	-	-	-	-	-	-	-	-	-	-	0,223	0,066	-	-	-
82-224-2267	-	-	-	-	-	-	-	0,200	**0,669**	-	-	-	-	-	-	-	-	-	-	0,198	-	-	0,032	-	0,016
82-310-501	-	-	-	0,260	0,163	0,053	-	-	-	0,153	-	0,081	-	0,108	-	-	-	-	-	**0,346**	0,032	0,077	-	-	0,134
82-120-512	-	0,136	-	0,164	-	-	-	-	-	-	**2,186**	0,058	0,148	1,054	0,324	-	0,913	-	-	-	-	-	-	-	0,001

	101	116	120	130	140	145	161	221	224	310	320	330	340	361	362	363	365	378	390	730	760	765	780	880	885
82-330-6962	-	-	-	-	0,045	*	*	-	-	0,013	-	1,620	0,014	0,129	0,053	*	0,060	*	0,327	*	-	-	-	-	-
82-340-8664	*	*	0,014	*	*	*	*	-	-	-	0,315	0,016	1,540	0,148	0,299	0,316	0,256	-	-	-	-	-	-	-	0,046
82-361-8820	-	*	*	*	*	*	*	-	-	-	0,444	0,260	0,234	1,163	0,044	-	0,348	-	-	-	-	-	-	-	-
82-362-1974	-	-	-	-	-	-	-	-	-	-	0,198	0,148	0,910	1,329	1,926	*	0,338	*	*	-	-	-	-	-	-
82-363-3815	-	-	*	-	0,076	-	-	-	-	-	0,447	-	1,546	-	-	1,753	0,117	*	-	*	-	-	-	-	*
82-365-10949	*	-	-	-	-	-	0,223	-	0,034	-	-	-	-	0,101	-	0,388	0,779	0,256	-	-	-	-	-	-	*
82-370-4608	0,073	-	*	-	-	-	-	-	-	0,044	0,273	-	0,032	-	-	0,704	0,005	0,616	-	-	-	0,176	-	0,178	-
82-371-17	0,057	0,072	-	0,084	0,000	-	-	0,122	-	-	-	-	-	-	0,059	*	0,173	*	1,283	0,273	-	-	-	-	-
82-380-9773	-	-	0,046	-	0,012	-	-	0,183	0,023	-	-	-	-	-	-	-	-	-	-	1,518	-	-	-	-	0,188
82-740-5825	-	*	-	-	-	-	-	-	-	-	-	-	-	-	-	0,169	-	-	-	0,147	1,570	0,214	0,112	0,112	-
82-745-1487	0,149	0,202	*	0,107	0,099	0,373	0,035	-	-	0,045	-	-	*	*	*	-	-	-	-	0,430	0,287	1,102	-	-	-
82-745-1497	0,089	0,426	-	0,018	0,243	0,336	0,080	-	-	0,616	-	-	-	-	-	-	-	-	-	0,046	0,196	1,228	*	0,006	-
82-765-509	*	*	*	0,008	0,041	0,118	0,051	-	-	-	-	-	-	-	-	-	-	-	-	-	0,287	0,095	1,171	0,066	-
82-880-175	-	0,231	*	-	-	-	-	-	-	0,013	0,173	-	0,089	-	-	0,029	-	-	-	0,391	0,067	0,009	0,047	0,176	0,185
82-885-1996	-	-	0,097	*	-	-	0,142	-	-	-	-	*	*	*	*	-	-	-	-	0,627	*	0,008	*	*	1,300

On constate que 74 notices sur 90 sont attribuées à un bon code de classification, soit un pourcentage de 82 %.
La fonction L non normalisée donne des résultats inférieurs, notamment pour les notices traitant de robotique : elle permet de bien classer 69 notices, soit un pourcentage de 77 %.
Ces chiffres peuvent être rapprochés de ceux de MARON - 52 % sur les documents extérieurs à l'échantillon de départ - et de HOYLE - 78 % - (58).

