CEA-N-1795

- Note CEA-N-1795 -

Centre d'Etudes Nucléaires de Saclay
Service de Documentation

INDEXATION AUTOMATIQUE

CONSTRUCTION AUTOMATIQUE DES THESAURUS

CLASSIFICATION AUTOMATIQUE

par

Alexandre ANDREEWSKY, Christian FLUHR

- Juin 1975 -

CEA-N-1795 - ANDREEWSKY Alexandre, FLUHR Christian

INDEXATION AUTOMATIQUE, CONSTRUCTION AUTOMATIQUE DES THESAURUS CLASSIFICATION AUTOMATIQUE

Sommaire.- Dans ce travail on rappelle d'abord les principes de l'indexation automatique telle qu'elle a été définie par les auteurs dans plusieurs travaux précédents. On fait ensuite la comparaison et la synthèse des travaux des auteurs et des travaux d'une équipe soviétique de l'Institut INFORM-ELECTRO de Moscou. On examine les problèmes mathématiques et linguistiques de la construction automatique des thésaurus et de la classification automatique.

1975 36 p.

Commissariat à l'Energie Atomique - France

CEA-N-1795 - ANDREEWSKY Alexandre, FLUHR Christian

AUTOMATIC INDEXING, COMPILING AND CLASSIFICATION

Summary.- A review of the principles of automatic indexing, as defined by the authors in several earlier papers, is followed by a comparison and summing-up of work by the authors and by a Soviet staff from the Moscou INFORM-ELECTRO Institute. The mathematical and linguistic problems of the automatic building of thesaurus and of automatic classification are examined.

1975 36 p.

Commissariat à l'Energie Atomique - France

Note CEA-N-1795

DESCRIPTION-MATIERE *(mots clefs extraits du thesaurus SIDON/INIS)*

en francais	*en anglais*
THERMINOLOGIE NORMALISEE	STANDARDIZED TERMINOLOGY

- Note CEA-N-1795 -

Centre d'Etudes Nucléaires de Saclay
Service de Documentation

INDEXATION AUTOMATIQUE
CONSTRUCTION AUTOMATIQUE DES THESAURUS
CLASSIFICATION AUTOMATIQUE

par

Alexandre ANDREEWSKY, Christian FLUHR

Table des matières

Cette note a été partiellement exposée au Congrès de l'ADBS le 5 Décembre 1974 et au Séminaire Simon le 3 Mars 1975.

INDEXATION AUTOMATIQUE, CONSTRUCTION AUTOMATIQUE DES THESAURUS ET CLASSIFICATION AUTOMATIQUE

I - INTRODUCTION

Nous nous proposons de discuter les problèmes de l'indexation automatique, de la construction automatique de thésaurus et de la classification automatique, obtenus par traitement automatique du texte intégral, résumés ou documents divers.

Les méthodes permettant de traiter ces problèmes sont assez complexes et sont étudiées depuis plusieurs années par des équipes travaillant dans des pays géographiquement très éloignés (USA, URSS, Angleterre, France...). Nous ferons en particulier référence à des travaux effectués par une équipe d'informaticiens soviétiques, dont les résultats en partie se recoupent avec les nôtres et pour le reste sont complémentaires des nôtres, mais qui disposent d'un corpus documentaire considérable - 800 000 documents d'électrotechnique de 300 à 1000 mots - ce qui représente un support statistique très solide aux théories qu'ils ont élaborées [1].

Dans ce qui suit, nous n'utiliserons pas la terminologie classique de la documentation étant donné qu'en indexation et classification automatiques les problèmes sont dès le départ pensés différemment. Nous montrerons en particulier, qu'indexation et classification automatiques des documents ont un caractère dual. C'est pourquoi nous en faisons un exposé parallèle.

---

1) Institut Inform-Electro (MOSCOU).

Les méthodes d'indexation et de classification automatiques font appel à des procédés ensemblistes, statistiques et linguistiques, dont nous allons essayer de mettre en évidence les traits communs étant entendu que les analyses linguistiques bien que conduites d'une manière analogue présentent nécessairement des différences d'une langue à l'autre.

Notons pour commencer une attitude commune à toutes les études de ce type. Non seulement on s'efforce de tirer par programme le maximum d'informations du traitement des textes et résumés mais encore on modifie automatiquement les algorithmes en cours de traitement en utilisant des méthodes dites d'apprentissage. Précisons que l'effort manuel fait dans le domaine de l'indexation, loin d'être inutile, permet d'effectuer des comparaisons entre qualités des indexations manuelles et automatiques.

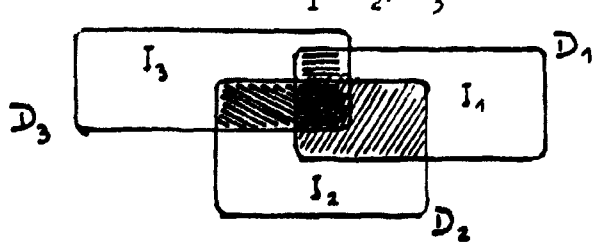
II - L'INDEXATION AUTOMATIQUE

II.1 - Définition.

La notion d'index en tant que fiche d'identité différentielle d'un document peut être définie en termes ensemblistes et statistiques.

- Enoncé ensembliste : Dans cette approche, on considère que l'index est constitué par le "complémentaire dans chaque document, de la réunion des intersections prises deux à deux".

Par exemple, sur le schéma ci-dessous I_1, I_2, I_3 sont respectivement les index des documents D_1, D_2, D_3.



La "fiche d'identité ou index" qui n'est qu'un représentant partiel du contenu sémantique du document, peut cependant suffire à identifier un document à partir d'une question documentaire et nous verrons un peu plus loin comment on s'efforce d'optimiser cette identification.

Pour pouvoir effectuer des opérations ensemblistes, on doit s'entendre sur ce qu'il conviendra d'appeler "éléments" des ensembles-documents. Ces éléments sont les concepts sémantiques quantifiés qui peuvent avoir plusieurs représentations différentes, ce qui implique une analyse morpho-syntaxique et morpho-sémantique approfondie permettant entre autres d'identifier les homographes (une représentation pour plusieurs sens), les synonymes (plusieurs représentations pour un seul sens) les éléments d'un même paradigme, la reconnaissance des référents des divers pronoms etc...(*). En pratique, la composition de l'index, représente un compromis entre le document tout entier et le minimum strictement discriminant. Plus l'index est pauvre et plus il devient difficile d'avoir une réponse correcte à des questions trop concises.

