**N° d'ordre : 33 / 2002-M / MT**


UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE HOUARI BOUMEDIENE

**FACULTE DE MATHEMATIQUES**

# THESE

**Présentée pour l'obtention du grade de** :

# MAGISTER

En **: MATHEMATIQUES**

**Spécialité** : RECHERCHE OPERATIONNELLE

**Par : KOUICI SALIMA**

**Thème**

## Réseaux de neurones artificiels et analyse en composantes principales pour l'amélioration de la recherche d'information

**Soutenue le 17/12/2002, devant le jury** :

| | | |
|---|---|---|
| Mr BERRACHEDI Abdelhafid | Maître de Conférence, USTHB | Président |
| Mr ABBAS Moncef | Professeur, USTHB | Directeur de thèse |
| M. BOUCHEMAKH Isma | Maître de Conférence, USTHB | Examinatrice |
| Mr. MAAMRA M. Saïd | Chargé de cours, USTHB | Examinateur |
| Mr MOULAI Mustapha | Chargé de cours, USTHB | Examinateur |
| Mr MOULAI Mustapha | Chargé de cours, USTHB | Examinateu |

# *Dédicaces*

***A** mon mari toujours présent pour me soutenir et mon adorable fille Hayet que j'adore,*

***A** mes très chers parents à qui je dois tout ce que je réussis,*

***A** mes frères et mes sœurs que je retrouve toujours dans les moments difficiles,*

***A** Mes beaux parents, mes beaux frères et mes belles sœurs pour leurs estime et encouragements,*

***A** Mes grands-parents pour leurs prières et meilleurs souhaits pour moi,*

***A** toutes mes amies,*

*Je dédie ce modeste travail*

# REMERCIEMENTS

J’exprime toute ma reconnaissance à Monsieur M. ABBAS professeur à l’USTHB et je le remercie vivement pour la compétence et la patiente avec lesquelles il a dirigé ce travail.

J’adresse tous mes remerciements à Monsieur A. BERRACHEDI pour l’honneur qu’il me fait en présidant le jury de cette thèse.

Je tiens à remercier également Madame I. BOUCHEMAKH et Messieurs S. MAAMRA et M. MOULAI qui ont accepté de faire partie du jury de cette thèse.

Je remercie aussi tous les collègues du CERIST pour leur aide précieuse.

# Sommaire

**Partie I : Introduction et état de l'art**

**Partie II : Modèles neuronaux, analyse des données et distribution Ziphienne pour la classification, la cartographie et le découpage des données documentaires**

# Partie I

## Introduction et état de l'art

# I- Introduction

## 1- Problématique

Vu la croissance des systèmes de diffusion et de distribution de l'information et la hausse exponentielle du nombre de chercheurs, l'information scientifique et technique passe par une phase d'explosion à travers la masse des publications disponibles. En effet, la publication est au premier rang des activités scientifiques d'un chercheur, non seulement dans un but de valorisation et de protection de ses droits intellectuels, mais aussi, pour la diffusion de l'information.

Suite à cette explosion, l'utilisateur (chercheur ou étudiant recherchant de l'information) se trouve face à un volume prodigieux d'informations dans lequel il espère localiser l'information pertinente le plus vite possible sans courir le risque de perdre de l'information ou du temps. De plus, la masse importante d'informations représente une matière première intéressante pour obtenir d'autres informations (élaborées) pouvant servir comme indicateurs stratégiques pour la veille et la décision.

Pour faciliter l'accès aux informations, les publications scientifiques sont signalées dans des bases de données dites "*bases de données bibliographiques*" où chaque enregistrement, appelé *notice bibliographique*, correspond à un document. Cette notice bibliographique est composée de champs signalétiques (contenant des données descriptives du document à savoir : titre, auteur(s), date d'édition,... etc.) et de champs analytiques (comprenant : résumé et mots clés). La recherche bibliographique dans ces bases de données se base sur les mots clés en les considérant comme étant les meilleurs indicateurs du contenu d'un document.
Ainsi, l'utilisateur propose sa requête constituée de mots clés jugés représentatifs pour sa recherche, ensuite, relie ces mots par des relations binaires (ou, et, sauf,...) constituant ainsi l'*équation de recherche*.

Cette approche, qui apparaît à première vue intéressante, est limitée par plusieurs contraintes à savoir :

- l'utilisateur peut omettre un ou plusieurs mots pertinents,
- l'évolution des concepts,

- point de vue de l'utilisateur différent de celui de l'auteur,
- l'effet de l'indexeur[1] .

**2- Objectifs**

Notre objectif est de proposer un nouveau modèle de recherche évitant à l'utilisateur la formulation de requête et le choix de mots clés en le guidant vers l'information pertinente qu'il recherche.

Partant du principe que la classification est quelque chose d'inhérent pour l'esprit humain, ce modèle se base principalement sur la *classification thématique* de la masse d'informations contenues dans les bases de données bibliographiques (classification de mots clés et des documents par thème) et la représentation des classes thématiques obtenues sur une carte reflétant leurs interactions et rapprochements.

Ainsi, ce modèle offre à l'utilisateur l'ensemble des thèmes recouvrant le domaine scientifique de la base de données bibliographiques. Chaque thème est relié à un ensemble de mots clés, un ensemble d'auteurs et un ensemble de références bibliographiques. Par la suite, ces thèmes sont représentés sur un espace bidimentionnel, constituant ainsi la *carte thématique*. Cette carte représente une bonne synthèse du contenu de la base de données bibliographiques et un point de départ intéressant guidant l'utilisateur vers le thème qui l'intéresse ensuite vers l'information qu'il recherche.

Enfin, le but principal visé est d'orienter l'utilisateur dans sa recherche dans la base de données bibliographiques en lui évitant la formulation d'une équation de recherche (recherche par mots clés).

D'où l'idée de concevoir un prototype de système de recherche permettant :

- La classification thématique et la représentation cartographique,
- Le stockage des informations élaborées (classes thématiques et carte) et leur représentation sous un format hypertexte,

[1] Documentaliste qui identifie les mots clés des documents.

- La mise en place d'une architecture mixte SGBD[2] - hypertexte permettant à l'utilisateur via une interface, de naviguer dans la base, en se basant sur la métaphore de la carte.

Ce système constitue un outil d'aide intéressant pour :

- guider l'utilisateur à l'essentiel de l'information,
- minimiser le temps de recherche de l'information,
- éviter à l'utilisateur la formulation de l'équation de recherche,
- réduire le risque de perte de l'information,
- permettre le filtrage de l'information,
- résoudre le problème de la dispersion de l'information.

De plus, il est important de signaler que la classification thématique offre aux décideurs une vision globale de l'environnement scientifique, en leur permettant de voir les tendances thématiques de la recherche et les interactions entre ces dernières. La carte thématique constitue donc un *indicateur stratégique.*

Plus encore, d'autres traitements sont inclus dans le système comprenant la génération d'autres indicateurs (quantitatifs) en se basant sur la fréquence des informations dans la base. Ces indicateurs permettent de faire des études comparatives (la comparaison de la production des personnes, des institutions …etc.).

## II- Notions de base

### 1- L'information

L'information est un ancien concept qui reste toujours d'actualité grâce à son importance pour toute communauté scientifique. La définition de ce concept est présente dans un grand nombre de travaux, allant des définitions classiques, telles que :

"L'information se présente sous forme de données plus au moins concrètes, un renseignement, un fait, un concept. Ce peut être un chiffre de population, un énoncé, une image, une photo, un film, un son. Ces données sont repérables et transmissibles grâce à un support concret…" **[15]**

ou "L'information est une connaissance inscrite (enregistrée) sous forme écrite (imprimée ou numérisée), orale ou audiovisuelle "**[17]**

[2] SGBD: Système de Gestion de la Bases de Données bibliographiques.

jusqu'aux nouvelles définitions prenant en considération l'impact des nouvelles technologies sur l'évolution de ce concept :
"L'information est une émission, réception, création, retransmission, de signaux groupés oraux ou écrits, sonores, visuels ou audiovisuels, en vue de la diffusion et de la communication d'idées, de faits de connaissances, d'analyses, de concepts, de thèses, de plans, d'objets, de projets, d'effets de toutes sortes dans tous les domaines"[43].

Dans ce travail, l'information est présentée de manière simple et suffisante pour ce qui va suivre :
L'information représente des données pouvant apporter de la connaissance sur un domaine donné. Elle est caractérisée par un sens, une forme et un support matériel.

**1.1-Typologie de l'information**

L'information peut être classée selon plusieurs critères à savoir :

- Son contenu : information économique, sociale, ...etc.
- Son public : information publique, scientifique, pédagogique.
- Sa nature : sonore, textuelle, graphique.
- Sa diffusion : interne, publiée, restreinte, souterraine (dite littérature grise),
- …etc.

**1.2- L'information scientifique et technique**

L'information considérée dans ce travail est "l'Information Scientifique et Technique" désignée habituellement par le sigle IST et définie par Pascal Prenon comme : " le résultat et le témoin de l'activité de la communauté scientifique"[8].

**2- Le document**

Le document est tout simplement le support matériel de l'information. Il est défini dans les dictionnaires sur la base de son contenu ou de son rôle. Dans le dictionnaire Larousse, le terme document représente " Ecrit servant de preuve ou de titre". Quant au dictionnaire Quillet-Flammarion ce terme désigne : " toute chose qui peut servir à nous renseigner".

Les normes internationales donnent une définition plus large. Dans la norme ISO :
" Le document est un ensemble cohérent et fini, d'informations structurées, lisibles, à usage défini sur un support donné".

Comme c'est le cas pour l'information et grâce à la progression très rapide des nouvelles technologies de l'information, principalement les réseaux, le concept document est caractérisé par une importante évolution : Il devient composite par l'intégration du son et de la vidéo en plus du texte et des graphiques à son contenu et il passe à la numérisation qui enrichit son contenu, permet son interactivité et facilite sa diffusion.

**2.2-Typologie des documents**

Les documents sont également classés sur la base de plusieurs critères :

Nature de l'information :

- Numérique,
- Textuelle,
- Sonore,
- Iconique,
- Graphique.

Nature du support :

- Imprimé (papier),
- Photographique(microformes),
- Magnétique,
- …etc.

Son rôle :

- Document de référence (annuaire, encyclopédie…),
- Document secondaire (bulletin de résumé, bibliographie…),
- …etc.

Concernant le document imprimé, une classification classique et connue (non exhaustive) est la suivante :

- Périodique,
- Thèse et mémoire,
- Rapport,
- Compte rendu de manifestation scientifique,

- Livre et ouvrage,
- Norme,
- Manuel,
- …etc.

## III- Représentation de l'information

### 1- L'indexation

La représentation de l'information se fait grâce à l'opération d'indexation qui est la "description du contenu d'un document à l'aide d'un langage documentaire pour faciliter la mémorisation de l'information dans un fichier en vue d'une recherche ultérieure"**[11].**
Une définition plus détaillée est la suivante :
"L'opération qui consiste à décrire et à caractériser un document à l'aide de représentation des concepts contenus dans ce document, c'est à dire à transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse. La translation en langage documentaire[3] se fait grâce à des outils d'indexation tels que le thesaurus[4], classifications,…etc."**[20].**

L'indexation consiste donc, en l'extraction d'un ensemble de concepts appelés *mots clés* qui représentent le mieux le contenu sémantique du document et la translation de ces derniers, grâce à un vocabulaire normalisé en vue de permettre la recherche d'information par sujets.

Cette opération comporte deux phases essentielles :
- La reconnaissance des concepts qui représente l'information contenue dans le document,
- La représentation de ces concepts dans un langage normalisé.

Si la deuxième phase de l'indexation n'est pas effectuée, on parle alors de *l'indexation en langage libre ou en langage naturel.*

[3] "Langage artificiel constitué de représentations de notions et de relations entre ces notions, et destiné, dans un système documentaire, à formaliser les données contenues dans les documents et dans les demandes des utilisateurs" norme AFNOR.
[4] "le vocabulaire d'un langage d'indexation contrôlé organisé formellement de façon à expliciter les relations à priori entre les notions" norme ISO 2788.

**2- Base de données bibliographiques**

Avant de définir ce concept, il serait préférable de présenter la notion *de référence ou notice bibliographique* qui représente une description du document : auteur, titre, date d'édition, note,... etc. Elle est dite *analytique* [5]si elle comprend, également, les mots clés et le résumé de ce dernier.

Une *base de données bibliographiques* est un fichier informatisé regroupant un ensemble de références bibliographiques organisées de manière cohérente. Elle est constituée pour permettre une recherche informatisée des informations.

Les bases de données bibliographiques sont gérées grâce à des systèmes informatiques appelés *SGBD* (Systèmes de Gestion de Bases de Données). L'objectif de ces systèmes est d'assurer le stockage, l'accès et la manipulation de grands volumes de données. Ils doivent garantir leur intégrité, fiabilité, sécurité et confidentialité.

**3- Base de données en texte intégral**

La grande révolution caractérisant les supports d'information, en particulier du point de vue capacité, associée aux nouvelles attentes et exigences des utilisateurs de l'information ont encouragé la naissance d'un nouveau type de bases de données. Ces bases de données sont les ***bases de données textuelles,*** dites aussi en ***texte intégral*** ou encore ***plein-texte*** comme elles sont définies par Pierre-Marie Belbenoit-Avich. "Les bases de données plein-texte comme leur nom l'indique ne donnent pas simplement des références bibliographiques, même si elles sont parfois couplées aux bases bibliographiques,…Mais, d'une manière précise, une base plein-texte fournit le texte intégral des documents" **[22].**

Le principal avantage de ces bases est qu'elles fournissent à ces utilisateurs non pas des références bibliographiques, seulement, mais l'information recherchée dans son intégralité.

[5] Par la suite l'appellation " notice bibliographique" sera utilisée pour désigner une notice analytique.

## 4- L'indexation automatique

L'indexation telle qu'elle a été présentée auparavant est réalisée manuellement, autrement dit par une personne appelée indexeur. Dans ce cas, les termes d'indexation ou les mots clés sont choisis par l'indexeur, donc, dépendent de son niveau de connaissance du sujet traité par le document. Par conséquent, il est possible que des termes d'indexation pertinents soient omis ou des termes trop génériques soient choisis. De plus, "deux sujets différents ne choisissent qu'à 70% des mots clés identiques pour indexer un même document en utilisant le même thesaurus"**[46].**

Les inconvénients de l'indexation manuelle rajoutés à l'apparition des bases de données en texte intégral ont favorisé le passage à ***l'indexation automatique*** pour la représentation des documents qui consiste en l'extraction par ordinateur des termes d'indexation représentant le contenu des documents textuels. Elle peut être considérée "comme une forme d'acquisition de connaissances" **[45].**

L'indexation automatique se base sur les méthodes suivantes :

### 4.1- Méthodes d'extraction des termes d'indexation

#### 4.1.1- Méthode statistique

Cette méthode repose sur l'idée que plus un mot apparaît dans un document, plus il est représentatif de son contenu. Ainsi, elle se base sur la fréquence d'apparition des termes dans le texte à indexer. Idée suggérée dés 1957 par H.P. LUHN **[26].** Néanmoins, quelques problèmes se posent, à savoir :

- Les termes les plus fréquents dans le texte ne sont pas, nécessairement, intéressants pour sa représentation (comme : "un", "des", "le", "la", "les", "à", "et"…etc.). Cette classe de termes est connue sous le nom de ***Mots vides.***

- Les termes peuvent être présents dans le texte sous leurs différentes variantes morphologiques (féminin, pluriel,…etc.).