BIBLIOGRAPHIE

1. ACKOFF (R. L.). - Towards a behavioral theory of communication. In : Management Sci., 4, 1958, 218-234.

2. ANDREEWSKY (A.). FLUHR (C.). - Indexation automatique : maintenance et gestion d'un système documentaire : 1ère partie : aspects théoriques.
- Saclay : Centre d'études nucléaires, 1973.

3. ANDREEWSKY (A.). FLUHR (C.). RAMBOUSEK (J.). - Automati--sation de l'analyse discriminante, de l'indexation, de la recherche hiérarchisée des documents et de l'aide à la décision.
- Saclay : Centre d'études nucléaires, 1973.

4. ATLAN (H.). - Du bruit comme principe d'auto-organisation. In : Communication, 18, 1972, 21-36.

5. ATLAN (H.). - L'Organisation biologique et la théorie de l'infor--mation.
- Paris : Hermann, 1972.

6. ATLAN (H.). - L'Evolution des concepts de temps et d'information en biologie.
In : Dix visions sur la communication humaine ; Lyon, Presses universitaires de Lyon, 1981.

7. BAR-HILLEL (Y.). - An examination of information theory.
In : Language and information : selected essays on their theory and application / Y. Bar-Hillel ; Reading, Addison-Wesley, 1964.

8. BELKIN (N. J.). - Information concepts for information science. In : J. Doc., 34, 1978, 55-85.

9. BELZER (J.). - Information theory as a measure of information content.
In : J. A.S.I.S., 24, 1973, 300-304.

10. BOYCE (B. R.). MARTIN (D.). - The Brillouin measure of an author's contribution to a literature in psychology.
In : J. A.S.I.S., 32, 1981, 73-76.

11. BRILLOUIN (L.). - La Science et la théorie de l'information.
- Paris : Masson, 1959.

12. BRINER (L. L.). - Identifying keywords in text data processing. In : Annual technical symposium Assoc. Comput. Mach. - Nat. Bur. Stand., 15, 1976, 85-90.

13. BRINER (L. L.). - A mathematical theory of indexing.
In : Information age in perspective, A.S.I.S. Annual meeting, 41, 1978, 55-58.

14. BROOKES (B. C.). - The Shannon model of IR systems.
In : J. Doc., 28, 1972, 160-162.

15. BROOKES (B. C.). - Measurement in information science : objec--tive and subjective metrical space.
In : J. A.S.I.S., 31, 1980, 248-255.

16. CALLON (M.). COURTIAL (J.-P.). TURNER (W. A.). BAUIN (S.). - De l'opération de traduction à la constitution de réseaux problé--matiques : l'analyse des mots associés dans la littérature scien--tifique et technique.
- Paris : Centre de sociologie de l'innovation, 1982.

17. CARNAP (R.). BAR-HILLEL (Y.). - An outline of a theory of semantic information.
In : Language and information : selected essays on their theory and application / Y. Bar-Hillel ; Reading, Addison-Wesley, 1964.

18. CAWKELL (A. E.). - Simplified information theory and data trans--mission : 1 & 2.
In : Electrical Engng, 39, 1967, 212-218 & 302-309.

19. CAWKELL (A. E.). - A measure of "efficiency factor" : commu--nication theory applied to document selection systems.
In : Inform. Process. Manag., 11, 1975, 243-248.

20. CHERRY (C.). - On human communication : a review, a survey, and a criticism.
- New York : Technology Press of Massachusetts Institute of Tech--nology ; John Wiley, 1957.

21. CHONEZ (A.). - Bibliographie.
In : Documentaliste, 18, 1981, 238-239.

22. CLEVERDON (C.) et al. - Factors determining the performance of indexing systems.
- Cranfield : College of aeronautics, 1966.