L'indexation automatique permettant facilement la réindexation des documents, on peut par approximations successives optimiser

---

(*) Dans ce travail seules sont décrites les méthodes d'identification des concepts qui font appel aux techniques de l'indexation automatique (voir plus loin II.2.3 homographes, synonymes). La reconnaissance des autres concepts fait l'objet de publications ultérieures.

la longueur des index et parvenir à un compromis entre le gain de place et une qualité de réponse documentaire adaptée aux besoins de l'utilisateur.

Précisons que ce gain de place peut selon les méthodes être de 10 à 100 (texte de 1000 mots représenté par un index de 10 à 100 mots).

- Enoncé statistique: Ce dernier reprend en fait particlement la formulation ensembliste qu'il pondère par des fréquences qui sont évaluées sur des concepts qui là encore peuvent avoir plusieurs représentations différentes.

La formulation statistique peut s'énoncer comme suit : L'intensité P avec laquelle un document D_i est caractérisé par un concept C_K (ou un ensemble de concepts) est d'autant plus grande que C_K est fréquent dans D_i et rare dans $D_j \neq D_i$, ce qui peut s'écrire

$$P(D_i \text{ si } C_K) = \frac{P(C_K \text{ dans } D_i)}{\sum_{J=1}^{J=N} P(C_K \text{ dans } D_j)} \qquad (A)$$

où N est le nombre de documents du corpus documentaire.

A peu de choses près l'expression écrite représente la formule de Bayes(*) couramment utilisée en théorie de la décision (c'est la formule de Bayes dans le cas particulier où tous les documents sont distincts). En apparence simple, elle fait surgir cependant toute une cascade de problèmes.

(*) voir dans l'annexe 1, l'exposé plus rigoureux de la théorie justifiant l'emploi de cette formule et voir aussi la bibliographie [1], [3], [4], [5].

A - Tout d'abord, les P (documents si concept) étant calculées, on doit définir les paramètres qui permettront de dire si un concept donné doit être conservé dans un index ou non. Cela peut se faire par exemple en comparant P (D_i si C_K) à sa valeur moyenne calculée sur tous les D_j c'est à dire :

$$\overline{P_K} = \frac{1}{K} \cdot \sum_{J=1}^{J=N} P(D_j \text{ si } C_K) \qquad (B)$$

On peut décider ensuite que si $P(D_i \text{ si } C_K)$ est plus grand que $\overline{P_K}$, alors C_K peut figurer dans l'index de D_i. (voir annexe 1 et bibliographie). Ceci est en accord avec les principes fondamentaux de la théorie de l'information qui proclame qu'un concept apporte d'autant plus d'informations qu'il se rencontre rarement dans le corpus documentaire. Toutefois ce principe ne peut être accepté sans réserves. Si dans un texte d'électrotechnique on trouve une expression du type "ce nonobstant le courant passera tout de même", on peut affirmer que "ce nonobstant" qui est fortement discriminant et peu répandu (en électrotechnique) fourni en principe une information importante mais qui n'est pas qualitativement pertinente dans la mesure où elle concerne le style de l'auteur et non le thème qui nous intéresse. Pour surmonter ce type de difficultés, on utilise d'une manière (explicite ou implicite) des fonctions de poids dites sémantiques qui multiplient la présence des concepts par un coefficient positif ou nul évalué à l'aide d'une théorie sémantique qui précise la notion "d'importance sémantique des concepts".

On peut en particulier distinguer deux types de fonctions de poids sémantiques :

- fonctions de poids calculées à partir d'une analyse gramma-

ticale et qui peuvent être par exemple du type

$f(C_K) = 0$ si C_K = adverbe/préposition/article/...

$f(C_K) = 1$ si C_K = substantif/adjectif/verbe/...

- fonctions de poids associées au "degré de précision des concepts" et qui peuvent être obtenues automatiquement de plusieurs façons, en particulier à l'aide de la fonction d'entropie (voir plus loin II-3).

B - La présence dans la formule (A) de N = nombre de documents, et la comparaison de P (D_i si concept) à P (D_j si concept) pour tout J pose des problèmes de gestion très complexes et représente une des grandes difficultés des systèmes d'indexation automatique. En effet, si le système est automatique, il ne peut construire les index et le thésaurus qu'à l'aide de méthodes comparatives qui brassent l'ensemble des documents du corpus documentaire. Cela ne peut être accepté que si la procédure de comparaison utilise des paramètres qui au bout d'un certain temps deviennent stationnaires. Des considération linguistiques permettent de penser que dès que le corpus documentaire devient représentatif de la diversité de la langue traitée, les grandeurs que nous avons choisies et qui permettent d'effectuer la sélection des concepts deviennent stationnaires. Ce point est particulièrement important dans la mesure où tout le problème de la gestion d'un système documentaire : indexation des nouveaux documents, réindexation des anciens, remise à jour du thésaurus (voir plus loin) est lié à la stationnarité de ces grandeurs. La propriété de stationnarité s'applique aussi aux néologismes. Si ces derniers correspondent à des concepts nouveaux importants, l'intérêt qu'on leur porte suit une courbe ascendante en début d'utilisation et qui devient en péri-

cipe rapidement stationnaire. Il convient à ce propos de remarquer que l'évolution d'une discipline scientifique quelconque n'est pas directement corrélée à l'apparition d'un certain taux de néologismes. En effet, à partir d'un certain stade d'encombrement terminologique il arrive très souvent qu'on évite d'introduire des mots nouveaux et cela se produit même quand on a affaire à une discipline dont le développement est impétueux. Par ailleurs, le néologisme représente souvent un moyen de pallier au manque d'inspiration scientifique (ésotérisme).

Indiquons encore que dans certaines disciplines comme la chimie, la médecine, la règle de formation de nombreux néologismes est précise et dans ce cas leur importance informationnelle peut s'évaluer directement à l'aide d'une analyse morpho-syntaxique.

II.2 - La construction automatique des thésaurus.

Une indexation automatique optimisée implique la construction automatique d'un thésaurus extensible ce qui comprend les procédures distinctes suivantes :

II.2.1 - Analyse morphologique.

L'analyse morphologique effectue la reconnaissance des mots et leur stockage optimisé[1] ; ensuite, en formulant un certain nombre d'hypothèses sur la nature grammaticale et sémantique des mots,

(1) "optimisé" signifie par exemple que toutes les formes conjuguées d'un verbe ne sont pas stockées mais générées par programme à partir d'une racine (lorsque le verbe est régulier). Même chose pour les substantifs, adjectifs, etc...

elle permet de restreindre le champ d'investigation des analyseurs syntaxiques et sémantiques. Dans certains cas un découpage linguistiquement pertinent en racine plus terminaison peut représenter une information importante dans la mesure où les mots de même racine ont de nombreuses parentés sémantiques.