- La présence de mots composés intéressants pour l'indexation du texte (comme par exemple : "intelligence artificielle" ou "système expert" en informatique ).

Pour la résolution du premier problème, l'indexation automatique fait recourt à l'intégration, dans la machine, de la liste des mots vides sous forme d'un anti-dictionnaire contenant : articles, prépositions,... etc.

Quant au deuxième problème c'est l'opération de troncature à droite qui est souvent utilisée.
Ainsi un terme d'indexation correspond à une forme radicale d'un mot. Le nombre de caractères optimum de la forme radicale à garder se situe dans l'intervalle [5,8] avec une préférence de 7.

Enfin, pour le repérage des mots composés, il suffit d'utiliser la fréquence conjointe des paires de termes relativement à leurs fréquences individuelles**.** C'est à dire, que si deux termes sont souvent associés dans un texte, en plus, leur forme associée est très fréquente dans le texte, ils peuvent former une unité intéressante pour l'indexation. Une autre solution consiste à utiliser certaines propositions qui servent à lier les concepts entre eux de façon privilégiée, comme par exemple : "étude des besoins", "évaluation du personnel", …etc.

**4.1 2- Méthode lexicale**

Comme son nom l'indique, cette méthode s'appuie sur un lexique. Elle se base sur l'extraction des termes présents et dans le document et dans ce lexique. Cette approche représente plusieurs avantages. En effet, un lexique est constitué par des experts dans le domaine de la base. De plus, il contient les associations des termes avec leurs différentes formes morphologiques, les liens de synonymie, de hiérarchie et de composition.

Les principales inconvénients de cette approche résident dans la difficulté de constituer un lexique (qui se fait d'une manière générale manuellement), la taille très importante de ce lexique qui nécessite un espace mémoire important, l'organisation complexe de ce dernier induisant un temps long de réponse et le fait qu'il reste toujours incomplet surtout lorsqu'il traite d'un domaine très spécialisé.

Selon FLUHR **[32]**, l'expérience a montré qu'une meilleure satisfaction est obtenue lorsque les deux méthodes sont couplées pour effectuer l'indexation automatique. Autrement dit, la

considération de la fréquence des termes dans le texte ainsi que leur présence dans un lexique. Comme c'est le cas du système SPIRIT de la société SYSTEX.

**5- Pondération des termes d'indexation**

La pondération des termes d'indexation se base sur l'idée que les termes qui sont très rares dans l'ensemble des documents mais très fréquents dans un document particulier représentent plus ce dernier par rapport à d'autres qui sont très fréquents et dans l'ensemble et dans ce document en particulier. De ce fait, le poids d'un terme d'indexation pour un document, selon SPARCK JONES, est proportionnel à sa fréquence d'apparition dans ce dernier "fréquence relative" et inversement proportionnel à sa fréquence globale dans les documents "fréquence absolue".

La fréquence absolue inverse d'un terme tj est donnée par la formule suivante:

$$FABj = \text{Log}\left( N / Fj \right) + 1$$

où :

N: Le nombre total de documents dans la base,

Fj: Le nombre de documents contenant le terme tj.

Concernant le poids du terme tj par rapport au document Di, il est donné par la formule:

$$Pij = Fij \times FABj$$

où :

Fij : La fréquence d'apparition du terme tj dans le document Di .

## IV- Système de recherche d'information

L'objectif des Systèmes de Recherche d'Information, notés SRI, est de permettre à l'utilisateur de rechercher et d'accéder à l'information qui l'intéresse. Ainsi, en plus des fonctions de mémorisation et d'organisation des données, un SRI doit inclure des fonctions de recherche et de consultation de l'information retrouvée.

Autrement dit, un système de recherche d'information est un ensemble de programmes qui interprètent les questions, recherchent les informations dans des fichiers et retournent les informations trouvées à la personne qui a posé la question.

D'une façon générale, la recherche se fait grâce à une mesure de ressemblance entre la requête ou simplement la demande de l'utilisateur et la représentation des documents dans la base de données. Pour pouvoir mesurer cette ressemblance, la requête de l'utilisateur est sensée être représentée de la même façon que la représentation des documents. Ainsi, les résultats de recherche dépendent de la représentation des documents, de la précision de la requête et du modèle de recherche permettant la mesure de la ressemblance.

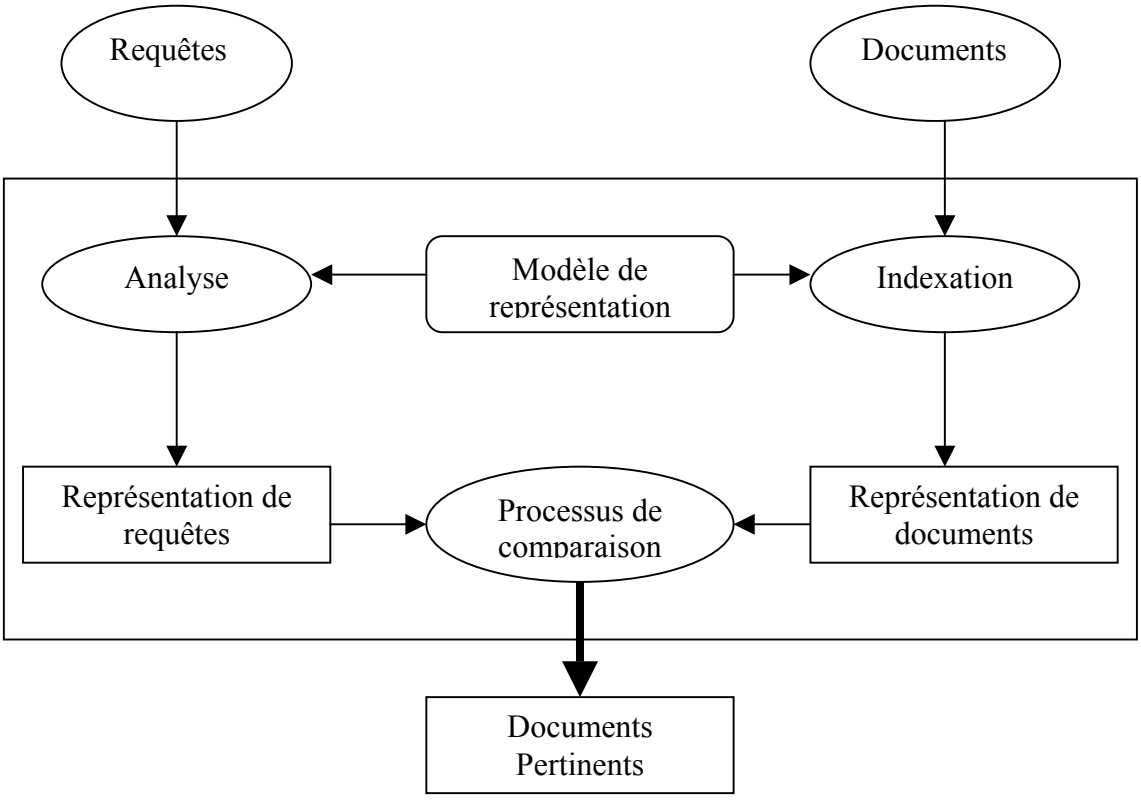
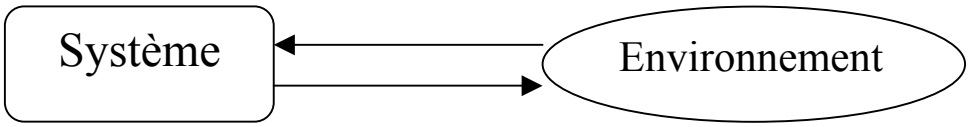



Schéma du mécanisme de recherche utilisant l'indexation

**V- Modèles de recherche d'information**

Plusieurs travaux se sont intéressés à la modélisation des Systèmes de Recherche d'Information. Il existe des auteurs qui ont traité l'aspect structurel, autrement dit, la représentation les différents composants d'un SRI, leur hiérarchie et leurs différentes interactions. D'autres auteurs se sont intéressés à la logique de la recherche d'information (c'est à dire comment se fait la recherche par le système). Ce qui englobe, la modélisation de la demande de l'utilisateur, la modélisation des documents et la mise en correspondance entre la demande et les documents.

Concernant l'aspect structurel des SRI, le modèle le plus simple rencontré est proposé par Christine Michel**[5]**. Ce modèle représente le SRI comme entité indivisible soumise à des entrées imposées par l'environnement (l'utilisateur) transformées en sorties satisfaisant certains critères. Ce modèle est dit système "boite noire". Il se schématise comme suit:




Un système "boite noire"**[5]**

Concernant l'aspect logique, il prend une part plus importante dans littérature. Ainsi, Il existe plusieurs modèles de recherche d'information proposés.

Les approches classiques utilisent des modèles mathématiques pour la représentation des documents et des requêtes et le calcul de la similitude entre ces deux derniers. Les modèles les plus connus sont : le modèle booléen, le modèle utilisant la logique floue, le modèle vectoriel et le modèle probabiliste.

Concernant les nouvelles approches ou tendances des SRI, des changements englobant l'aspect structurel et logique sont proposés, comme ceux se basant sur la reformulation automatique des requêtes ou la prise en compte du profil utilisateur.

## 1- Modèles classiques de recherche d'information

Ces modèles diffèrent, principalement, dans leur mode de représentation des requêtes et la manière de faire la correspondance requête/document par le système.

### 1.1- Modèle booléen

Dans ce modèle, la requête est constituée d'un ensemble de mots clés reliés entre eux par des opérateurs d'intersection, d'union ou de différence (AND, OR et NOT).

**Exemples**

**Requete1:**

" Cancer" **AND**" Sang"

Pour chercher des documents qui traitent du "***cancer du sang*".**

**Requète2:**

("Informatique "**OR**" Statistique") **AND** "Analyse de données"

Pour chercher des documents qui traitent de ***l'analyse des données du point de vue informatique ou statistique.***

Dans ce modèle, le système utilise une fonction d'équivalence entre la requête exprimée en langage booléen et les documents. Cette relation est soit vraie soit fausse (binaire). Les documents donnés en sortie sont ceux qui réalisent une valeur vraie (c'est à dire ceux qui répondent exactement à la requête). Par conséquent, les documents résultats ne sont pas soumis à un ordre de pertinence.

### 1.2- Modèle utilisant la logique floue

La logique floue a été formulée par A. Zadeh dans le milieu des années 60. Elle vient de la constatation que la variable booléenne, qui ne peut prendre que deux valeurs (vrai ou faux) est mal adaptée à la représentation de la plupart des phénomènes courants. Alors que la logique classique considère qu'une proposition est soit vraie soit fausse, la logique floue distingue une infinité de valeurs de vérité (entre 0 et 1). Par conséquent, la logique floue est une généralisation de la logique booléenne classique. Elle ajoute cependant une fonctionnalité déterminante : la possibilité de calculer un paramètre en disant simplement dans quelle mesure il doit se trouver dans telle ou telle zone de valeur.

Ainsi, le modèle de recherche basé sur la logique floue est une extension du modèle booléen. Car, les requêtes sont, aussi, exprimées en langage booléen (c'est à dire des mots clés reliés par les opérateurs booléens).

Cependant le calcul de l'équivalence (requête/document) prend en considération le poids des mots dans les documents. Autrement dit, l'association d'un mot de la requête avec un document n'est pas binaire (1 s'il apparaît 0 sinon), mais elle est exprimée par le poids du mot dans ce document.

Par conséquent, la valeur de la fonction d'équivalence entre la requête et le document n'est, également, pas binaire. Mais elle est comprise entre 0 et 1.

**Exemple**

Soient :

- Di un document de la base,
- t1 et t2 deux mots clés.

Les valeurs des fonctions d'équivalence des requêtes suivantes avec le document Di sont données par :

| **Requête** | **Fonction d'équivalence** |
|---|---|
| T1 **AND** t2 | Min (Pi(t1), Pi(t2)) |
| T1 **OR** t2 | Max (Pi(t1), Pi(t2)) |
| **NOT** (t1) | 1- Pi(t1) |

Pi(tj) est le poids du mot tj ( $j \in \{1,2\}$)dans le document Di.

Le document Di est pris comme résultat de recherche si la valeur de la fonction d'équivalence est supérieure à un certain seuil. Ce seuil peut être fixé par l'utilisateur.

Il est tout à fait évident que dans ce modèle, les résultats (documents) peuvent être présentés selon un ordre de pertinence ou de similitude avec la requête. Ceci grâce à la valeur de la fonction d'équivalence.

### 1.3- Modèle vectoriel

Ce modèle considère les documents ainsi que les requêtes comme des points dans un espace multidimensionnel ou chaque dimension correspond à un mot clés. Ainsi à chaque document (resp. requête) correspond un vecteur binaire dont la $i^{ème}$ composante est égale à 1 si le mot ti est présent dans le document (resp. requête) et 0 sinon.

Le mécanisme de recherche consiste à repérer les documents dont les vecteurs correspondants sont les plus proches du vecteur requête. Ce qui implique une mesure de proximité entre vecteurs. Les mesures les plus utilisées sont le **produit scalaire** entre vecteurs et la **mesure cosinus** de l'angle entre les vecteurs.

- L'utilisation du produit scalaire induit la formule suivante:

$$\text{Similitude } (Di,Q) = \sum_{j=1..I} qj \times dij$$

telle que:

Di ($d_{i1}$, $d_{i2}$, ….. $d_{ij}$, …… $d_{iI}$) représente le vecteur associé au document i.

Q ($q_1$, $q_2$, $q_j$, $q_I$) représente le vecteur requête.

I est le nombre de mots clés dans la base.

- Quant à la mesure cosinus, elle induit :

$$\text{Similitude } (Di,Q) = \frac{\sum_{j=1..I} qj \times dij}{\left(\sum_{j=1..I} (qj)^2\right)^{1/2} \times \left(\sum_{j=1..I} (dij)^2\right)^{1/2}}$$

Le choix des documents résultat se fait sur la base d'une comparaison entre la valeur de la similitude document/requête et un certain seuil.

Une amélioration de cette approche consiste à considérer non pas la présence et l'absence des mots dans le document et la requête, pour leur représentation vectorielle, mais leurs poids dans ces derniers.

Ainsi dans la formule précédente :

- dij représente le poids du mot clés j dans le document Di ,
- et qj représente le poids du mot clés j dans la requête Q.