23. CONANT (R. G.). - Detecting subsystems of a complex system.
In : IEEE Trans. Systems, Man, Cybernetics, 2, 1972, 550-553.

24. CONVERT (G.). - Entropie et théorème de Bayes en théorie de l'estimation.
In : R. tech. Thomson-CSF, 14, 1982, 5-17.

25. COOPER (D.). LYNCH (M. F.). - Review of variety generation techniques ... : consolidation report on variety generation research funded by the British Library Research and Development Depart--ment : 1971-1980.
- Sheffield : Postgraduate school of librarianship and information science, 1980.

26. COSNIER (J.). - Le Statut du langage dans la communication humaine.
In : Dix visions sur la communication humaine ; Lyon, Presses universitaires de Lyon, 1981.

27. CRAMPES (J.-B.). - Aide à l'interrogation d'un dictionnaire de données.
In : R. Autom. Inform. Rech. opér., Inform., 14, 1980, 87-95.

28. DA ROCHA PARANHOS (W. M. M.). - Application of an entropy measure for journal evaluation and its comparison to other quanti--tative measures.
Th. : Ph. D. : Case Western Reserve Univ., Cleveland : 1981.

29. DEBILI (F.). - Analyse syntaxico-sémantique fondée sur une acqui--sition automatique de relations lexicales-sémantiques.
Th. : Sc. math. : Paris 11, Orsay : 1982.

30. DELERIS (M.). - Analyse - modélisation - optimisation d'un sous-ensemble de production de zinc.
Th. : Doct.-ing. : Toulouse : 1981.

31. DOLBY (J. L.). - On the notions of ambiguity and information loss.
In : Inform. Process. Manag., 13, 1977, 69-77.

32. DOYLE (L. B.). - The Microstatistics of text.
In : Inform. Stor. Retr., 1, 1963, 189-214.

33. DUFOUR (J.). - Méthodes et méthodologie d'analyse de systèmes complexes : application aux procédés industriels et aux systèmes macro-économiques.
Th. : Sc. : Lyon : 1979.

34. DUSSAUCHOY (A.). - Résultats récents en théorie de l'informa--tion : application à l'analyse structurale.
In : Structures économiques et économétrie, Colloque C.N.R.S., Lyon, 1980.

35. ESCARPIT (R.). - Théorie générale de l'information et de la communication.
- Paris : Hachette, 1976.

36. FAIRTHORNE (R. A.). - Documentary classification as a self-organizing system.
In : Information theory, Symposium, London, 1960 ; London, Butter--worths, 1961, 426-436.

37. FAIRTHORNE (R. A.). - Morphology of "information flow".
In : J. Assoc. Comput. Mach., 14, 1967, 710-719.

38. FAIRTHORNE (R. A.). - Empirical hyperbolic distributions (Brad--ford - Zipf - Mandelbrot) for bibliometric description and pre--diction.
In : J. Doc. , 25, 1969, 319-343.

39. FANO (R. M.). - Information theory and the retrieval of recorded information.
In : Documentation in action, based on 1956 Conference on docu--mentation at Western Reserve University ; New York, Reinhold, 1956.

40. FLUHR (C.). - Présentation technique de SPIRIT.
- Paris : Compagnie internationale de services en informatique, 1982.

41. FOREST (F.). - Une application de l'information hyperbolique à la recherche documentaire.
Th. 3eme cycle : Informatique : Paris 6 : 1974.

42. FOREST (F.). - Une approche théorique des problèmes de bruit et de silence en recherche documentaire.
In : Journées mancelles Information et questionnaires, Le Mans, 1980 ; Structures de l'information, Publ., 18, 105-116.

43. FOREST (F.). - Recherche documentaire : une formalisation du comportement des différents interlocuteurs.
In : Documentaliste, 19, 1982, 16-19.

44. GARFIELD (E.). - Information theory and other quantitative factors in code design for document card systems.
In : J. Chem. Doc., 1, 1961, 70-75.
Reproduit dans : Current Contents, 44, 1977, 8-19.