Pour parvenir à un tel découpage, on enregistre par exemple une liste de terminaisons en indiquant explicitement les mots qui font exception à la règle de décomposition générale.

La liste des terminaisons peut être constituée soit manuellement, soit automatiquement en utilisant des méthodes de tri statistique. Pour cela on peut par exemple procéder comme suit : on suppose les mots rangés par ordre alphabétique et on prend à partir du début d'un mot la chaine de lettres la plus longue commune à au moins R mots et à au plus S mots où R et S sont des paramètres que l'on rajuste en cours d'expérience. On obtient ainsi une liste de mots ayant en commun une chaîne début ou base B. On choisit dans cette liste n'importe quel mot de la forme B+T et on regarde si T existe pour un grand nombre de mots du dictionnaire et qui seront de la forme B_1+T, B_2+T, B_x+T ... On vérifie ensuite qu'en moyenne le nombre de mots E ayant à chaque fois une base commune B_i est compris entre R et S. La liste des désinences linguistiquement pertinentes étant constituée d'une façon ou d'une autre, on peut associer à chaque désinence une suite d'informations qui précisent les cas particuliers, les valeurs sémantiques possibles, les valeurs grammaticales, (*) etc...

(*) La résolution grammaticale constitue une étape très importante et est largement décrite dans les travaux [1], [2], [3] et [5]

II.2.2 - Constitution du lexique des concepts mots sémantiquement pertinents.

Pour cela on crée un dictionnaire de mots vides. Ou bien cette liste de mots vides est constituée manuellement, ou bien elle est obtenue automatiquement par comparaison des textes scientifiques à des textes littéraires, ou bien enfin, on fait la réunion des index obtenus à l'aide d'une sélection grammaticale complétée par l'emploi des formules statistiques (A) et (B).

Dans tous les cas on a affaire à des méthodes de contraste :

- contraste textes scientifiques - textes littéraires
- contrastes interne au corpus documentaire même.

II.2.3 - Constitution automatique des liaisons sémantiques.

Comme nous l'avons déjà signalé, l'identification des concepts pose en analyse automatique du contenu deux problèmes symétriques importants :

- identification des homographes : une seule représentation pour plusieurs concepts
- identification des synonymes : plusieurs représentations pour un seul concept.

Sans ces deux identifications on ne peut faire aucun comptage correct des concepts et on ne peut obtenir de réponses correctes à des questions documentaires formulées en langue naturelle.

Pour réaliser ces identifications d'une manière entièrement automatisée, on doit faire l'hypothèse que les documents stockés représentent des entités sémantiques de profil nettement défini et que par conséquent le complémentaire d'un mot M dans un document ou dans

l'index(*) de ce document, représente une partie du champ sémantique d'un des concepts associé au mot M (en général un mot est homographe). C'est l'hypothèse qui a été utilisée par les spécialistes soviétiques déjà cités et par nous-mêmes.

Le problème difficile qu'il y a à résoudre dans ce cas est de déterminer le nombre de champs sémantiques partiels du concept C dont la réunion est suffisamment représentative du champ sémantique global de C, car le calcul du champ sémantique global ne peut se faire sur l'ensemble du corpus documentaire (ce qui serait très couteux). Aussi faut-il déterminer statistiquement la fraction du corpus documentaire sémantiquement représentative (c'est la méthode utilisée par les spécialistes soviétiques d'Inform-Electro).

Par ailleurs, il n'est pas possible de réunir les champs sémantiques des homographes de sens différents. Par conséquent la détermination des homographes doit être faite avant, ou au moins en même temps que la détermination des champs sémantiques.

- Détermination des homographes

La détermination des homographes se fait en premier lieu par une analyse grammaticale. Dans notre cas on dispose d'un programme d'analyse grammaticale, obtenu par apprentissage, (voir Document de linguistique quantitative N° 21 Edition Dunod). Toutefois la différence des fonctions grammaticales n'est pas un critère suffisant de séparation de l'homographie sémantique. En effet si pour le mot "car" l'information "substantif" ou "conjonction" est décisive et correspond aussi à une différence sémantique très nette, pour le mot

(*) Il s'agit d'un index initial construit en première approximation par simple élimination des mots vides sans tenir compte de l'homographie de la synonymie ou des relations de spécificité.

"programme" l'information "substantif ou "verbe conjugué" ne correspond pas à une homographie sémantique réelle car au fond le mot "programme" dans les deux cas assure une fonction sémantique analogue. On peut s'en rendre compte grâce à la transformation toujours possible pour les verbes de spécialité : je programme → je fais un programme. Par conséquent, l'analyse grammaticale doit être complétée par un certain nombre de procédures complémentaires permettant d'améliorer le caractère sémantique de la décision grammaticale. On peut ensuite, dans le but de résoudre l'homographie des mots ayant une catégorie grammaticale bien définie (substantif en général) utiliser plusieurs méthodes :

. première méthode (semi automatique) : (1)

On coche à la main dans les index tous les homographes de même sens. Après quoi, on réunit automatiquement les champs sémantiques partiels des dits homographes. Ce travail se fait sur une certaine fraction "pertinente" du corpus documentaire et cette fraction doit être déterminée statistiquement.

. seconde méthode (automatique) : (2)

On peut aussi tenter de séparer les homographes en évitant tout prétraitement manuel. Pour cela, l'automate choisit un mot quelconque dans un document, il le repère dans tous les index d'une certaine fraction du corpus documentaire, et il construit ses champs séman-

---

(1) cette méthode a été proposée par les spécialistes soviétiques d'Inform-Electro et par nous-mêmes. Elle a effectivement été expérimentée avec succès par les spécialistes d'Inform-Electro.

(2) proposée par les spécialistes d'Inform-Electro et nous-mêmes les premières expériences effectuées par les spécialistes d'Inform-Electro ne permettent pas encore de conclure.

tiques partiels S_i. On réunit ensuite deux champs partiels S_i et S_j si $S_i \cap S_j$ est suffisamment important par rapport à $S_i \cup S_j$.

L'expression :

$$T_{ij} = \left| \frac{\text{card } S_i \cap S_j}{\text{card } S_i \cup S_j} - 1 \right|$$

permet d'évaluer l'importance de l'intersection par rapport à la réunion. On l'utilise en définissant un seuil T_o tel que si $T_{ij} < T_o$, on réunit S_i et S_j.