Sachant que la pondération des mots clés dans une requête, est donnée par l'utilisateur.

### 1.4- Modèle probabiliste

Ce modèle se base sur le calcul de la probabilité conditionnelle en faisant intervenir le processus d'indexation.

Un document D est présenté comme suit :

D(t1,t2,…..tn) sachant que :

ti=1 si le mot **i** indexe **D**,

ti=0 sinon.

Les probabilités calculées sont :

P(ti/Pert) : la probabilité que le mot i apparaisse dans un document sachant que le document est pertinent pour la requête.

P(ti/nonPert) : la probabilité que le mot i apparaisse dans un document sachant que le document est non pertinent pour la requête.

En utilisant la formule de Bayes avec les deux hypothèses suivantes:

- La distribution des termes dans les documents pertinents est la même que leur distribution par rapport à la totalité.
- Les deux variables aléatoires «document pertinent» et «document non pertinent» sont indépendantes.

La fonction de recherche se base sur le calcul de probabilité de pertinence d'un document D, notée P(pert/D).

$$P(pert/D) = P(D/pert)*P(pert) / P(D)$$

$$P(nonpert/D) = P(D/nonpert)*P(nonpert) / P(D)$$

avec :

$$P(D)= P(D/pert)*P(pert)+ P(D/nonpert)*P(nonpert)$$

Si t1,t2,….tI sont indépendantes alors :

P(D/pert)= P(t1/pert)* P(t2/pert)* ……*P(tI/pert)

avec :

P(ti/pert)=ri/R

sachant que :

R : nombre de documents pertinents,

ri : nombre de documents pertinents contenant ti.

P( ti/nonpert)=(ni-ri)/(N-R)

sachant que :

ni : nombre de documents contenant ti,

N : nombre de documents.

## 2- Nouveaux modèles de recherche d'information

### 2.1- Modèle de recherche basé sur le profil des utilisateurs

Ce modèle concerne la recherche d'information dans les bases de données en texte intégral. Il vise à améliorer la précision et le rappel du système de recherche par rapport à une requête d'un utilisateur. Il se base sur le constat que des utilisateurs qui s'intéressent au même sujet n'attendent pas, nécessairement, la même information. En effet, certains attendront des documents synthétiques et d'autres détaillés, certains pédagogiques et d'autres professionnels,…etc.

De plus, une enquête menée auprès d'un certain nombre d'utilisateurs a prouvé que la majorité de ces derniers (ayant répondu) font rarement une lecture séquentielle et globale mais souvent partielle et dépend de ce qu'ils souhaitent en faire (publier un article, monter un cours, se tenir informer,... etc.)**[1]**.

De ces fais, ce modèle vise à incorporer des nouveaux éléments complétant la représentation classique des requêtes et des documents qui comprend une simple liste de descripteurs (mots clés). Il se base sur les phases essentielles suivantes :

1- Découper les documents en unités documentaires en respectant les contraintes suivantes:
   - L'unité doit être significative hors contexte (autonome),
   - L'unité ne doit pas dépasser quelques pages,
   - L'unité doit être homogène (constitue ou fait partie d'une unité délimitée par l'auteur dans le plan du document).

2- Caractérisation de ces unités (fonction d'aiguillage) selon deux types de propriétés celles qui s'appliquent à l'ensemble du document et celles qui lui sont spécifiques :

| | |
|---|---|
| Type | Résumé<br>Table des matières<br>Introduction<br>Description de contexte<br>Description de thème<br>Environnement<br>Expérimentation<br>Résultats<br>Discussion<br>Méthode<br>Conclusion<br>Bibliographie |
| Forme discursive | Descriptive, narrative, argumentative, discours rapporté. |
| Style de présentation | Littéraire<br>Littéraire avec données numérique<br>Données numériques<br>Calcul<br>Représentation |

3- Caractériser le profil de l'utilisateur selon quatre propriétés :

| | |
|---|---|
| Niveau éducationnel | Maîtrise/ DEA/Recherche |
| Champ disciplinaire | SIC/Informatique/Agronomie/Pharmacie/… |
| Etape de recherche | Constitution d'une bibliographie<br>Définition du sujet<br>Faisabilité<br>Expérimentation<br>Interprétation des données<br>Rédaction<br>Repérage des approches expérimentales<br>Plan de travail<br>Compréhension de la problématique<br>Etat de l'art<br>Synthèse bibliographique<br>Dégagement des nouveaux axes de recherche<br>Mise à jour des connaissances |
| Type de recherche | Recherche pointue<br>Recherche Généraliste |

4- Compléter la requête par le profil de l'utilisateur et la représentation du document par la caractérisation de ses unités documentaires.

5- Faire la correspondance entre la requête et la représentation du document.

L'utilisateur reçoit en résultat un ensemble d'unités documentaires à partir desquelles il peut remonter aux documents globaux.

**Expérimentation: Profil-Doc**

Les hypothèses de ce modèle étant théoriques, un prototype d'un tel système devait être réalisé. Un tel Projet a été proposé, pour la première fois, par Sylvie Lainé Cruzel en septembre 1992 **[5]** sous l'intitulé Profil-Doc. En 1996**[5]**, plusieurs membres du Laboratoire RECODOC de l'université Claude Bernard Lyon1 ont publié sur le sujet d'où la description totale du projet et la réalisation d'un prototype.

Ce prototype est composé :

1- d'une base de données d'unités documentaires,
2- d'une interface d'interrogation,
3- et d'une interface de consultation.

**2.2- Modèle de recherche basé sur la reformulation de la requête**

L'objectif de ce modèle est de limiter les problèmes liés aux mauvais choix de termes dans la requête. Il se base sur la reformulation de cette dernière en utilisant les deux approches suivantes :

La première approche se base sur la notion de voisinage d'un terme. Cette notion est issue d'études sur le langage naturel. Le voisinage est dégagé d'un thesaurus dans lequel les termes sont reliés par des relations sémantiques (synonymie et hiérarchie).

La seconde approche se base sur le mécanisme de la réinjection de la pertinence. Les documents résultat sont présentés à l'utilisateur qui fournit un jugement sur leur pertinence. Ce jugement est utilisé pour la pondération des termes de la requête. Cette nouvelle requête permet une nouvelle recherche en considérant le jugement de l'utilisateur.

Ces deux approches sont complémentaires d'où la possibilité de les combiner. La principale limite de ce modèle est qu'il représente une adaptation à court terme aux besoins des utilisateurs. Autrement dit, un utilisateur ne peut bénéficier des résultats obtenus par un autre utilisateur.

# Partie II
# Modèles neuronaux, analyse des données et distribution Ziphienne pour la classification, la cartographie et le découpage des données documentaires

# I- Les réseaux de neurones artificiels

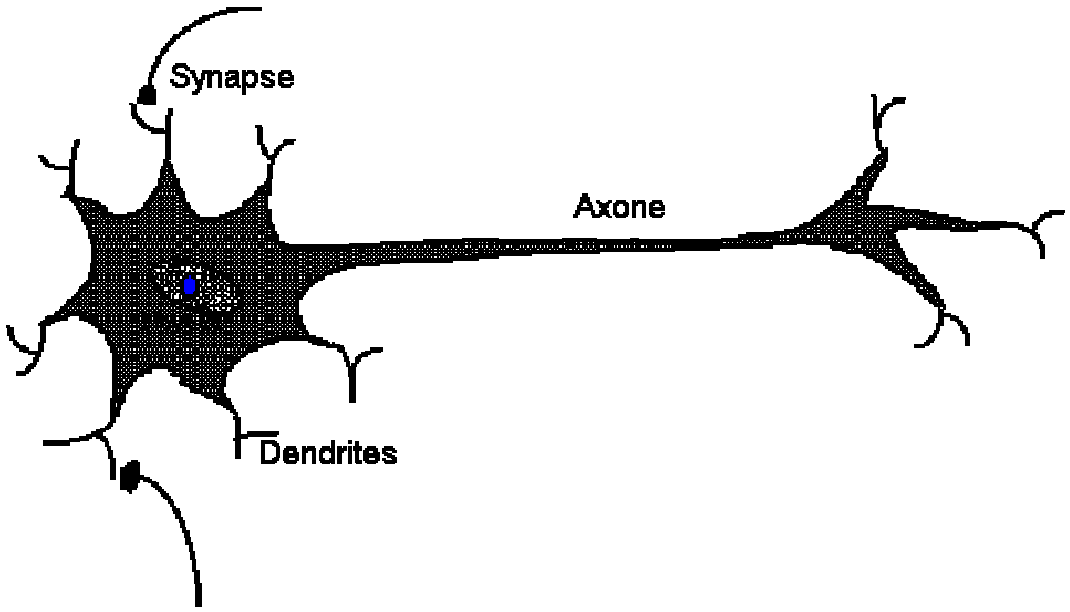
## 1- Introduction

Le cerveau humain possède des caractéristiques, intéressantes, absentes de l'architecture des machines parallèles malgré leurs grandes capacités de calcul. Parmi ces dernières, il existe en particulier son parallélisme, sa grande capacité d'apprentissage, de généralisation, d'adaptation,…etc. , d'où l'idée d'essayer de récupérer certaines de ces caractéristiques. Dans ce sens et inspirés des réseaux de neurones biologiques, les réseaux de neurones artificiels cherchent à utiliser les principes organisationnels que l'on pense être mis en œuvre dans le cerveau humain.

## 2- Neurone biologique

Un neurone est une cellule biologique composée d'un corps cellulaire dit *soma* et de deux types de prolongements : les *dendrites* et l'*axone.*

- Le soma possède un noyau qui contient l'information génétique de l'organisme contenant le neurone.
- Les dendrites permettent au neurone de recevoir des signaux provenant d'autres neurones.
- L'axone permet de transmettre l'information générée dans le corps cellulaire sous forme de signaux vers d'autres neurones. L'extrémité de l'axone possède la forme d'une arborisation dont chaque branche se termine par un bouton synaptique autour duquel se trouvent les synapses. Ces derniers constituent les unités de communication axo-dentritique.




**Neurone biologique[42]**

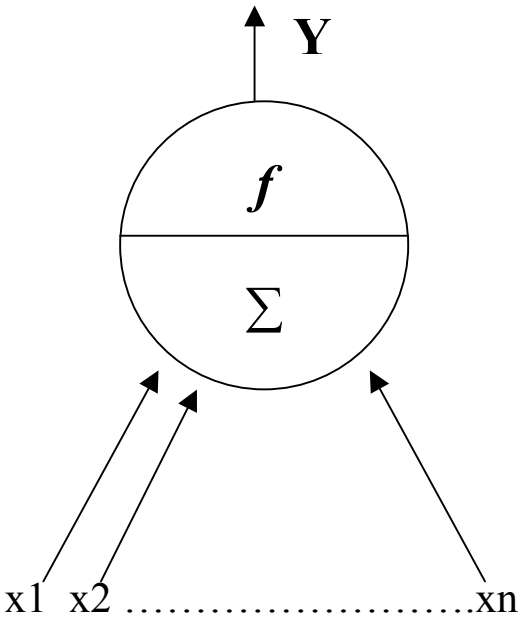
### 3- Neurone formel

Le premier modèle d'un neurone formel a été introduit en 1943 par Mac Culloch et Pitts**[42]**. Ce neurone calcule une somme pondérée de ses n entrées : $x_1, x_2, \ldots x_n$ *(potentiel du neurone)*, renvoie 1 si cette somme est supérieure à un certain seuil θ et 0 sinon. D'où :

$$Y = f(\sum_j wjxj-\theta)$$

où f est la fonction d'activation et wj le poids de la connexion associée à la $j^{ème}$ entrée (poids synaptique).

La valeur de la fonction Y, qui représente la sortie du neurone, est appelée parfois *activité du neurone.*




**Neurone formel [7]**

Ce modèle a été généralisé, par la suite, par le choix d'autres fonctions d'activation, telles que les fonctions linéaires par morceaux, les sigmoïdes et les gaussiennes. Les fonctions les plus utilisées dans les réseaux de neurones artificiels (RNA) sont les sigmoïdes qui sont des fonctions continues, strictement croissantes et définies par :

$$G(x)= (1+e^{-\beta x})^{-1}$$

**4- Réseau de neurones artificiels**

Un réseau de neurones artificiels est un ensemble de neurones formels interconnectés et évoluant dans le temps par des interactions réciproques. Les neurones du réseau sont de deux types :

- Les neurones cachés : leurs sorties ne font pas partie des sorties du réseau.
- Les neurones de sortie : leurs sorties constituent les sorties du réseau.

Un RNA peut être représenté par son **graphe de connexions** qui est un graphe orienté et pondéré. Les sommets du graphe sont les neurones formels du réseau et les arcs sont les connexions entre sorties des neurones et entrées d'autres neurones. Chaque arc est muni d'un poids.

## 5- Typologie des Réseaux de neurones artificiels

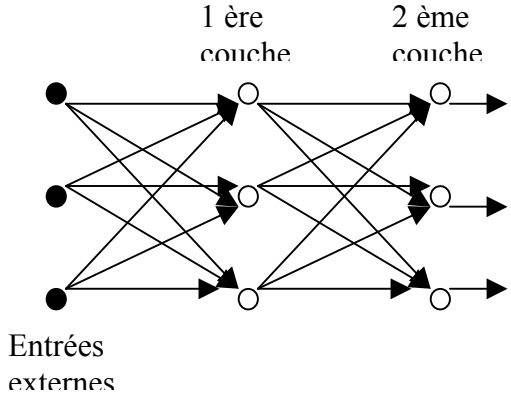
Selon leur représentation par le graphe de connexions, les RNA sont regroupés en deux catégories :

### 5.1- Réseaux non bouclés

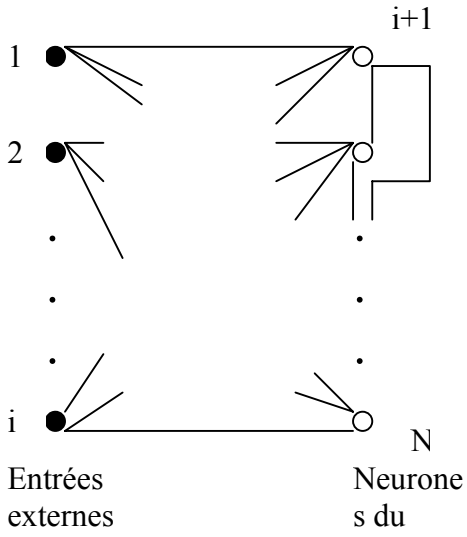
Les réseaux non bouclés sont les réseaux dont le graphe de connexions est acyclique. Ils sont appelés *réseaux statiques*. Ils sont divisés, à leur tour, en deux familles: les *réseaux à couches* et les *réseaux complètement connectés*. Les réseaux à couches, appelés perceptron mitli-couches (mono-couche dans le cas d'une seule couche), sont les plus répondus. Dans ce cas, les sommets du graphe de connexions (neurones) sont organisés en couches. Les neurones d'une couche reçoivent leurs entrées de ceux de la couche précédente et transmettent leurs sorties à la couche suivante. Pour la deuxième famille des réseaux non bouclés qui est la famille des réseaux complètement connectés, chaque neurone du réseau reçoit les entrées externes du réseau et les sorties de tous les neurones de numéros inférieurs.