45. GARFIELD (E.). - Information theory and all that jazz : a lost reference list leads to a pragmatic assignment for students.
In : Current Contents, 44, 1977, 5-7.

46. GARFIELD (E.). - Citation indexing : its theory and application in science, technology and humanities.
- New York : J. Wiley, 1979.

47. GARLAND (K.). - An application of information theory for mate--rials selection and collection evaluation.
Th. : Ph. D. : Case Western Reserve Univ., Cleveland : 1980.

48. GOFFMAN (W.). NEWILL (V. A.). - Communication and epidemic processes.
In : Proc. Royal Soc., 298, 1967, 316-334.

49. GOOD (I. J.). - Discussion.
In : Documentary classification as a self-organizing system / R. A. Fairthorne.
In : Information theory, Symposium, London, 1960 ; London, Butter--worths, 1961.

50. GUAZZO (M.). - Retrieval performance and information theory.
In : Inform. Process. Manag., 13, 1977, 155-165.

51. GUIASU (S.). THEODORESCU (R.). - Incertitude et information.
- Québec : Presses de l'Université Laval, 1971.

52. HAYES (R. M.). - The Measurement of information from a file.
In : Statistical association methods for mechanized documentation, Symposium, Washington, 1964 ; Nat. Bur. Stand. Miscellaneous Publ., 269, 1965, 161-162.

53. HAYES (R. M.). - Weighted entropy : a literature review.
- Los Angeles : Graduate school of library and information science, 1981.

54. HAYES (R. M.). BORKO (H.). - Mathematical models of informa--tion system use.
In : Inform. Process. Manag., 19, 1983, 173-186.

55. HENRY-LABORDERE (A.). - Analyse des données : applications et méthodes pratiques.
- Paris : Masson, 1977.

56. HIROU (P.). Communication personnelle, 1982.

57. HOLLNAGEL (E.). - Is information science an anomalous state of knowledge ?
In : J. Inf. Sci., 2, 1980, 183-187.

58. HOYLE (W. G.). - Automatic indexing and generation of classi--fication systems by algorithm.
In : Inform. Stor. Retr., 9, 1973, 233-242.

59. JACQUESSON (A.). SCHIEBER (W. D.). - Term association analysis on a large file of bibliographic data, using a highly-controlled indexing vocabulary.
In : Inform. Stor. Retr., 9, 1973, 85-94.

60. JARDINE (N.). SIBSON (R.). - Mathematical taxonomy.
- London : John Wiley, 1971.

61. KAMPE DE FERIET (J.). - La Théorie généralisée de l'information et la mesure subjective de l'information.
In : Théories de l'information, actes des Rencontres de Marseille-Luminy, 1973 ; Berlin , Springer, 1974.

62. KAMPE DE FERIET (J.). - Les Deux points de vue de la théorie de l'information : information a priori, information a posteriori.
In : Théorie de l'information : développements récents et appli--cations, Colloque international du C.N.R.S., Cachan, 1977 ; Paris, C.N.R.S., 1978.

63. KESSLER (M. M.). - Bibliographic coupling extended in time : ten case histories.
In : Inform. Stor. Retr., 1, 1963, 167-187.

64. LANGLOIS (R. N.). - Systems theory and the meaning of infor--mation.
In : J. A.S.I.S., 33, 1982, 395-399.

65. LEGENDRE (L.). LEGENDRE (P.). - Ecologie numérique : 1 : Le traitement multiple des données écologiques.
- Paris : Masson, 1979.

66. LOSFELD (J.). - Information fournie par un ensemble d'observa--teurs et applications aux questionnaires et à l'analyse des don--nées.
Th. : Sc. math. : Lille : 1974.