T_o doit être déterminé statistiquement et ce de telle manière que tous les homographes soient trouvés et qu'on n'en trouve pas plus qu'il n'en existe en réalité. On peut aussi essayer de déterminer T_o d'une façon naturelle en cherchant les "noyaux" des champs sémantiques de l'homographe, c'est à dire les "accumulations" des champs sémantiques partiels.

Ce dernier point pose des problèmes combinatoires importants. Une bonne approche de ce problème consiste à supposer que les champs sémantiques partiels S_i dont les cardinaux sont les plus grands occupent un rôle privilégié dans les noyaux à condition de trier les champs S_i pour lesquels on a $S_i \cap S_j \sim \phi$.

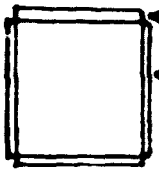
- Reconnaissance des synonymes

Les concepts correspondants aux différents homographes étant reconnus, et leurs champs sémantiques déterminés, on peut procéder à la reconnaissance des synonymes, des relations "générique-spécifique" et plus généralement des diverses relations de parenté sémantique.

La méthode élaborée(*) peut être exposée simplement à l'aide des schémas suivants :

(*) proposée et expérimentée avec succès par les spécialistes d'Inform-Electro.

- synonymie



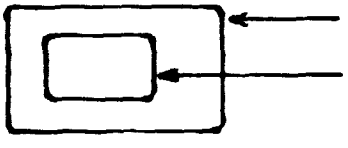
champ sémantique S_i du concept C_i

champ sémantique S_j du concept C_j

Si $S_i \cap S_j \sim S_i \sim S_j$ alors $C_i \sim C_j$

et on a une relation de synonymie.

- spécificité-généricité



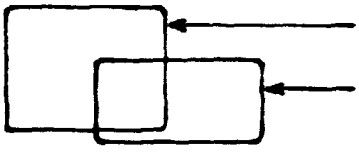
Champ sémantique S_i du concept C_i

Champ sémantique S_j du concept C_j

Si $S_i \cap S_j = S_j$ et donc $S_j \subset S_i$,

alors C_j est terme spécifique de C_i

- parenté sémantique



Champ sémantique S_i du concept C_i

Champ sémantique S_j du concept C_j

C_i et C_j sont dans une relation de parenté

car $S_i \cap S_j \neq \emptyset$.

Les trois cas que nous venons de décrire peuvent être discutés d'une façon plus rigoureuse à l'aide par exemple de la relation de Takimoto :

$$T_{ij} = \left| \frac{\text{card } (S_i \cap S_j)}{\text{card } (S_i \cup S_j)} - 1 \right|$$

et en introduisant un seuil T_o de décision déterminé expérimentalement :

si $T_{ij} < T_o \Rightarrow$ synonymie

$T_{ij} > T_o \Rightarrow$ parenté

La spécificité-généricité étant reconnue à l'aide de la double relation

$$S_j \subset S_i \quad \text{et} \quad T_{ij} > T_o.$$

Une méthode analogue peut encore être utilisée dans la reconnaissance des locutions. Soient en effet deux mots M_1 et M_2 tels que leur apparition consécutive M_1, M_2 soit assez fréquente dans les documents.

Soit C_1 le concept relatif à M_1 et S_1 le champ sémantique de M_1

Soit C_2 le concept relatif à M_2 et S_2 le champ sémantique de M_2

Soit S_{12} le champ sémantique de M_1 M_2.

Si $S_1 \cup S_2 \simeq S_{12}$, alors on peut dire que M_1 M_2 n'est pas une locution et il n'y a pas de concept C_{12}.

Si $S_1 \cup S_2$ est très différent de S_{12}, alors on peut dire que M_1 M_2 est une locution et il y a un concept C_{12}.

Là encore la comparaison de $S_1 \cup S_2$ et de S_{12} peut se faire à l'aide de l'expression :

$$T_{1,2,12} = \left| \frac{\text{Card } (S_1 \cup S_2) \cap S_{12}}{\text{Card } (S_1 \cup S_2) \cup S_{12}} - 1 \right|$$

pour laquelle on doit déterminer statistiquement un seuil T_o.

II.3 - Les diverses manières d'obtenir les fonctions de poids sémantique. (*)

En principe, la fonction de poids sémantique doit être d'autant plus grande que le mot est spécifique dans la langue de spécialité étudiée. Lorsqu'on dispose d'un dictionnaire de spécialité, on peut supposer que la spécificité est d'autant plus faible qu'un mot participe directement ou indirectement à la définition des autres mots. La spécificité est maximale si un mot ne participe à la définition d'aucun autre mot. Cette définition de la spécificité

(*) pour ce paragraphe voir aussi la bibliographie (3), (4), (5).

permet d'envisager un certain nombre de procédures pratiques pour le calcul des fonctions de poids sémantiques. En principe, ces procédures sont équivalentes mais en fait elles peuvent donner des résultats très différents.

1) Si le dictionnaire de spécialité dont on dispose est sur support informatique, les fonctions de poids d'après ce que l'on vient de voir peuvent être calculées automatiquement à partir du dictionnaire.

2) Si l'on dispose d'un thésaurus sur support informatique, on peut prendre pour fonction de poids sémantique la fonction $f_i = m - n_i + 1$ où n_i est le nombre de termes spécifiques du mot M_i ou du concept C_i, m est le sup. de n_i. On voit que si $n_i = m$, f_i est minimale et égale à 1.

3) A l'aide de la fonction $H(C_i) = -\sum_{J=1}^{J=N} P(D_j/C_i)\log_2(D_j/c_i)$

dont le sens est facile à comprendre en donnant à $H(C_i)$ certaines valeurs limites : si C_K ne figure que dans un seul document, $H(C_K) = 0$ et si C_K est uniformément réparti sur l'ensemble des documents $H(C_K) = \log N$.

Toutefois l'utilisation de la fonction $H(C_i)$ doit se faire en prenant certaines précautions étant donné qu'elle peut être très sensible à une reconnaissance correcte du concept C_i. En effet, on peut par exemple dans un corpus donné calculer H(action), valeur à laquelle il faudra adjoindre dans certains cas H(réaction). Ici l'identification du concept doit être poussée très loin.

4) Enfin on peut utiliser les champs sémantiques définis précédemment et prendre comme fonction de poids sémantique f_i du concept C_i l'expression

$$f_i = m - \text{card } S_i + 1$$

où S_i est le champ sémantique de C_i, et m le cardinal du champ sémantique le plus grand.

Il est clair qu'entre toutes ces méthodes il y a une "équivalence" théorique mais qu'en pratique elles peuvent donner des résultats différents. Seule une étude linguistique et statistique poussée peut permettre de rajuster ces diverses méthodes.