**Remarque**

Dans le cas d'un perceptron multi-couches les neurones de la première couche reçoivent les entrées externes du réseau.




**Réseaux à couches**

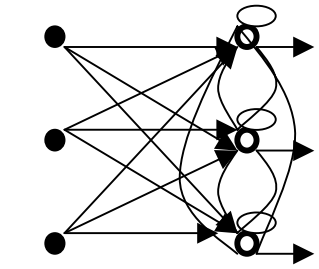


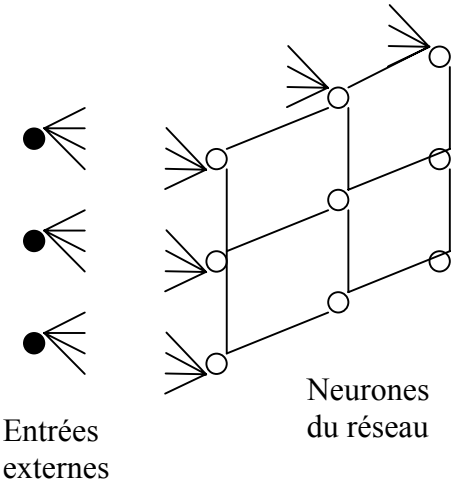

**Réseau complètement connecté**

### 5.2- Réseaux bouclés

Les réseaux bouclés sont caractérisés par un graphe de connexions contenant des cycles. Ils sont appelés *réseaux dynamiques* ou *réseaux récurrents*. Dans cette catégorie nous distinguons les réseaux à compétitions, les SOM,…etc.

Les réseaux les plus répondus de cette catégorie ont la structure suivante :




**Réseaux à compétition** **SOM**

## 6- L'apprentissage dans les réseaux de neurones

La caractéristique fondamentale du cerveau humain est sa capacité d'apprentissage. Dans le cadre des réseaux de neurones artificiels, l'apprentissage est ″le problème de la mise à jour des poids des connexions mises en œuvre dans un réseau afin de réussir la tâche qui leur est assignée″ **[42].**

L'apprentissage permet au réseau de neurones de s'adapter à son environnement. Il se base, en général, sur une collection d'exemples pour la modification des poids, pour arriver après l'adaptation à fournir la sortie appropriée à l'entrée introduite.

Pour définir un processus d'apprentissage, il faut :

- Modéliser l'environnement, en d'autres termes, savoir le type d'informations fournies pour permettre l'apprentissage. Cette modélisation constitue le *paradigme d'apprentissage.*
- Ensuite, comprendre la règle qui régie la mise à jour des poids : *règle d'apprentissage*

Il existe trois principaux paradigmes d'apprentissage :

- **Supervisé** : on fournit au réseau la sortie attendue pour chaque patron d'entrées. Les poids sont définis de façon à avoir une sortie aussi proche que celle fournie.
- **Non supervisé** : aucune information sur la sortie n'est donnée. L'apprentissage se base, dans ce cas, sur la structure des données et leurs corrélations.
- **Hybride** : combine les deux paradigmes suscités. Une partie des poids peut être déterminée grâce à un apprentissage supervisé et l'autre à un apprentissage non supervisé.

Aussi, il existe plusieurs règles d'apprentissage :

- La règle de correction d'erreurs,
- Apprentissage de Boltzmann,
- Règle de Hebb,
- Règle d'apprentissage par compétition,
- Règle de Delta,
- Règle de Widrow-Hoff,
- ....etc.

## 7- Problèmes résolus grâce aux réseaux de neurones artificiels

Les réseaux de neurones s'appliquent aux problèmes de :

- Classification,
- Catégorisation,
- Approximation de fonction,
- Prédiction/prévision,
- Optimisation,
- Contrôle.

## II- La méthode de classification K Means Axiales

### 1- Introduction

La méthode K Means Axiales est une méthode de classification proposée par Alain Lelu dans le cadre de sa thèse de doctorat en 1993. Elle s'inspire de l'algorithme K Means réalisé par Mac Queen en 1967 **[4]** et se base sur une loi d'apprentissage, de réseaux de neurones artificiels, modifiée qui est la loi d'apprentissage d'Oja.

Par conséquent, avant la présentation de la méthode K Means Axiales, il faut commencer par introduire l'algorithme de Mac Queen et la loi d'Oja.

### 2- La méthode K Means

Cette méthode fait partie de la famille des méthodes de classification à centres mobiles. Elle permet de classer les T lignes d'un tableau X en K classes.

L'algorithme de cette méthode se présente comme suit :

Soit X = {x } un tableau de T lignes x et de I colonnes et K le nombre de classes fixé à priori.

**Algorithme K Means**

**[0]** Initialiser au hasard les K centres de gravité des classes, notés m(k) :

$$m^0(k) = [\mu_1{}^0(k), \mu_2{}^0(k), \ldots, \mu_I{}^0(k)]$$

**[1]** Pour chaque ligne x du tableau telle que x = [ $\xi_1, \xi_2, \ldots, \xi_I$] :

- Calculer ses K distances aux points m(k)

  $$\delta^2 (x, m(k)) = (\xi_1 - \mu_1(k))^2 + (\xi_2 - \mu_2(k))^2 + \ldots\ldots + (\xi_I - \mu_I(k))^2$$

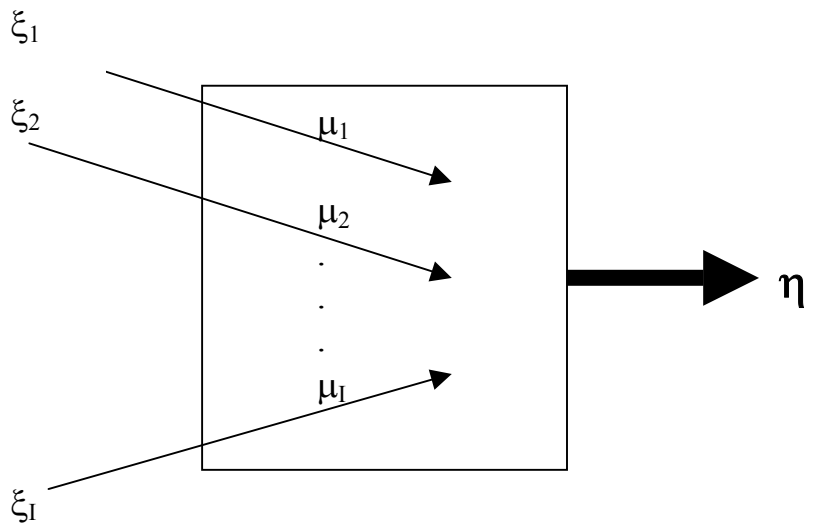
- Incorporer x à la classe $\underline{k}$ pour laquelle cette distance est minimale, incrémenter le cardinal $t_{\underline{k}}$ de la classe $\underline{k}$ et mettre à jour la position du centre de gravité m($\underline{k}$).

Mac Queen a démontré la convergence des K vecteurs m(k) vers des points fixes mais n'a rien dit sur la décroissance de $\phi = \sum_k \sum_x \delta^2 (x , m(k))$ qui représente un critère de qualité de la classification.

En plus, l'initialisation au hasard ne garantit pas que le nombre de classes soit égal à K, puisque certaines classes peuvent être vides. D'où la préférence d'initialiser les vecteurs m(k) par les K premières lignes x. Ce qui garantie l'existence d'au moins un point par classe.

### 3- La loi d'apprentissage d'Oja

Oja schématise un neurone par une boite noire soumise à un vecteur d'entrée x ($\xi_1$, $\xi_2$, , $\xi_I$) et répondant par une valeur de sortie $\eta$. Les poids synaptiques du neurone sont représentés par le vecteur $m(\mu_1, \mu_2, \ldots, \mu_I)$.




**Shématisation du Neurone formel d'Oja**

Le vecteur m des poids synaptiques evolue (du temps t au temps t+1) selon la loi d'Oja qui est fonction de x, $\eta$ et de son état au temps t:

$$m^{t+1} = m^t + \alpha^t \eta^t (x - \eta^t m^t)$$

tel que:

- $\sum_t (\alpha^t)^2$ est convergente,
- $\sum_t \alpha^t$ diverge.

La sortie du neurone est égale au produit sacalaire entre le vecteur $m^t$ et le vecteur x:

$$\eta^t = <x, m^t>$$

Oja a montré que la loi d'évolution du vecteur $m^t$ le fait tendre vers le premier vecteur propre de la matrice $X^TX$ et fait tendre $||m||$ vers 1.

**4- La loi d'apprentissage d'Oja modifiée**

Alain Lelu a montré **[4]** qu'un ensemble de K neurones du type Oja ayant des vecteurs de poids synaptiques m(k) (k = 1..K) utilisant la même loi d'activation (produit scalaire) permet d'avoir les K premiers vecteurs et valeurs propres $\lambda(k)$ en utilisant la loi d'apprentissage suivante :

Pour chaque entrée x :

$$m^{t+1}(k) = m^t(k) + (\eta^t(k)/\tau^t(k))(x - e^t(k)) \qquad (1)$$

$$\text{Avec } e^t(k) = \eta^t(k)\, m^t(k) - \sum_{j=1,K \quad j \neq k} \eta^t(j)\, m^t(j)\, \tau^t(k)/(\tau^t(k) - \tau^t(j))$$

$$\tau^{t+1}(k) = \tau^t(k) + \eta^t(k)^2$$

Il faut noter que la valeur propre $\lambda(k)$ et reliée avec $\tau(k)$ par la formule:

$$\tau(k) = t\lambda(k)$$

Pour extraire une valeur approchée du premier vecteur propre, il suffit d'utiliser un seul neurone (K=1).

Ainsi, la loi d'apprentissage (1) devient:

$$m^{t+1}= m^t+(\eta^t/\tau^t)( x- \eta^t m^t) \qquad (2)$$
$$\tau^{t+1}= \tau^t+ (\eta^t)^2$$

**5- La méthode K Means Axiales**

La méthode K Means Axiales s'inspire de la méthode K Means et se base sur la loi d'apprentissage d'Oja modifiée. Cette méthode considère que les données sont normalisées. Chaque classe est représentée par son axe d'enertie passant par l'origine et identifié par le vecteur unitaire noté m(k). La similarité entre une donnée et une classe se calcule par le produit scalaire <m(k), x > qui représente le cosinus de l'angle (x,m(k)).

**D'ou l'algorithme de classification K Means Axiales**

**[0]** Initialiser au hazard K axes m(k) de classes:

$m^0(k)= [\mu_1^0(k), \mu_2^0(k), \ldots, \mu_I^0(k)]$ avec $||m^0(k)||=1$

Ainsi que les K valeurs $\tau(k)$: $\tau^0(k)=0$

**[1]** Pour chaque ligne x du tableau $x = [\xi_1, \xi_2, \ldots, \xi_I]$ où $||x||=1$

- Calculer les K projections:

  $\eta(k)= <m^t(k), x >$

- Incorporer x à la classe <u>k</u> pour laquelle cette projection est maximale et mettre à jour la position de l'axe m(k):

  $m^t(k)= m^{t-1}(k)+( \eta(k) /\tau^t(k))(x-\eta(k) m^{t-1}(k))$

  avec $\tau^t(k)= \tau^{t-1}(k)+ (\eta(k))^2$

**[2]** Après épuisement des T lignes : fin

## 6- Version non adptative (iterative) de la méthode K means axiales

Comme il a été déjà suscité, la méthode de Mac Queen offre des résultats satisfaisants mais ne garantie pas la décroissance du critère $\phi$. En 1965 **[4]**, Forgy a montré qu'une version non adaptative de cette méthode offre cette garantie. D'ou l'idée d'Alain Lelu de proposer une version iterative ou non adaptative de la méthode K Means Axiales. Cette version se présente comme suit:

**L'algorithme iteratif K means axiales**

**[0]** Initialiser au hazard K axes m(k) de classes:

$m^0(k)= [\mu_1{}^0(k), \mu_2{}^0(k), \ldots, \mu_I{}^0(k)]$ avec $||\, m^0(k)\,|| =1$

Ainsi que les K valeurs $\tau(k)$: $\tau^0(k)=0$

**[1]** Pour chaque ligne x du tableau $x = [\, \xi_1, \xi_2, \ldots, \xi_I]$ où $||\, x\,|| =1$

- Calculer les K projections:

  $\eta(k) = <m(k), x>$

- Incorporer x à la classe $\underline{k}$ pour laquelle cette projection est maximale.

**[2]** Après épuisement des T lignes :

- Cacluler le critère $\tau = \sum_k \tau(k)$ avec $\tau(k) = \sum_{x \in \text{classe } k} < m(k),x>$
- Si la croissance de ce critère est inferieure à un certain seuil : fin.

  Sinon mettre à jour la position de l'axe m(k):

  $m^t(k)= m^{t-1}(k)+( \eta(k)/\tau^t(k))(x-\eta(k)m^{t-1}(k))$

  aller en 1.

Lelu a montré que la croissance du critère $\tau$ est positive ou nule lorsque les passages sont faits sur un ensemble fini de données.

## 7- Application de la méthode K means axiales pour le traitement des données documentaires

### 7.1- Spécifités des données documentaires

Les données documentaires sont cacartérisées par les spécifités suivantes:

- *Elles sont volumineuses:* Les bases de données bibliographiques recensent de nos jours un volume très important de références bibliographiques.
- *Elles sont de dimentionnalité considérable:* Les vocabulaires d'indexation dans les bases documentaires s'évaluent en milliers de termes.
- *Elles sont très diluées:* Les tableaux croisant les références bibliographiques aux termes d'indexation sont extrêmement vides du fait que chaque référence est représentée par une très faible proportion de termes par rapport au vocabulaire global.
- *Elle sont de type pick-any:* Expréssion d'origine englo-saxonne qui désigne que le volume de mots clés choisis par référence bibliographique varie en fonction de l'indexeur et de la politique d'indexation adoptée.

### 7.2- Choix de la méthode K means axiales

Les spécifités suscitées des données documentaires et textuelles nous ont amenés à choisir les modèles neuronaux en général et la méthode de classification K Means Axiales, en particulier, car ils conviennent le mieux à ce type de données comparés aux autres méthodes de classification.

### 7.3- Principe de la méthode pour la classification des données documentaires

Pour l'application de cette méthode sur les données documentaires, l'ensemble des documents (références bibliographiques) est considéré comme un nuage de points dans l'espace géométrique où chaque dimension correspond à un mot clé. Les classes thématiques induites à la suite de son application sont sous forme de demi-axes passant par l'origine et pointant vers les zones de forte densité. D'une manière plus simple, chaque référence bibliographique correspond à un vecteur unitaire (contenant que des zéros et des uns ) dit vecteur-référence. L'$i^{ème}$ coordonnée de ce vecteur est égale à "un" si le $i^{ème}$ mot clé indexe le document correspondant à cette référence et égale à "zéro" dans le cas contraire.