67. LOUIS-GAVET (G.). - Diverses applications issues d'une fonction f de compactage basée sur une étude mathématique du langage natu--rel : compactage de données, comparaison de textes, Hash-coding.
In : R. Autom. Inform. Rech. opér., Inform., 12, 1978, 47-71.

68. LUSSATO (B.). - Théorie de l'information et processeur humain. - Saint Sulpice de Favières : Ed. Jean-Favard, 1980.

69. MACKAY (D. M.). - Information, mechanism and meaning. - Cambridge, Mass. : M.I.T. Press, 1969.

70. MAESTRACCI (J.-M.). - Aide à l'interrogation dans un système de documentation automatique. Th. 3ème cycle : Math. appliq. : Lille 1 : 1971.

71. MANDELBROT (B.). - An informational theory of the statistical structure of language. In : Communication theory, Symposium, London, 1952 ; London, Butterworths, 1953.

72. MARON (M. E.). - Automatic indexing : an experimental inquiry. In : J. Assoc. Comput. Mach., 8, 1961, 404-417.

73. MARON (M. E.). - A logician's view of language-data processing. In : Natural language and the computer ; New York, McGraw-Hill, 1963.

74. MARSHAK (J.). - Problems in information economics. In : Management controls ; New York, McGraw-Hill, 1964.

75. MAX (J.). - Théorie de l'information appliquée aux mesures. In : Techniques de l'ingénieur, exposé R 353, 1982, 1-15.

76. MAZUR (M.). - Les Principes de la théorie qualitative de l'information. In : Réflexions sur de nouvelles approches dans l'étude des systèmes, Paris, 1975 ; Paris, E.N.S.T.A., 1976.

77. MEETHAM (A. R.). - Communication theory and the evaluation of information retrieval systems. In : Inform. Stor. Retr., 5, 1969, 129-134.

78. MEYER-EPPLER (W.). - Grundlagen und Anwendungen der Informationstheorie. - 2. Aufl. - Berlin : Springer, 1969.

79. MEYRIAT (J.). - Exposé. Séminaire de Sciences de l'information, Paris, 1982.

80. MILLER (G. A.). - The Magical number seven, plus or minus two : some limits on our capacity for processing information. In : The Psychology of communication : seven essays / G. A. Miller ; New York, Basic Books, 1967.

81. MOLES (A.). - Théorie de l'information et perception esthétique. - Paris : Denoël-Gonthier, 1972.

82. MOLES (A.). - Préface. In : Théorie mathématique de la communication / W. Weaver, C. E. Shannon ; Paris, Retz-C.E.P.L., 1975.

83. MUGUR-SCHACHTER (M.). - Le Concept nouveau de fonctionnelle d'opacité d'une statistique : étude des relations entre la loi des grands nombres, l'entropie informationnelle et l'entropie statis--tique.
In : A. Inst. Henri Poincaré, Sect. A, 32, 1980, 33-71.

84. PIETILAINEN (P.). - Relation of resemblance in information retrieval.
In : Inform. Process. Manag., 18, 1982, 55-59.

85. PIETILAINEN (P.). - Local feedback and intelligent automatic query expansion.
In : Inform. Process. Manag., 19, 1983, 51-58.

86. PINSON (G.). - Vers un modèle "hologrammorphique" de l'infor--mation.
In : Dix visions sur la communication humaine ; Lyon, Presses universitaires de Lyon, 1981.

87. POOLE (R. W.). - An introduction to quantitative ecology.
- New York : McGraw-Hill, 1974.

88. POPPER (K. R.). - La Connaissance objective.
- Bruxelles : Complexe ; Paris : distr. P.U.F., 1978.

89. QUASTLER (H.). - A primer on information theory.
In : Symposium on information theory in biology, Gatlinburg, 1956 ; London, Pergamon, 1958.

90. RENYI (A.). - Calcul des probabilités avec un appendice sur la théorie de l'information.
- Paris : Dunod, 1966.

91. RICHETIN (M.). - Analyse structurale des systèmes complexes en vue d'une commande hiérarchisée.
Th. : Sc. : Toulouse : 1975.