III - INDEXATION, QUESTION DOCUMENTAIRE, CLASSIFICATION AUTOMATIQUE.

Le thésaurus précédemment décrit permet d'identifier les concepts : séparation des homographes et recherche des synonymes et de dresser la liste de l'ensemble des concepts utilisés dans l'indexation. Cette opération étant supposée faite, on peut constituer le tableau des correspondances documents-concepts que l'on peut représenter de la manière suivante :

concepts / documents	C_1	C_2	C_3	C_5
D_1	$P(C_1/D_1)$	$P(C_2/D_1)$	$P(C_3/D_1)$	$P(C_4/D_1)$
D_2	$P(C_1/D_2)$	$P(C_2/D_2)$	...	...
D_3	...			
D_4	...			

Les $P(C_i/D_j)$ sont les fréquences pondérées des C_i dans les D_j. L'indexation, la question documentaire, et la classification automatique peuvent être obtenues comme résultats d'algorithmes différents opérant sur le tableau ci-dessus avec cependant un certain nombre de différences importantes que nous allons indiquer au fur et à mesure.

. L'index d'un D_j est constitué de C_i tels que $P(C_i/D_j) \geqslant$ à une certaine valeur moyenne V_i, par exemple :

$$V_i = \frac{1}{N^o} \sum_{K=1}^{K=N} P(C_i/D_K) \qquad \text{(voir annexe 1)}$$

Cependant, les C_i sont constitués (au moins dans notre traitement) en même temps que les index. De plus, il s'agit d'index optimisés utilisés aussi pour la question documentaire, mais pas nécessairement pour la classification automatique. Pour cette dernière, on peut prendre des index obtenus par exemple par simple sélection linguistique (grammaticale, morphologique, sémantique).

. La question documentaire, est un ensemble Q de concepts C_i donnés(*) Q peut être considéré comme un document "privilégié" et on doit rechercher tous les D_i tels que $Q \cap D_i \neq \emptyset$ et ranger les D_i par ordre de "proximité décroissante" ce qui rend indispensable la définition d'une "proximité". Il est clair qu'à l'aide d'un ensemble de questions documentaires bien choisies, on peut constituer une classification qui toutefois peut être différente de la classification naturelle induite par le stock documentaire.

. La classification naturelle, cherche en quelque sorte les documents pôles autour desquels le stock documentaire s'agglutine. Soulignons le encore une fois, il peut y avoir intérêt pour la classification automatique à constituer des index plus larges que pour l'indexation.

Pour la question documentaire ou pour la classification, nous utiliserons des distances et des proximités qui s'apparentent à l'une des expression suivantes :

(*) Les concepts C_i peuvent être reliés par des relations semantiques, logiques ou syntaxiques.

A - Distance.

. $\frac{\text{card } D_i \cap D_j}{\text{card } D_i \cup D_j}$ (importance relative de l'intersection par rapport à la réunion).

B - Proximités.

. card $D_i \cap D_j$ (importance absolue de l'intersection)

. $\frac{\text{card } D_i \cap D_j}{\text{card } D_j}$ (importance relative du cardinal de l'intersection au cardinal de l'un des ensembles choisi comme "prioritaire")

. P(D_i si concepts de $D_i \cap D_j$) ici encore l'un des documents (D_j) est prioritaire.

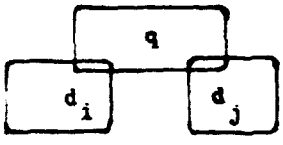
La symétrisation des "proximités" permet évidemment de définir des distances.

III.1 - La question documentaire.

Pour comparer une question q aux documents stockés et pour trouver les documents sémantiquement les plus proches de q, on peut utiliser des grandeurs non-symétriques qui donnent à q un rôle prioritaire. Par exemple : $\frac{\text{card } d_i \cap q}{\text{card } q}$. L'emploi d'une expression de ce type permet de sélectionner les documents d_i qui ont une partie commune avec q relativement importante par rapport à q et quelconque par rapport à d_i. Cependant, afin de tenir compte des fonctions de poids sémantique et des fréquences, on doit recourir à des expressions plus complexes que nous allons décrire maintenant. Il est entendu cependant que la recherche de tous les documents ayant un concept commun avec q (c'est à dire recherche des d_i tels que card $d_i \cap q > 0$) constitue une première présélection indispensable. Après quoi on peut procéder de plusieurs façons différentes mais dans ce paragraphe nous ne décrivons qu'une seule méthode basée sur la formule de Bayes modifiée.

D'autres méthodes sont décrites dans le paragraphe suivant consacré à la classification.

Afin de rendre possible la comparaison des proximités de q, d_i et de q, d_j lorsque $d_i \cap d_j = \not\emptyset$, c'est à dire lorsqu'on est en présence du cas représenté par le schéma suivant,



$d_i \cap d_j = \emptyset$

$q \cap d_i \neq \emptyset$ et $q \cap d_j \neq \emptyset$

on modifie les documents d_i en les complétant par tous les concepts figurant dans les intersections $q \cap d_j$ et ne figurant pas dans $q \cap d_i$, et en affectant à ces concepts des fonctions de poids sémantiques très faibles que l'on détermine en relation avec l'expérience. On obtient ainsi des documents d_i^* obtenus à partir des documents d_i. Les d_i peuvent encore être obtenus d'une autre façon en complétant les d_i par tous les concepts de q n'appartenant pas à $d_i \cap q$, affectés là encore d'une fonction de poids sémantique faible. Dans ce cas on a $d_i^* \cap q \equiv q$, $\forall_i$ (les pondérations de q sont différentes suivant i). Dans les deux cas, si l'on appelle $< d_i^* \cap q >$ le produit probabiliste des concepts communs à d_i et q on peut utiliser la formule de Bayes modifiée :

$$P(d_i^*/<d_i^* \cap q>) = \frac{P(d_i^*).P(<d_i^* \cap q>/d_i).(P_{min})^{card\left[d_i^* - d_i\right]}}{\sum_{J=1}^{J=N} P(d_j^*).P(<d_j^* \cap q>/d_j)(P_{min})^{card\left[d_j^* - d_j\right]}}$$

où l'on a remplacé $P(d_i^* \cap q / d_i^*)$ par

$$P(d_i^* \cap q / d_i).(P_{min})^{card\left[d_i^* - d_i\right]}$$

et où $P(d_i) = P(d_i^*) = \frac{1}{N}$

Il est clair que $p(d_i^*/<d_i^* \cap q>)$ n'est pas une distance puisqu'elle n'est pas symétrique.