Aussi, chaque classe thématique correspond à un vecteur (dit vecteur-classe) pointant vers une zone de forte densité.

L'application de la méthode K Means Axiales pour la classification thématique des données bibliographiques passe par les étapes suivantes :

### 7.4- Etapes de la méthode

**Etape1 : Initialisation**

- du nombre de classes K,
- des K vecteurs-classes par les K premiers vecteurs-références,
- du seuil d'apprentissage,
- du seuil de typicité des documents (reps. du seuil de typicité des mots clés).

**Etape2 : Projection des références**

Chaque document i normé est projeté sur les K demi-axes. La projection (orthogonale) $\eta_i(k)$ correspond au produit scalaire entre le vecteur-référence correspondant au document i et le vecteur-classe m(k).

**Etape3 : affectation des documents**

Chaque document i est affecté à la classe $\underline{k}$ pour laquelle sa projection $\eta_i(\underline{k})$ est maximale.

Avant de passer à la phase suivante de la méthode, nous tenons à préciser que les étapes 2 et 3 présentées ci-dessus correspondent à une itération de la méthode K Means Axiales.

**Etape 4 : Test de stabilité et mise à jour des classes thématiques**

Un test de stabilité doit être effectué afin de décider de l'arrêt de l'algorithme ou du passage à une autre itération après la mise à jour de la position des demi-axes.

**Etape 5 : Détermination des classes**

- Un document i appartient à une classe k' si sa projection orthogonale sur le demi-axe correspondant est supérieure au seuil de typicité des documents.
- Un mot clé j appartient à une classe k' si la $j^{eme}$ coordonnée du vecteur correspondant à cette classe est supérieure au seuil de typicité des mots clés.

**Remarques**

1- La méthode K Means Axiales permet de construire des classes recouvrantes.
C'est à dire que:
   - Chaque référence peut appartenir à plusieurs classes.
   - Chaque mot clé peut appartenir à plusieurs classes.

2- Dans chaque classe thématique les mots clés peuvent être pondérés et ordonnés en fonction des coordonnées de cette classe sur les axes qui correspondent aux mots clés. Ainsi, l'intitulé de chaque classe thématique peut être représenté par le mot clé correspondant à la plus grande coordonnée. Aussi, les documents d'une même classe peuvent être pondérés et ordonnés en fonction de leurs projections orthogonales sur cette classe.

**7.5- Algorithme détaillé de la méthode**

Notons par:

T : nombre de références bibliographiques.

I: nombre des mots clés indexant les références.

K: nombre des classes thématiques.

$x_1, x_2, x_3, \ldots\ldots, x_T$: les T vecteurs-références de dimension I.

$m_1, m_2, m_3, \ldots\ldots, m_K$: les K vecteurs-classes de dimension I.

$\eta$ : matrice de projection sur les axes (de dimension $T \times K$).

$\eta_{max.}$ : vecteur des projections maximales (de dimension K.)

F: Matrice d'affectation des références de dimension ($T \times K$.)

RM: matrice de couverture des références par les classes, de dimension $K \times T$.

R: matrice de couverture des mots clés par les classes, de dimension $K \times I$

***{initialisation}***

**(0)** Pour i=1…T faire

pour j=1…I faire

$x_i[j] = \begin{cases} 1 & \text{si la référence i est indexée par le mot j} \\ 0 & \text{si non} \end{cases}$

Pour k=1….K faire

$\tau[k]=0$

pour j=1… I faire

$m_k[j]= x_k[j]$

$\underline{\tau} =0$ , $\theta =0$

***{Calcul des projections}***

**(1)** Pour i=1…T faire

Pour k=1…K faire

$\eta[i,k]= \langle x_i , m_k \rangle$

***{Calcul des projections maximales}***

**(2)** Pour i=1…T faire

$\eta_{max.}[i]=$ Maximum de $\eta[i,k]$

$1 \leq k \leq K$

***{ Affectation des documents }***

**(3)** Pour i=1…T faire

Pour k=1…K faire

Si $\eta[i,k] = \eta_{max.}[i]$ alors

$F[i,k]= 1$

sinon $F[i,k] =0$

***{ test de  stabilité et mise a jour des vecteurs-classes }***

**(5)** Pour k=1...K  faire

Pour i =1 to T faire

Si F[i,k] =1 alors $\tau[k]=\tau[k]+<m_k,xi>^2$

Pour k=1...K  faire $\underline{\tau}= \underline{\tau}+\tau[k]$

**Si** $\underline{\tau} - \theta >$ seuil  d'apprentissage alors

Pour k=1…K faire

Pour i=1..T faire

Si F[i,k]=1 alors

$m_k = m_k + (\eta\ [i,k] / \tau[k])\ (x - \eta[i,k]mk)$

aller en **(1)**

***{Détermination des classes recouvrantes }***

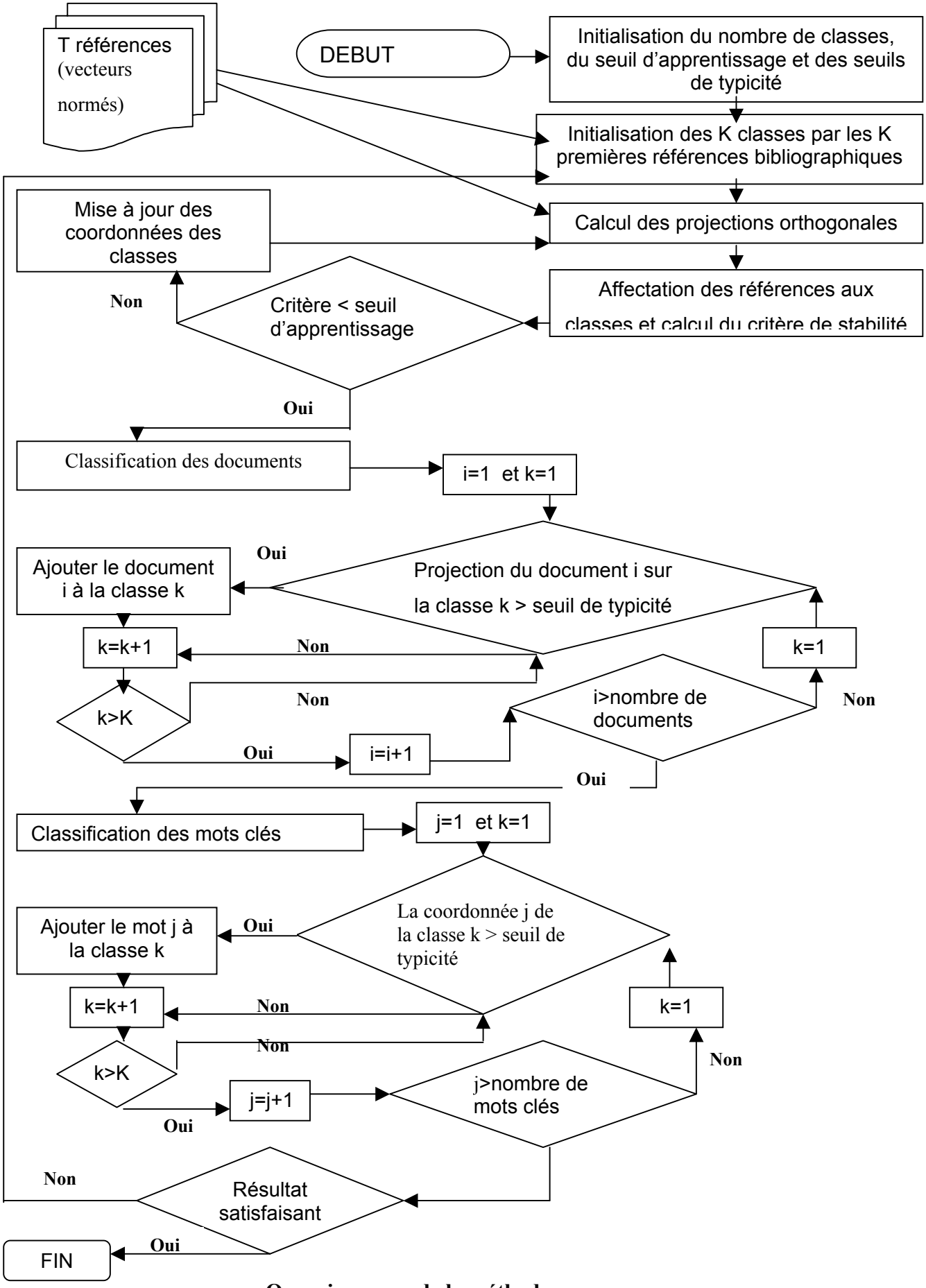
**(6)**Pour i=1…T et k=1…K faire

$$RM[k,i]=\begin{cases} 1 & \text{si } \eta[i,k]\geq \text{seuil de typicité} \\ 0 & \text{si non} \end{cases}$$

Pour j=1…I et k =1…K faire

$$R\ [k,j]=\begin{cases} 1 & \text{si } m_k[j]\geq \text{seuil de typicité} \\ 0 & \text{si non} \end{cases}$$

## 7.6- Organigramme de la méthode




**Organigramme de la méthode**

### 7.7- Conclusion

L'application de la méthode K Means Axiales sur les références disponibles dans une base de données bibliographiques permet de dégager l'ensemble des classes thématiques abordées dans cette base.

Chaque classe thématique ou simplement thème est relié à un ensemble de références et un ensemble de mots clés.

Comme il a été suscité, les mots clés ainsi que les références, d'une même classe, sont pondérés et ordonnés. L'intitulé de la classe thématique correspond au mot ayant le plus grand poids.

En plus, les auteurs des références d'une même classe sont, également, associés à cette dernière.

## III- L'analyse de données pour la représentation des classes thématiques

### 1- Analyse de données

L'analyse de données est une technique relativement récente, constituée essentiellement dans la décennie 1960-1970. Elle consiste en la manipulation de grand volume de données afin de parvenir à des conclusions bien fondées. Sachant de toute réalité est un résultat complexe de très nombreux facteurs, l'analyse de données permet de raisonner sur un nombre quelconque de variables, d'où le nom analyse multivariée. Elle consiste en deux démarches principales, la classification automatique de variables statistiques et l'analyse factorielle utilisant les propriétés des espaces vectoriels pour la description des individus et des variables.

Elle a été présentée par DODGE dans son dictionnaire encyclopédique en statistique comme suit : "De nombreuses activités scientifiques commencent par un recueil de données, qu'on peut généralement classer dans un tableau à double entrée de grande taille. L'analyse de données vise à permettre à l'utilisateur de ces tableaux d'extraire facilement le nombre d'informations qui lui sont nécessaires……. Elle regroupe des méthodes très nombreuses et très différentes d'analyse statistique"[13].

Parmi ces méthodes, existe la méthode d'analyse en composantes principales. Cette méthode se présente comme suit :

### 2- Analyse en composantes principales

L'analyse en composantes principales est une méthode puissante et très utilisée pour l'exploration de la structure des données multidimensionnelles. Elle permet d'obtenir une représentation approchée des données sur un espace de dimension faible comparée à la dimension d'origine.

Pour décrire le principe de cette méthode, les éléments suivants sont introduits (notons que les natations utilisées, dans cette partie, sont celles de B. SAPORTA [19] ).

## 2.1- Tableau de données

Les données multidimensionnelles sont regroupées dans un tableau X à n lignes et p colonnes. Il comporte les observations de p variables sur n individus. Ainsi, chaque ligne i du tableau est associée à un individu et chaque colonne j à une variable. L'élément xij sur la ième ligne et la jème colonne correspond à la valeur prise par la jème variable sur le ième individu.

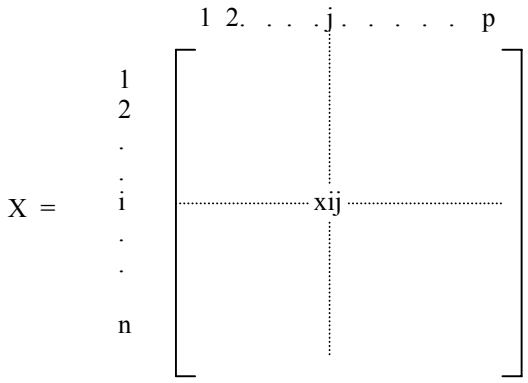



**Tableau de données**

Chaque variable est identifiée grâce à un vecteur noté xj, de dimension n, rassemblant les valeurs de cette dernière sur les n individus. Ce vecteur correspond à la $j^{eme}$ colonne de X. De même, les individus sont identifiés grâce aux vecteurs ei de dimension p correspondant aux lignes de X.

## 2.2- Pondération des individus

En pratique, les individus ne sont pas tous, nécessairement, de la même importance. D'où la nécessité d'introduire des poids reflétant cette différence d'importance. Ces poids sont rassemblés dans une matrice D diagonale de taille n. L'élément Dii de cette matrice correspond au poids pi du ième individu. Ces poids vérifient :

- pi positif pour tout i,
- la somme des pi est égale à 1.

## 2.3- Centre de gravité du nuage des individus

Le centre de gravité du nuage des points, correspondant aux individus, n'est rien d'autre qu'un point G (vecteur de taille p) regroupant les moyennes arithmétiques des p variables.

**2.4- Tableau centré des données**

A un tableau X de données correspond un tableau Y de données centrées.

Ce tableau vérifie, pour tout i et tout j:

$Yij = Xij - Gj$

Ainsi :

$Y = X - \underline{1}\, G'$

tel que $\underline{1}$ est le vecteur de taille n dont tous les éléments sont égaux à 1 et G' le vecteur transposé de G.

**2.5- Matrice de variance covariance des variables**

La matrice V de variance-covariance correspondante au tableau X est définie par :

$V = X'DX - GG' = Y'DY$

telle que X' correspond à la matrice transposée de X et Y' à celle de Y.

**2.6- Matrice de corrélation**

La matrice regroupant les coefficients de corrélation entre les p variables est notée R.

Elle est donnée par :

$$R = D_{1/S} V D_{1/S}$$

telle que $D_{1/S}$ est la matrice diagonale des inverses des écarts-types.

**2.7- Tableau centré réduit des données**

Au tableau de données X correspond le tableau Z centré et réduit. Ce tableau vérifie pour chaque couple d'indices (i , j):

$Zij = (Xij - Gj) / Sj$

Sj est l'écart-type de la $j^{ème}$ variable.

**2.8- Espace vectoriel des individus**

Les individus sont considérés comme des points (vecteurs ) d'un espace vectoriel noté F. Le centre de gravité du nuage de points est le point défini par le vecteur G. Afin de mesurer les distances entre les points du nuage, l'espace vectoriel des individus F est muni d'une métrique notée M.