92. RIP (A.). - Scientometric studies of biotechnology.
In : Conference of the European association for the study of science and technology, Deutschlandsberg, 1982.

93. ROBERTSON (S. E.). - Theories and models in information re--trieval.
In : J. Doc., 33, 1977, 126-148.

94. ROUAULT (B.). - Essai de diverses méthodes de classification automatique en vue de la constitution d'un langage documentaire.
Th. 3ème cycle : Math. : Nancy 1 : 1972.

95. SALTON (G.). - Automatic indexing : a summary.
In : La Recherche sur la gestion de l'information en Europe, Conférence EURIM 5, Versailles, 1982.

96. SCHUTZENBERGER (M.-P.). - La Théorie de l'information.
In : Information et communication ; Paris, Maloine, 1983.

97. SEBEOK (T. A.). - The Informational model of language : analog and digital coding in animal and human communication.
In : Natural language and the computer ; New York, McGraw-Hill, 1963.

98. SHANNON (C. E.). - La Théorie mathématique de la communi--cation.
In : Théorie mathématique de la communication / W. Weaver, C. E. Shannon ; Paris, Retz-C.E.P.L., 1975.

99. SHAW (W. M., Jr.). - Entropy, information and communication.
In : Information choice and policies, A.S.I.S. Annual meeting, 42, 1979, 32-40.

100. SHAW (W. M., Jr.). - Information theory and scientific communi--cation.
In : Scientometrics, 3, 1981, 235-249.

101. SHAW (W. M., Jr.). - Statistical disorder and the analysis of a communication-graph.
In : J. A.S.I.S., 34, 1983, 146-149.

102. SPARCK JONES (K.). - A statistical interpretation of term speci--ficity and its application in retrieval.
In : J. Doc., 28, 1972, 11-21.

103. SPARCK JONES (K.). VAN RIJSBERGEN (C. J.). - Information retrieval test collections.
In : J. Doc., 32, 1976, 59-75.

104. STANCIU (L.). - The Epistemologic and praxiological functions of the information : 1 : Development of information theory, critical considerations, 2 : Unity between information theory, epistemology and praxiology, the cognitive-informational field.
In : Probl. Inf. şi Doc., 16, 1982, 10-16 & 66-74.

105. STARYNKEVITCH (D.). - Quelques programmes d'analyse lexico--graphique et de traitement de texte.
- S.l. : I.B.M.-France, 1979.

106. STILES (H. E.). - The Association factor in information retrieval.
In : J. Assoc. Comput. Mach., 8, 1961, 271-279.

107. THINES (G.). LEMPEREUR (A.). - Dictionnaire général des sci--ences humaines.
- Paris : Ed. universitaires, 1975.

108. VAN RIJSBERGEN (C. J.). - Information retrieval. - 2nd ed.
- London : Butterworths, 1979.

109. WEAVER (W.). - Contributions récentes à la théorie de la commu--nication.
In : Théorie mathématique de la communication / W. Weaver, C. E. Shannon ; Paris, Retz-C.E.P.L., 1975.

110. YANNAKOUDAKIS (E. J.). GOYAL (P.). HUGGIL (J. A.). - The Generation and use of text fragments for data compression.
In : Inform. Process. Manag., 18, 1982, 15-21.

111. YOVITS (M. C.) et al. - Information flow and analysis : theory, simulation, and experiments : 1 : Basic theoretical and conceptual development, 2 : Simulation, examples, and results.
In : J. A.S.I.S., 32, 1981, 187-210.

112. ZUNDE (P.). - Information theory and information science.
In : Inform. Process. Manag., 17, 1981, 341-347.

113. ZUNDE (P.). SLAMECKA (V.). - Distribution of indexing terms for maximum efficiency of information transmission.
In : Am. Doc., 18, 1967, 104-108.