Afin de pouvoir dire jusqu'à quel point le document qui répond le mieux à la question y répond effectivement, on peut intégrer la question au système documentaire et évaluer $P(q/d_i^* \cap q)$ à l'aide de la formule :

$$P(q/<d_i^* \cap q>) \frac{P(<d_i^* \cap q>/q)}{\sum_{J=1}^{J=N} P(d_i^* \cap q/d_j) + P(d_i^* \cap q/q)}$$

où l'on a posé $P(q) = P(d_j) = \frac{1}{N+1}$.

Cette formule reste valable si $<d_i^* \cap q> \equiv q$.

Remarque :

L'expression de $P(q/<q>)$ peut en principe être inférieure à celle de $P(d_i^*/<q>)$ pour certains i. En effet, les poids et les fréquences des concepts de $d_i^* \cap q \equiv q$ ne sont pas les mêmes dans d_i^* et dans q.

III.2 - Classification automatique.

Les méthodes que nous allons décrire supposent toujours que l'on dispose d'un thésaurus permettant d'identifier les concepts.

Première méthode. Si l'on cherche les documents qui ont simplement en commun au moins un certain profil sémantique S, alors le calcul de card $D_1 \cap D_2$ peut suffire, avec comme condition $S \subset D_1 \cap D_2$. Si l'on veut tenir compte de l'importance relative de S par rapport à D_1 et D_2, alors on peut prendre la distance de Takimoto sous sa

forme directe[1] :

$$T(D_1, D_2) = \frac{\text{card } D_1 \cap D_2}{\text{card } D_1 \cup D_2}$$

ou sous sa forme complémentaire

$$T^*(D_1, D_2) = \frac{\text{card } D_1 \cup D_2 - \text{card } D_1 \cap D_2}{\text{card } D_1 \cup D_2}$$

($T^* + T = 1$)

Seconde méthode. La seconde méthode généralise la première en introduisant des fonctions de poids sémantiques.

Soit $f(i/j)$ les fréquences pondérées par des fonctions de poids sémantiques des concepts i dans les documents j, et N le nombre total de concepts utilisés dans le corpus documentaire.

La généralisation du card $D_1 \cap D_2$ s'obtient à l'aide de la fonction[2] :

$$V(D_1, D_2) = \sum_{i=1}^{i=N} \left| f(i/1) - f(i/2) \right|$$

(on a $V(D_1, D_2) = V(D_2, D_1)$).

En fait, la somme écrite n'est réellement étendue qu'à l'ensemble des concepts figurant dans $D_1 \cup D_2$.

Par ailleurs, la connaissance de N-card D_1, N-card D_2 permet d'avoir parallèlement une idée du degré de généralité du document.

Le sens de $V(D_1, D_2)$ peut être explicité à l'aide du tableau suivant :

concepts / documents	C_1	C_2	C_3
D_1	$f(1/1)$	$f(2/1)$	$f(3/1)$
D_2	$f(1/2)$	$f(2/2)$	$f(3/2)$

(1) expérimentée par les spécialistes d'Inform-Electro.

(2) proposée par les spécialistes d'Inform-Electro et nous-mêmes.

$V(D_1, D_2)$ représente la somme des modules de la différence colonne par colonne des lignes D_1 et D_2. Lorsque les concepts sont comptés d'une manière purement ensembliste, alors $f(i/j) = 1$ si $C_i \in D_j$ et $f(i/j) = 0$ si $C_i \notin D_j$, auquel cas $V(D_1, D_2) =$ aire $D_1 \cap D_2$, et par conséquent $V(D_1, D_2)$ permet de saisir la proximité entre documents qui ont <u>au moins</u> certains concepts en commun, en tenant compte des pondérations sémantiques.

La distance $V(D_1, D_2)$ définie plus haut ne tient pas compte de l'importance relative de $D_1 \cap D_2$ par rapport à $D_1 \cup D_2$. On peut en tenir compte à l'aide de la fonction[1] :

$$V(D_i, D_j) = \frac{\sum_{K=1}^{K=N} \left| f(K/i) - f(K/j) \right|}{\sum_{K=1}^{K=N} \left[f(K/i) + f(K/j) \right] \left[\delta_{f(K/i), f(K/j)} + 1 \right]^{-1}}$$

où $f_{ki} \cdot f_{kj} = 1$ si $f_{ki} \cdot f_{kj} \neq 0$

$f_{ki} \cdot f_{kj} = 0$ si $f_{ki} \cdot f_{kj} = 0$

Enfin l'importance relative de $D_1 \cap D_2$ par rapport à D_1 seul ou par rapport à D_2 seul, peut être évaluée à l'aide de la formule :

$$V(D_i, D_j) = \frac{\sum_{K=1}^{K=N} \left| f(K/i) - f(K/j) \right|}{\sum_{K=1}^{K=N} f(K/i)}$$

---

(1) *Cette fonction a été définie par les spécialistes de l'Institut Inform-Electro et nous mêmes.*

Il est clair qu'on peut aussi utiliser l'expression de $\left|\sum d_i(<d_i \cap d_j>)-1\right|$ discutée à propos de la question documentaire. La symétrisation de cette formule conduit à une expression assez lourde.

III.3 - Utilisation pratique des distances et proximités ainsi définies.

En pratique, les distances et les proximités définies, peuvent être utilisées de plusieurs manières différentes. La plus simple, et qui a été effectivement expérimentée (Inform-Electro), consiste en premier lieu à se donner un seuil ρ_o arbitraire. Après quoi, si l'on appelle $\rho(D_1,D_2)$ la distance ou la proximité définie, on pose que la condition $\rho(D_1,D_2) \leqslant \rho_o$ implique que D_1 et D_2 appartiennent à la même classe. A chaque ρ_o donné correspond une certaine classification $C(\rho_o)$ telle que $C(\rho_o)$ à la forme : $C(\rho_o) = \bigcup_i C_i(\rho_o)$ avec $C_i(\rho_o) \cap C_j(\rho_o) = \emptyset$ et telle que $\forall D_j, \exists C_K(\rho_o)$ telle que $D_j \subset C_K(\rho_o)$.

Par ailleurs un document peut toujours appartenir à plusieurs classes relatives à des classifications différentes.

La constitution effective d'une classe $C(\rho_o)$ peut se faire de la façon suivante : on part d'un document quelconque D_i et on cherche tous les documents D_j tels que $\rho(D_i,D_j) \leqslant \rho_o$. Puis ensuite pour chaque D_j tous les $D_{k(j)}$ tels que $\rho(D_j, D_{k(j)}) \leqslant \rho_o$ et ainsi de suite jusqu'à arrêt du processus (ou épuisement du stock documentaire).