La métrique usuelle est égale à I (matrice identité). Néanmoins la métrique utilisée en statistique est autre que cette dernière. Elle est donnée par $M = D_{1/S^2}$ (matrice diagonale des inverses des variances). Cette métrique permet de résoudre le problème de la différence des unités de mesure des variables. De plus, elle représente le choix par défaut dans les logiciels d'ACP.

Ainsi la distance entre deux individus définis par les vecteurs ei et ej est donnée par :

$$d(ei,ej) = (ei-ej)'M(ei-ej)$$

Et leur produit scalaire par :

$$< ei,ej > = ei'Mej$$

**2.9- L'inertie totale du nuage des individus**

L'inertie totale du nuage des points correspond à la moyenne pondérée des carrées des distances des points au centre de gravité. Ainsi

$$Ig = \sum^{n}_{i=1} pi\ (ei-G)'M(ei-G)$$

L'inertie est également donnée par la formule suivante :

Ig = trace (MV)

**2.10- Espace des variables**

Comme pour les individus, les variables sont considérées comme des vecteurs (xj) d'un espace vectoriel de dimension n. Cet espace est noté E. Aussi, l'étude de proximité entre les variables nécessite la définition d'une métrique. La métrique communément utilisée dans ce cas est la matrice diagonale des poids D.

Ainsi le produit scalaire dans l'espace E est donné par :

$$<xk,xj>=xk'Dxj=\sum i\ pi\ xik\ xij$$

Il correspond à la covariance Skj entre ces deux variables si elle sont centrées.

De même la norme d'une variable xi selon la métrique D est donnée par :

$$||xi||_D^2 = Si^2$$

Et l'angle $\theta kj$ entre les deux variables xk et xj vérifie :

$$\cos \theta kj = < xk,xj> / ||xk||\ ||xj|| = Skj / ( Sk\ Sj )$$

Ce qui correspond au coefficient de corrélation des deux variables. xk et xj

Chaque vecteur xj de l'espace des variables correspond à un axe dans l'espace des individus. La combinaison linéaire des variables permet d'engendrer une nouvelle variable qui définie, à son tour, un axe dans l'espace des individus F. Soit $\Delta$ un axe de F engendré par un vecteur unitaire a normé (selon la métrique M). Les coordonnées des variables, notées ci, sur ce nouvel axe sont obtenues par projection orthogonale sur ce dernier.

Ainsi, pour tout individu i :

$$ci = < a,ei>= a'Mei$$

Ces coordonnées sont regroupées dans un vecteur c de taille n qui permet de former une nouvelle variable. D'où

$$c = X\ M\ a$$

En posant u = Ma on aura : c = Xu

La variance de cette nouvelle variable est donnée par :

$$V(c) = c'D\ c = u'X'DXu= u'\ V\ u$$

Par conséquent, à chaque nouvelle variable est associée :

- Un axe $\Delta$ de vecteur unitaire a,
- Un vecteur c de l'espace E,
- Et une forme linéaire u appelée facteur.

**3- Principe de la méthode ACP**

Comme c'est déjà suscité, l'objectif principal de cette méthode est de représenter le nuage des individus dans un espace de dimension plus faible que le nombre initial des variables.
Cet espace est choisi de façon à déformer le moins possible les distances entre les projections des points sur ce dernier. Sachant que la procédure de projection implique nécessairement une réduction de la somme des distances, le sous espace recherché, noté Fk, est celui réalisant une inertie maximale.

La projection des individus sur le sous espace Fk se fait grâce à un opérateur P qui n'est autre qu'une matrice vérifiant :

$$P^2=P \quad \text{et} \quad P'M=MP$$

Le nuage projeté est donné par le tableau XP'. La variance de ce tableau, dans le cas de variables centrées, est donnée par :

$$(XP)'D\,XP= P'VP$$

Ainsi, l'inertie du nuage projeté est donnée par :

trace ( P'VPM)= trace ( VMP ) ( car trace (AB)= trace (BA) et P2=P )

Il suffit juste de démontrer que trace (AB)= trace ( BA)

Sachant que la trace d'une matrice carrée est la somme des éléments diagonaux de cette matrice, supposons que L(m×m) est la matrice carrée AB et T (n×n) la matrice carrée BA

alors :

$$\text{trace (AB )} = \text{trace (L)} = \sum_{i=1}^{m} Lii$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} AijBji$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{m} BjiAij$$

$$= \sum_{j=1}^{n} Tjj = \text{trace ( T) = trace ( BA )}$$

Ainsi, il faut chercher P réalisant trace (VMP) maximale.

**Théorème ( Saporta, 1990 [19])**

Soit Fk un sous espace portant l'inertie maximale, alors le sous espace de dimension k+1 portant l'inertie maximale est la somme directe de Fk et du sous espace de dimension 1 M-orthogonale à Fk portant l'inertie maximale : Les solutions sont emboîtées.

Ainsi la détermination du sous espace recherché peut se faire de proche en proche. C'est à dire qu'il faut commencer par la recherche d'un sous espace de dimension 1 réalisant l'inertie maximale. Ensuite, d'un sous espace, de dimension toujours égale à 1, M-orthogonale au premier et portant l'inertie maximale….etc.

**3.1- Axes principaux**

D'après ce qui précède, il faut commencer par définir une droite réalisant l'inertie maximale. Cette droite correspond à un axe définit grâce à un vecteur unitaire, noté a.

Le projecteur M-orthogonale P sur cette droite vérifie : $P= a(a'Ma)^{-1} a'M$

Ainsi l'inertie du nuage est donnée par :

$Ig = \text{trace}(VMP)=\text{trace}(VM\, a(a'Ma)^{-1} a'M)$

$= \text{trace}(VM\, aa'M) / (a'Ma)$ car $a'Ma$ est un scalaire

$= \text{trace}(a'MVMa) / (a'Ma)$ car $\text{trace}(AB)=\text{trace}(BA)$

$= (a'MVMa) / (a'Ma)$ car $a'MVMa$ est un scalaire

Maximiser l'inertie totale revient à annuler la dérivée par rapport à a d'où :

$$[(a'MVMa) / (a'Ma)]' = [ (a'Ma)2MVMa-(a'MVMa)2Ma] / (a'Ma)^2 = 0$$

ainsi

$$MVMa = [ (a'MVMa) / (a'Ma)]Ma$$

d'où :

$$VMa = [ (a'MVMa) / (a'Ma)]a = \lambda\, a$$

Puisque M est régulière, a n'est rien d'autre qu'un vecteur propre de VM. Mais dans le but de maximiser l'inertie (égale à $\lambda$) a est le vecteur correspondant à la plus grande valeur propre. Plus encore, vu que VM est M-symétrique, ses vecteurs propres sont M-ortogonaux d'où le théorème.

**Théorème ( Saporta, 1990 [19])**

Le sous espace Fk de dimension k est engendré par les k vecteurs propres de VM associés aux k plus grandes valeurs propres.

### 3.2-Facteurs principaux

Comme c'est déjà suscité, à chaque axe défini par le vecteur unitaire a est associé un facteur u vérifiant u= Ma

Puisque :

$$VMa=\lambda a$$

alors

$$MVM\ a = \lambda\ Ma$$

ainsi

$$MVu = \lambda u$$

Par conséquent, les facteurs principaux sont les vecteurs propres de MV, $M^{-1}$- normés.

### 3.3- Composantes principales

A chaque facteur $u_i$ correspond une composante principale $c_i$ vérifiant :

$$ci = X\ ui$$

Le vecteur $c_i$ regroupe les cordonnées des individus sur l'axe défini par $a_i$.

**Remarque**

L'usage de la métrique $D_{1/S}$ correspond à diviser les valeurs prises par les variables sur les individus par les écarts-types des variables. Ce qui correspond à travailler sur la base du tableau centré réduit Z au lieu de X et d'utiliser la métrique usuelle I.

4- L'Analyse en composantes principales pour la représentation des classes thématiques

L'objectif de l'utilisation de l'ACP est la représentation des classes thématiques, obtenues à la suite de l'application de la méthode K Means Axiales, sur un plan constituant ainsi la "carte thématique". Cette carte permet de visualiser les rapprochements et les interactions entre ces classes. A Chacune des classes est affecté un ensemble de mots clés. Ces affectations sont représentées par le tableau R (matrice de couverture des mots clés par les classes)[6]. Les lignes de ce tableau représentent les classes thématiques alors que les colonnes correspondent aux mots clés.

$$R[k, j] = \begin{cases} 1 & \text{si mot j est dans la classe k} \\ 0 & \text{si non} \end{cases}$$

Ainsi, chaque classe correspond à un individu et chaque mot clés à une variable.

[6] Voir : Application de la méthode K means axiales pour le traitement des données documentaires: Algorithme détaillé de la méthode

## IV- Distribution Ziphienne pour le découpage de l'information

Comme c'est déjà suscité, les données bibliographiques, en général, et l'ensemble des mots clés en particulier, sont caractérisés par leur volume très important. Cette caractéristique constitue un problème majeur pour la classification thématique dans une base de données bibliographiques. Pour la résolution de ce problème, la solution la plus adaptée et la plus utilisée est la méthode de découpage des fréquences faibles. Cette méthode se base sur le principe que la fréquence d'un mot dans la base reflète son importance par rapport aux thématiques de la base. Avant de proposer la méthode de découpage, il est indispensable d'introduire :

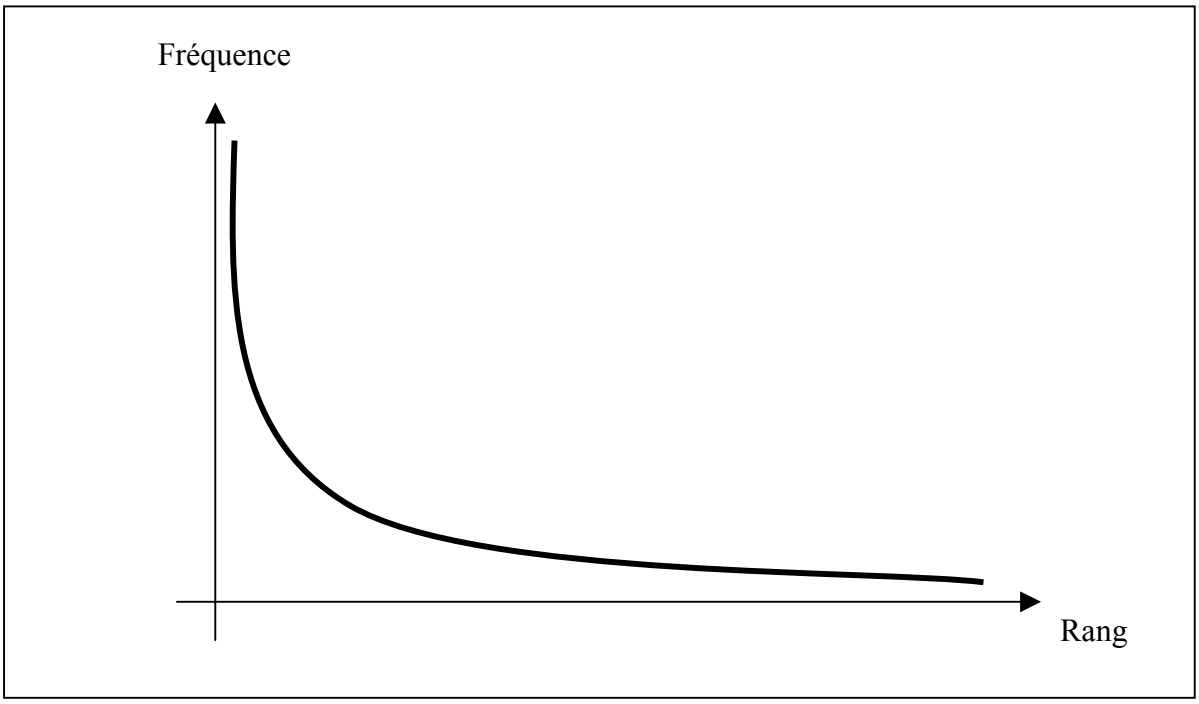
### 1- Distribution Ziphienne

En 1949 [16], G.K. Ziph a constaté une certaine régularité sur la fréquence d'apparition des mots. Ainsi, après rangement des mots en fonction de leurs fréquences décroissantes, il a pu définir une distribution entre le rang et la fréquence. Il existe dans la littérature plusieurs formulations de cette distribution. Parmi ces dernières, Lafrouge[16] a donné la formulation suivante :

$$G(r)=K/r^{a}$$

telle que :
G(r) désigne le nombre d'occurrences de la forme de rang r,
K et a des constantes.

Cette distribution est de type hyperbolique et possède une longue queue. Elle se présente comme suit :



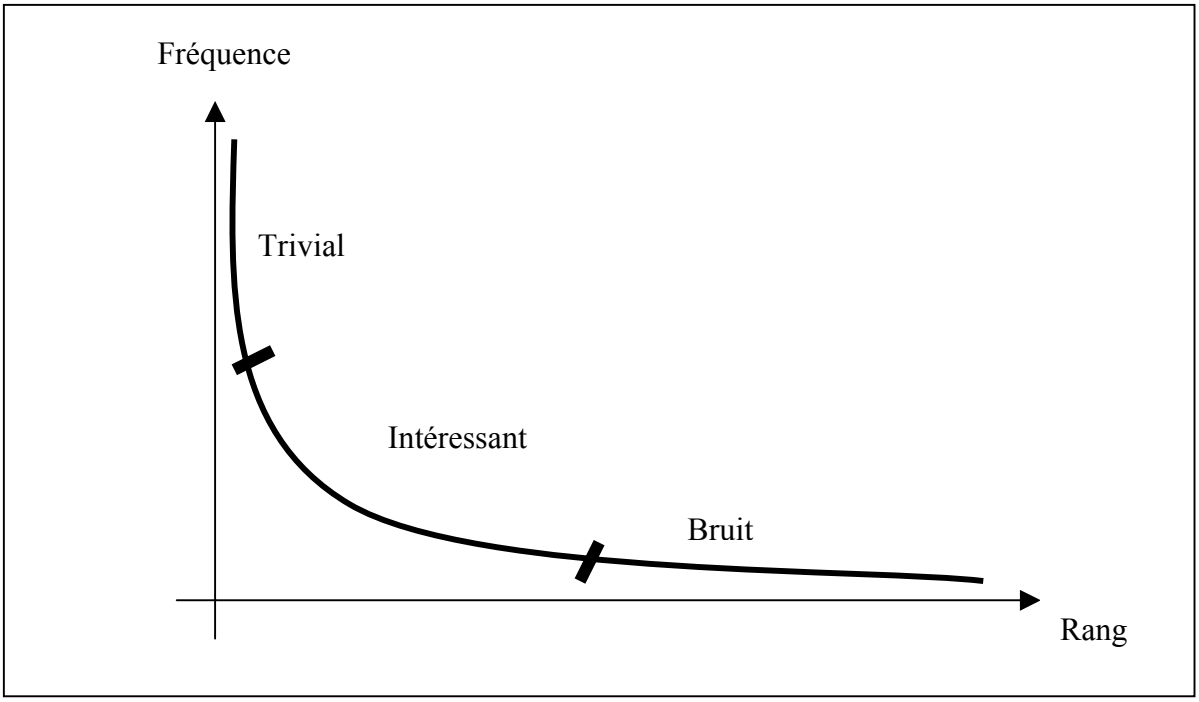

**Représentation de la distribution Ziphienne**

Cette courbe est découpée [18] en trois zones :

**Zone 1** : information triviale: ensemble de mots clés triviaux en fonction de la thématique de la base .
Exemple : Dans une base de données spécialisée en informatique, le mot clés «informatique» appartient à cette zone.