On peut procéder autrement en cherchant les noyaux (ou

accumulations) du stock documentaire. Pour cela on peut considérer que les documents dont les index ont le cardinal le plus élevé occupent une place importante dans les noyaux. On pourra donc prendre comme "représentant" des noyaux un ensemble $K(D)$ de documents, de cardinal élevé, et tel que si D_j, $D_K \in K(D)$, $D_j \cap D_K \sim \phi$

En d'autres termes on devra choisir ρ_o (statistiquement) de telle façon que lorsque $\rho(D_j, D_K) \geqslant \rho_o$ l'intersection $D_j \cap D_K$ contienne un nombre "insignifiant" de concepts (pondération sémantique comprise).

Les "représentants" supposés des noyaux sont ensuite utilisés comme documents de "départ" pour construire la classification.

Enfin, à l'intérieur de chaque classe on peut vérifier si le "représentant" choisi est aussi le document-barycentre de la classe, et cela en calculant l'expression

$$\min_i \sum_{J=1}^{J=N_p} \rho(D_j, D_j)$$

où N_p est le nombre de documents d'une classe p de la classification $C(\rho_o)$. Mais si N_p est trop grand la recherche du document barycentre devient trop coûteuse.

IV - CONCLUSION.

Compte tenu des expériences que nous avons faites et des expériences que nous avons pu voir réalisées à l'Institut Inform-Electro sur un corpus considérable, on peut formuler les conclusions suivantes :

1 - L'indexation automatique est plus souple que l'indexation manuelle. Elle permet en particulier de porter dans l'index des in-

formations qui à l'indexeur peuvent à un moment donné ne pas paraître pertinentes mais qui jouent un rôle contexte important. Elle est plus rapide que l'indexation manuelle, moins coûteuse et elle simplifie considérablement la remise à jour (réindexation) des documents.

Elle est plus objective en ce sens que l'algorithme d'indexation automatique fournit toujours les mêmes index dans un stock documentaire donné, alors que l'indexation manuelle peut conduire à des fluctuations dans la composition des index élaborés à des moments différents. Enfin elle donne des réponses aux questions documentaires qui testées par un lot important d'utilisateurs ont été déclarées dans la plupart des cas meilleures que les réponses aux systèmes indexés manuellement.

2 - La classification automatique permet d'obtenir les divers "pôles sémantiques" du stock documentaire. Outre qu'elle permet de voir les tendances générales qui se dégagent dans la spécialité étudiée, elle peut aussi être utilisée pour améliorer la recherche documentaire.

Pour finir indiquons encore que la conception et l'élaboration des systèmes d'indexation et classification automatique se heurtent aux difficultés inhérentes à la construction de tous les "grands systèmes" : modularité, transportabilité d'un ordinateur à l'autre, existence d'équipes multidisciplinaires (informaticiens, mathématiciens, linguistes), et enfin un système de saisie des données adaptées. A ce propos il convient de remarquer tout d'abord, que la plupart des centres documentaires importants possèdent des résumés sur supports informatiques, que de nombreux textes, livres, documents, revues, sont mis sur bandes magnétiques par les éditeurs,

et qu'enfin, la mise au point de lecteurs optiques permettra d'accélérer considérablement l'enregistrement des textes et rendra l'indexation automatique beaucoup plus accessible.

Appendice : Annexe 1

L'utilisation de la formule de Bayes

$$P(d_i/C_K) = \frac{P(C_K/d_i).P(d_i)}{\sum_{J=1}^{J=N} P(C_K/d_j).P(d_j)}$$

avec $P(d_i) = \frac{1}{N}$ si tous les documents sont différents.

suppose d'une part que l'on assimile le stock documentaire à une urne de documents et chaque document à une urne de concepts, et d'autre part cela suppose que l'on fait des tirages de documents et des tirages de concepts.

Rappelons que les évènements "tirages des d_i" doivent être incompatibles de même que les évènements "tirages des C_K".

L'hypothèse d'incompatibilité des évènements, implique toujours leur dépendance : $P(d_i.d_j) = 0$, $P(d_i) \neq 0$, $P(d_j) \neq 0$ implique que $P(d_i).P(d_i/d_j) = 0$ et donc que $P(d_i) \neq P(d_i/d_j) = 0$

Mais la dépendance ainsi réalisée dans le cas présent est définie par la nature de l'expérience de tirage effectuée (apparition de d_i exclue celle de tout d_j, $j \neq i$) et elle implique une certaine indépendance sémantique des documents : on doit être assuré que le tirage d'un d_i n'entraine pas sans qu'on le sache, le tirage d'un d_j, $j \neq i$. Cela doit se produire en particulier si on a stocké deux fois le même document (à quelques périphrases près). Sinon il n'est plus possible de poser $P(d_i) = \frac{1}{N}$ et on doit en fait écrire $P(d_i) = \frac{2}{N}, \frac{3}{N}, \ldots$ le numérateur représentant le nombre des d_i ou des presques d_i dans le stock documentaire.

Cette remarque vaut pour les concepts C_K pour lesquels

l'indépendance sémantique est encore plus difficile à garantir.

Toutefois ceci ne représente pas pour l'application que nous envisageons, un inconvénient majeur.

En effet, pour les documents, la répétition d'un document ou de documents très voisins est évitée au moment de l'introduction des documents dans le stock documentaire et les quelques erreurs commises jouent un rôle peu important si le nombre N de documents est grand. Par contre, la dépendance sémantique des mots est un phénomène impossible à éviter. Lorsque cette dépendance est du type "synonymie" (deux représentations pour un même sens) alors dans ce cas pour deux mots synonymes M_1 et M_2 les probabilités $P(M_1/D_i)$ et $P(M_2/D_i)$ ont des valeurs inférieures à celles de $P(M_1 \vee M_2/D_i)$. Les probabilités des mots si documents sont donc inférieurs à ce qu'elles devraient être. Si la dépendance est du type "homographie" (une représentation pour plusieurs sens) alors pour tout homographe M mal reconnu, les probabilités des mots si documents $P(M/D_i)$ seront supérieures à ce qu'elles devraient être.

Dans ces conditions dans la formule :

$$P(D_i/G) = \frac{P(D_i)P(G/D_i)}{\sum_{J=1}^{J=N} P(D_j)P(G/D_j)}$$

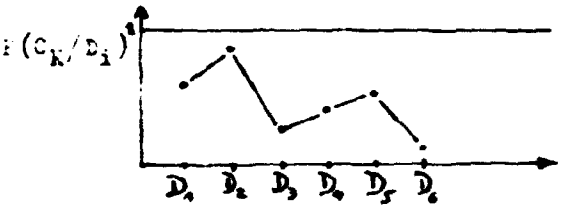
où $G = C_1, C_2 \ldots C_K$

et où par conséquent $P(G/D_i) = P(C_1/D_i).P(C_2/D_i)\ldots P(C_K/D_i)$

(voir la bibliographie [3], [4] [5])

les $P(C_K/D_i)$ sont des valeurs approchées des $P(C_k/D_i)$ réels. Il s'ensuit que les seuils permettant d'effectuer les sélections doivent être déterminés non seulement théoriquement mais expérimentalement.