**Zone 2** : information intéressante : ensemble des mots clés représentatifs du contenu de la base.
Exemple : Toujours dans une base spécialisée en informatique le mot clés «intelligence artificielle » appartient, logiquement, à cette zone.

**Zone 3** : Information marginale ou bruit : ensemble de mots clés non pertinents par rapport aux thématiques de la base.
Exemple : un mots clés « Annaba » par exemple peut être considéré comme un bruit dans une base en informatique.




**Découpage en zones de l'ensemble des mots clés[16]**

**2- Découpage en zones**

Le découpage d'une distribution symétrique se base sur la moyenne et l'écart type. Or, la distribution Ziphienne n'est pas symétrique d'où la difficulté de déterminer les indicateurs de découpage (représentant des rangs délimitant les zones ).

La solution de ce problème a été présentée par Lafouge [16] comme suit :

Soient :

r1 l'indicateur de coupure zone1/zone 2

r2 l'indicateur de coupure zone2/zone 3

alors :

r1= partie-entière ($D_2$)

r2= partie-entière ($D_{1/2}$)

avec :

$D2 = 1/B$ $\quad B = \sum_{i=1}^{I} (pi)^2$

$D1/2 = E^2$ $\quad E = \sum_{i=1}^{I} (pi)^{1/2}$

où I désigne le nombre total de mots clés et pi la probabilité d'apparition du mot clés i.

Dans ce travail, la probabilité pi est calculée grâce à la formule suivante :

$pi = fi / (\sum_{j=1}^{I} fj)$

**3- Conclusion**

L'utilisation de la distribution Ziphienne permet d'obtenir une bonne réduction du volume de l'information traitée. En effet, à travers la courbe de cette distribution, nous constatons le volume important de l'information bruit comparée à l'information intéressante ou triviale. Ainsi, la troncature de cette information permet une grande réduction du volume tout en gardant l'essentiel de l'information.

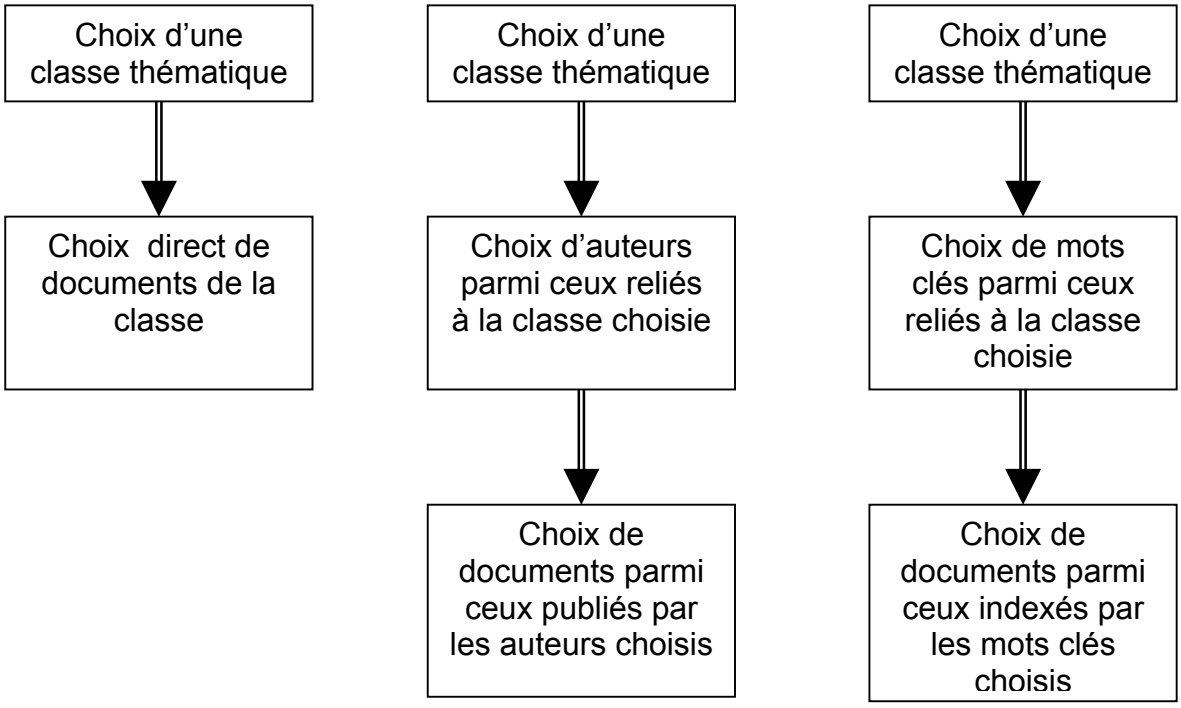
# Partie III

## Réalisation d'un prototype de système de recherche d'information en se basant sur la classification thématique et la cartographie des données

## I- Présentation du système

Le système proposé sera appelé METRISYS. L'Acronyme ***MetriSys*** est constitué de deux parties, la partie *Metri* pour designer la mesure et *Sys* pour système. L'objectif principal de ce système est de permettre une recherche guidée dans la base de données bibliographiques. Il se base principalement sur la *classification thématique* qui permet d'identifier l'ensemble des thèmes (classes thématiques) présents dans la base. Chaque thème est relié à un ensemble de mots clés, d'auteurs et de références bibliographiques. De plus, le système permet une *représentation graphique* des classes thématiques sur un espace bidimentiennel de façon à visualiser leurs rapprochements. Cette représentation des thèmes sur un plan constituera la *carte thématique.*

Ce système offre à l'utilisateur un nouveau modèle pour la recherche d'information. Ce modèle se base sur les classes thématiques pour guider l'utilisateur à l'information pertinente selon les modes suivants :




**Modes de recherche d'information dans le système METRISYS**

Il faut juste noter, que les deux derniers modes sont plus intéressants comparés au premier car ils permettent de réduire l'espace de recherche sur la base de la sélection de mots clés ou d'auteurs.

En plus de cette nouvelle approche, l'utilisateur dispose également de l'approche classique de recherche d'information dans laquelle il propose un ensemble de mots clés à partir desquels se fait la recherche.

Ce système englobe également d'autres traitements relativement simples permettant de dégager des informations élaborées à partir de la base de données ; comme le tri des formes[7] en fonction de leur fréquence ainsi que leur découpage en informations pertinentes, triviales et bruit. A partir de ces traitements plusieurs indicateurs peuvent être déduits. Ces indicateurs sont dits «indicateurs bibliométriques». Parmi ces derniers existe : l'indice de productivité d'un auteur, d'un pays, …etc.

## II- Facteurs à prendre en compte pour la réalisation du système

Pour la réalisation de notre système plusieurs facteurs doivent être examinés :

### 1- Diversité des systèmes de gestion des bases de données

Il existe une multitude de systèmes de gestion de bases de données bibliographiques. Ces systèmes permettent la saisie, la modification, l'importation, l'exportation, … dans les bases de données bibliographiques. D'où l'idée de concevoir un système susceptible de traiter les données indépendamment de leur système de gestion. Par conséquent, METRISYS doit être compatible avec les différents systèmes de gestion disponibles.

### 2- Diversité de la structure des données

La diversité des systèmes de gestion de bases de données induit, nécessairement, une diversité de leur structure logique(ordre logique des données : titre, auteur, date d'édition,…etc.) ainsi que la structure physique des données (.txt, .iso, .db,…etc.). Le système METRISYS doit prendre en considération cet aspect en permettant la reconnaissance de ces différentes structures.

[7] Unités d'information : auteurs, mots clés, pays d'édition…etc.

**3- Volume important des données**

Le système METRISYS doit permettre le traitement d'un volume important de données.

**4- Diversité des traitements à effectuer par le système**

Le système METRISYS doit englober les différents traitements suscités : la classification, la cartographie, le tri, le découpage des données et la représentation graphique.

III- Traitements effectués par le système

Le système METRISYS comporte plusieurs fonctionnalités à savoir :

**1 - Chargement des données**

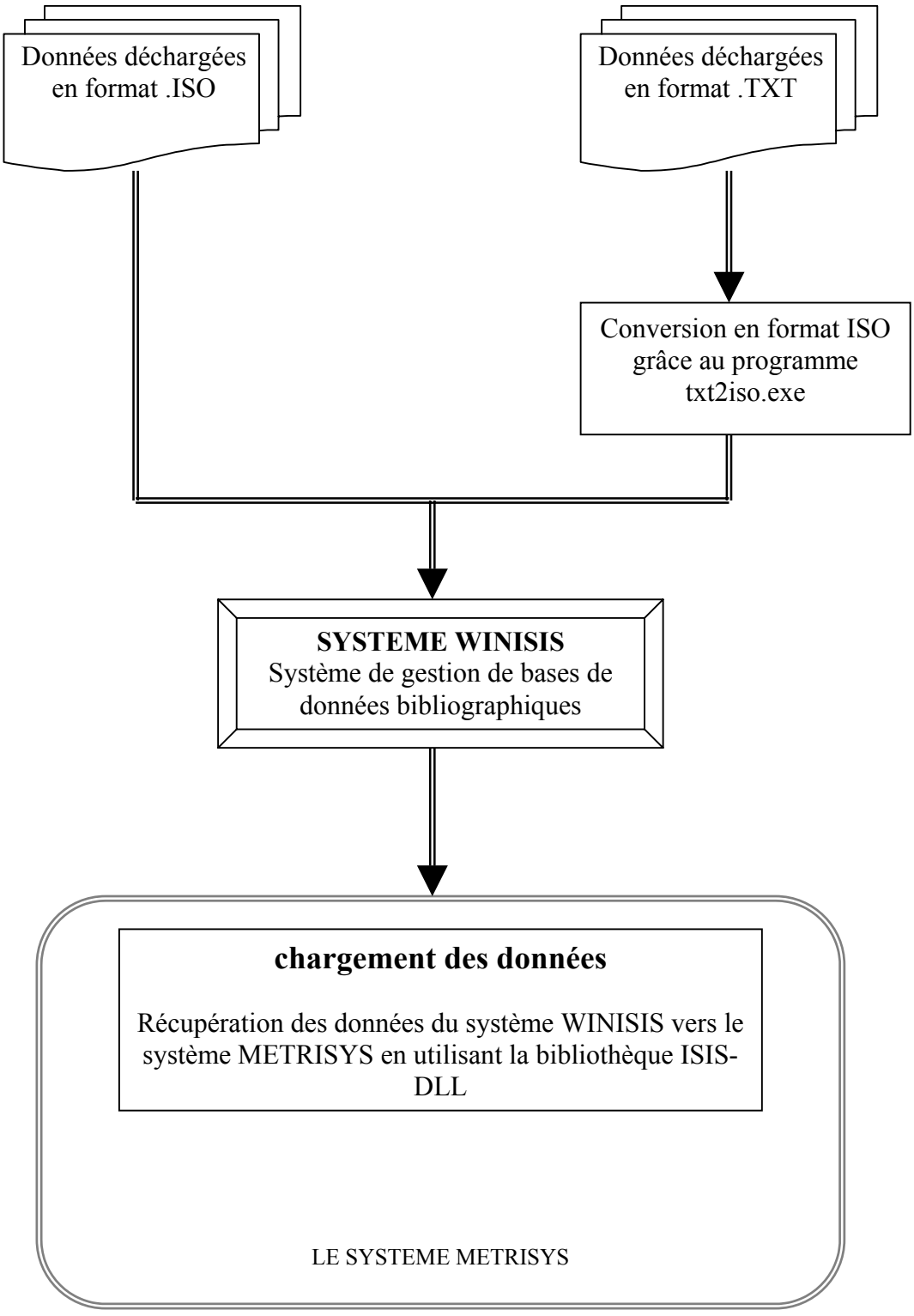
Les données bibliographiques sont récupérées par déchargement de fichiers on-ligne ou off-ligne, d'où la nécessité de reconnaissance de leurs structures physique et logique.

Concernant la structure physique des données, la majorité des fichiers sont déchargés en format **TXT** ou **ISO**. Néanmoins, les données ayant un format TXT peuvent être converties en format ISO grâce à un programme appelé txt2iso.exe. Par la suite, les données en format ISO sont intégrées au système WINISIS[8] qui permet de gérer ce type de données. Ainsi et grâce à une bibliothèque de fichiers DLL, appelée ISIS-DLL, qui constitue une interface entre le système WINISIS et tout autre système développé sous l'environnement DEPLHI ou C++ , notre système peut reconnaître la structure physique des données. Sachant qu'il est développé sous l'environnement DELPHI.

Concernant la structure logique des données, METRISYS peut reconnaître les séparateurs logiques présents dans le format ISO toujours en utilisant les fonctionnalités de l'interface ISIS-DLL.

Cette fonctionnalité se schématise comme suit :

[8] Système de gestion de bases de données bibliographiques.




**Chargement des données bibliographiques**

**2- Génération de données descriptives**

METRISYS permet la génération de tableaux de fréquences des formes et le tri de ces derniers en fonction de leurs fréquences.

**3- Découpage des données**

Comme c'est déjà suscité et selon la loi de Ziph, les informations pertinentes constituent une partie restreinte de l'ensemble des données. Ainsi, le système METRISYS permet de dégager ces informations en utilisant le découpage des données en trois zones : triviales, intéressantes et bruit. Cette restriction des données induit un gain important de temps et d'espace de traitement.

**4- Classification thématique**

La plus importante fonctionnalité du système est la **classification thématique**. Elle se base sur la méthode K Means Axiales qui permet la génération des classes thématiques couvertes par la base de données.

Les résultats sont stockés dans un fichier contenant les informations suivantes pour chacune des classes:

- l'intitulé de la classe (thème),
- liste des mots clés associés à la classe,
- liste des références associées,
- liste des auteurs associés,
- coordonnées de la classe sur la carte (déterminées par la suite).

**5- Représentation des classes thématiques sur une carte**

A la suite de l'application de la méthode d'analyse en composantes principales, les coordonnées de chaque classe sur un espace bidimentionnel sont déterminées. Ces dernières sont enregistrées dans le fichier suscité avec les autres informations relatives aux classes thématiques. Ainsi, les classes sont représentées sur un plan constituant ainsi **la carte thématique.**

**6- Recherche d'information**

Le logiciel propose à l'utilisateur deux approches de recherche.