La détermination des seuils se fait en considérant que les C_K les plus caractéristiques d'un D_i sont ceux pour lesquels $P(D_i/C_K)$ est le plus important ce que l'on peut voir sur le schéma suivant :



ici $\sup_i P(C_K/D_i) = p(C_K/D_2)$

donc

$C_K \in$ index de D_2

Une méthode moins sévère consiste à prendre les C_K pour lesquels $P(C_K/D_i) \geqslant$ seuil donné.

Par exemple on peut prendre cette constante égale à la moyenne des $P(C_K/D_j)$ pour tous les j. Dans ce cas, $C_K \in$ index de D_i. Si

$$P(C_K/D_i) \geqslant \frac{1}{N} \cdot \sum_{J=1}^{J=N} P(C_K/D_j) = \overline{P(C_K)}$$

où N est le nombre de documents du stock documentaire. Compte tenu du fait que $P(D_j) = \frac{1}{N}$ quelque soit j, alors dans ce cas particulier $\overline{P(C_K)} = \sum_{J=1}^{J=N} P(D_j)P(C_K/D_j) = P(C_K)$. C'est à dire $P(C_K)$ représente la fréquence du concept C_K dans le stock documentaire. Cette fréquence peut ne pas devenir stationnaire avec l'accroissement de N. Par contre la fréquence

$$\overline{P(C_K)}^{+} = \frac{1}{N^{+}} \cdot \sum_{J=1}^{J=N} P(C_K/D_j)$$

(N^{+} est le nombre de documents pour lesquels $P(C_K/D_j) > 0$, c'est à dire les documents dans lesquels C_K figure).

devient en principe stationnaire.

Les seuils ainsi définis, pour la sélection des concepts, peuvent être sensibles à la dépendance des concepts entre eux. Mais il faut se souvenir que l'index représente un compromis entre le minimum strictement discriminant et le document tout entier. Il s'ensuit

que le seuil $\overline{P(C_k)}^+$ par exemple, peut être modifié expérimentalement et multiplié par un coefficient θ positif qui s'il est inférieur à 1 permet de "récupérer" les concepts éliminés par suite d'une non-identification de certaines dépendances sémantiques. Si θ est supérieur à 1 on pourra éliminer des concepts, mais comme on ne peut à la fois à l'aide de θ "éliminer" et "récupérer", on s'efforcera en général de "récupérer" et d'obtenir ainsi des index un peu trop riches afin d'éviter le "silence" documentaire. Soulignons que la méthode de sélection des concepts que nous utilisons permet de contourner le problème de la dépendance des concepts et de venir ainsi à bout d'une des difficultés les plus importantes inhérentes à l'emploi de la formule de Bayes. On peut procéder de la même façon dans tous les problèmes de décision faisant appel au modèle bayesien.

REMARQUE :

Dans certains cas, pour éviter les répétitions, il arrive qu'on utilise un terme générique à la place d'un terme spécifique, par exemple : la maladie évolue alors normalement au lieu de : l'hépatite évolue alors normalement. En ce cas, il faut traiter le terme générique MALADIE comme un pronom et considérer qu'en fait il répète le mot HEPATITE.

Cependant la règle de décision permettant de considérer qu'un terme générique est un pronom référant à un de ses termes spécifiques peut être très complexe.

Annexe 2

En fait les champs sémantiques des synonymes et de certains antonymes risquent par cette méthode d'être très voisins. Si c'est le cas, c'est qu'à la question documentaire posée on doit effectivement associer les documents relatifs à l'antonyme.

Aussi bien pour les problèmes de synonymie que pour le calcul des fonctions de poids sémantiques, il peut être utile de calculer le degré de cooccurence. Une bonne mesure de cette cooccurence O_{ij} des concepts C_i et C_j est donnée par la formule :

$$O_{ij} = \frac{\sum_{K=1}^{K=N} \inf \left[P(C_i/d_K), P(C_j/d_K) \right]^2}{\left(\sum_{S=1}^{S=N} P(C_i/d_s) \right) \cdot \left(\sum_{r=1}^{r=N} P(C_j/d_r) \right)}$$

Le degré de cooccurence des concepts C_i et C_j est d'autant plus grand que sont équivalents (à un mot près) leur champs sémantiques :

(Dans un index I on a $S_i = C_j + S_i \cap S_j$)

Le degré de cooccurence permet en particulier de rendre plus homogène le calcul des fonctions de poids sémantiques.

BIBLIOGRAPHIE

1 - A. ANDREEWSKY — METHODES DE LA LINGUISTIQUE MATHEMATIQUE ANALYSE AUTOMATIQUE DU LANGAGE Note C.E.A. n° 1557 - Juillet 1972.

2 - A. ANDREEWSKY, C. FLUHR, A. BLOCH — EXPERIENCE DE CONSTITUTION D'UN PROGRAMME D'APPRENTISSAGE POUR LE TRAITEMENT AUTOMATIQUE DU LANGAGE Note C.E.A. n° 1606 (1) - Décembre 1972 Note C.E.A. n° 1606 (2) - Novembre 1973

3 - A. ANDREEWSKY, C. FLUHR — APPRENTISSAGE - ANALYSE AUTOMATIQUE DU LANGAGE APPLICATION A LA DOCUMENTATION DOCUMENTS DE LINGUISTIQUE QUANTITATIVE N° 21 Edition DUNOD 1973.

4 - A. ANDREEWSKY, C. FLUHR — INDEXATION AUTOMATIQUE MAINTENANCE ET GESTION D'UN SYSTEME DOCUMENTAIRE lère partie : Aspects Théoriques Note C.E.A. n° 1694 (1) - Décembre 1973

5 - A. ANDREEWSKY, C. FLUHR — INDEXATION AUTOMATIQUE DES DOCUMENTS INTERROGATION EN LANGUE NATURELLE GESTION - MAI 1974 SEMINAIRE DE L'IRIA.

Manuscrit reçu le 3 mars 1975

Edité par
le Service de Documentation
Centre d'Etudes Nucléaires de Saclay
Boîte Postale n° 2
91190 - Gif-sur-YVETTE (France)