**-** La première, dite classique, passe par les étapes suivantes :

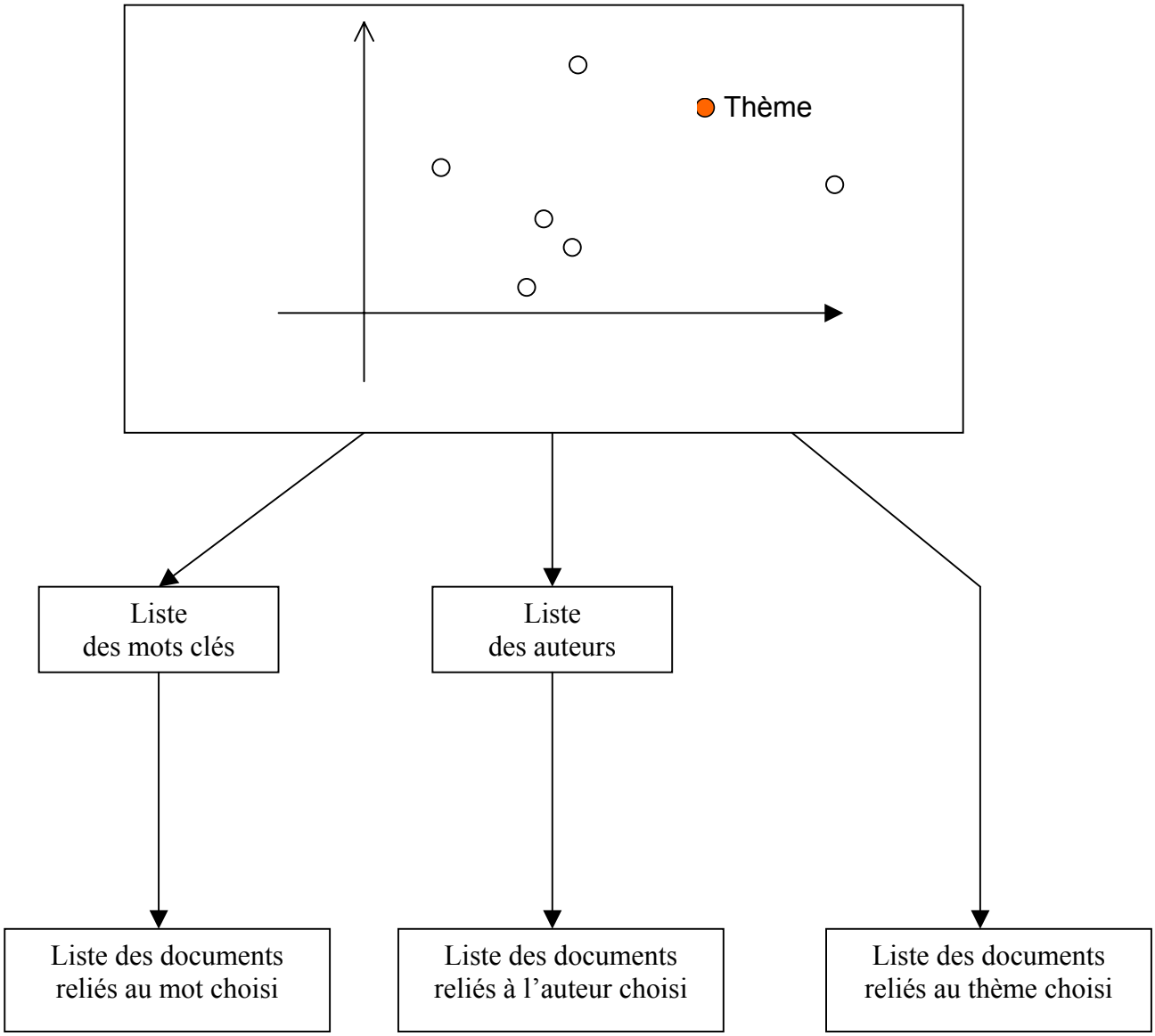
- Le système offre un index de mots clés ou d'auteurs.
- L'utilisateur choisi le mot clés ou l'auteur qui l'intéresse.
- Le système induit l'ensemble des références reliées à ce mot ou cet auteur choisi.

**-** La deuxième approche se base sur les classes thématiques et propose à l'utilisateur deux interfaces de recherche.

**i-** La première interface est constituée principalement de la carte thématique. Chaque point de la carte (qui correspondant à un thème) est un lien **hypertexte** permettant l'accès à la liste des mots clés, des auteurs ou des références associés à ce dernier. Ainsi l'utilisateur peut choisir la liste qui l'intéresse, ensuite, l'élément qui l'intéresse de la liste pour arriver enfin aux données bibliographiques. Autrement dit, l'utilisateur exploite la carte thématique présentée sous un format hypertexte pour naviguer dans la base bibliographique et faire des recherches.

**ii-** La deuxième interface de recherche suit le même chemin que la première sauf que celle ci donne l'ensemble des classes présenté sous forme d'une liste (sans la visualisation de leurs rapprochements).

Les chemins suivis, suite à l'utilisation de la première interface se schématisent comme suit :




**Module recherche (l'interface graphique)**

## IV- Schéma général du système

Le système METRISYS peut être représenté par le schéma général suivant :

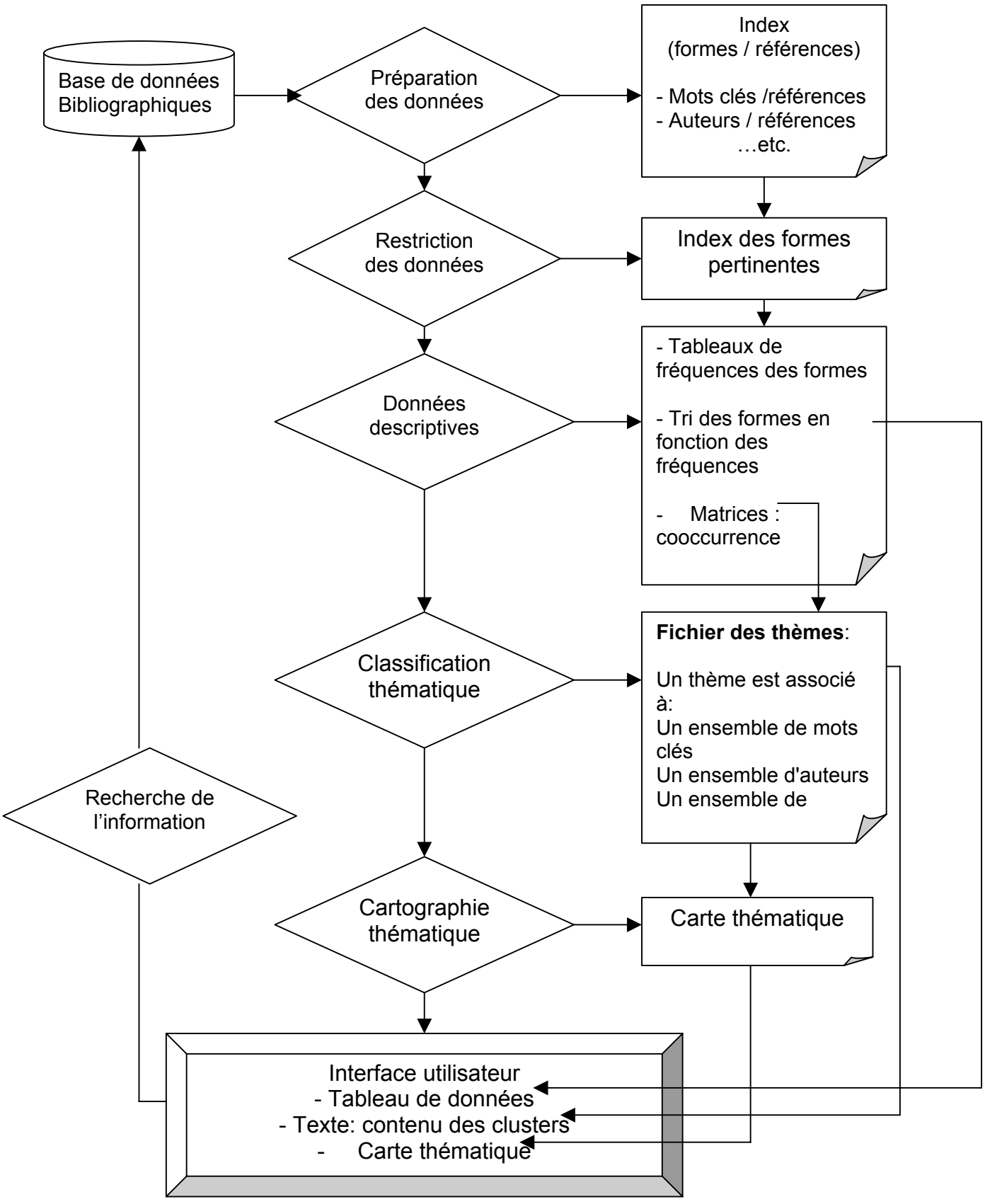



**Schéma général de METRISYS**

## V- Réalisation du système

### 1- Environnement de développement

L'environnement de développement choisi est l'environnement Delphi. Ce choix est justifié par les avantages suivants :

1- Delphi apporte une grande souplesse au développeur.
2- Il génère un fichier exécutable (.exe) qui ne nécessite aucun autre fichier pour être exécuté.
3- Il offre un compilateur optimisé qui donne une application rapide sans qu'il soit nécessaire de faire un effort pour optimiser les programmes.
4- Il offre une facilité de création et de gestion de bases de données.

…etc.

### 2- Le système WINISIS

WINISIS est un système conçu spécifiquement pour la gestion informatisée des bases de données non numériques et structurées (constituées principalement de texte). Un de ses principaux avantages est la manipulation d'un nombre illimité de bases de données, chacune pouvant se composer d'éléments complètement différents.

Les éléments d'informations sont enregistrés dans ***les champs,*** à chacun est assignée une ***étiquette*** numérique qui indique son contenu. Ce système permet la gestion de champs (et donc d'enregistrements) de longueur variable ; cela permet, d'une part, une utilisation optimale du stockage sur disque et d'autre part une grande liberté pour définir la longueur maximale de chaque champ. Le champ peut contenir des ***sous-champs,*** dont chacun est identifié par un séparateur ***de sous-champ*** (ensemble de deux caractères précédant l'élément ). En outre un champ peut être ***répétitif,*** c.-à-d. qu'un enregistrement peut contenir plusieurs fois ce champ.

#### *WINISIS offre les fonctions suivantes :*

1- Définition de bases de données contenant les rubriques désirées,
2- Saisie de nouveaux enregistrements dans une base de données,
3- Modification, correction ou effacement d'enregistrements,
4- Affichage d'une partie ou de la totalité des enregistrements selon les besoins,
5- Tri des enregistrements dans n'importe quel ordre.

6- Impression de catalogues et/ou d'index complets ou partiels d'une base.

7- …etc.

### 3- L'interface ISIS-DLL

**ISIS-DLL** représente une bibliothèque de liens dynamiques compatible avec le système WINISIS. Elle constitue un outil de développement d'applications sous l'environnement Windows 3.1, Windows 3.11, Windows 95 ou Windows NT en utilisant Visuel Basic, Delphi, C ou C++. Elle comporte un ensemble de fonctions permettant de manipuler les entités d'une base de données ISIS (Fichier Maître, Fichier Inversé...) à travers l'application développée. Elle représente, donc, une interface reliant les applications aux bases de données ISIS. Elle compote les fichiers suivants :

| | |
|---|---|
| isis001.bas | Constantes et types utilisés en Visuel Basic |
| isis001.pas | Constantes et enregistrements utilisés en Delphi |
| isis001.h | Constantes et structures utilisées en C et C++ |
| Isisdll.h | Fonctions prototypes en C et C++ |
| Isis32.dll | Bibliothèque ISIS-DLL |
| Isis32.bas | Module de déclaration des fonctions en Visuel Basic |
| Isis32.pas | Unité de déclaration des fonctions en Delphi |
| Isis32.lib | Bibliothèque de déclaration des fonctions en C et C++ |

### 4- Interface hypertextuelle

Durant les dernières années plusieurs recherches ont porté sur l'ergonomie des interfaces. En effet, l'interface utilisateur constitue le chemin d'entrée au système de recherche. Ainsi, la convivialité de cette dernière joue un très grand rôle dans la satisfaction des utilisateurs qui constitue le but principal de toute recherche dans ce domaine.

Dans ce sens, le logiciel METRISYS propose une interface graphique contenant des liens hypertexte permettant la navigation de la carte thématique vers les mots clés, auteurs ou références pour arriver à une information pertinente pour le sujet recherché.

La réalisation de cette interface se base principalement sur la notion d'hypertexte qui se présente comme suit :

**4.1- L'hypertexte**

L'hypertexte est un document informatisé composé de nœuds reliés entre eux par des liens. La nature de ces nœuds peut être aussi bien textuelle, visuelle, sonore ou encore audiovisuelle. Il constitue un nouveau système d'inscription et d'enregistrement et un nouveau moyen de classer et d'organiser des connaissances et des données symboliques. Il donne à l'utilisateur la possibilité de choisir son cheminement à l'intérieur d'un document. En cliquant à l'aide de sa souris sur le mot ou l'icône qui l'intéresse, l'utilisateur est immédiatement dirigé vers la partie du document qui s'y rattache. Ainsi, l'utilisateur construit son propre parcours de lecture en fonction de ses préoccupations et de ses intérêts. L'hypertexte est, par conséquent, un document virtuel qui n'est jamais globalement perceptible. Cette propriété en fait un document "interactif" dans lequel le lecteur tient une place prépondérante.

## VI- Description générale des Menus

Le système MetriSys offre un ensemble de menus pour l'activation de ses fonctionnalités et un ensemble de boites de dialogues pour la sélection ou la définition des divers paramètres ou options d'analyse.

### 1- Fenêtre principale

L'exécution de MetriSys induit l'apparition de la *fenêtre application,* dite *fenêtre principale*, qui se présente comme suit :


Cette fenêtre comporte :

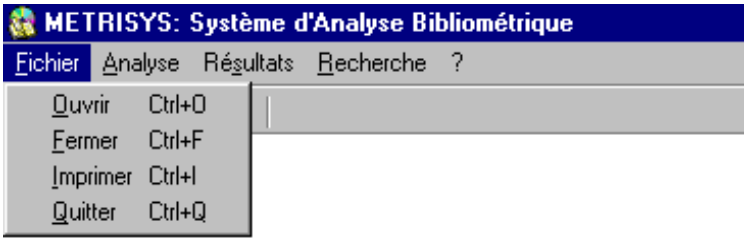
- Le menu principal composé de 5 menus : Fichier, Analyse, Résultats, Recherche et ?
- Une barre d'outils
- Et une barre d'état

**- Menu  Fichier**

Le menu fichier comprend les fonctionnalités suivantes:

- Ouvrir et balayer une base de données,
- Imprimer des références d'une base de données,
- Fermer la base de données ouverte,
- Quitter l'application.

Ce menu se présente comme suit :

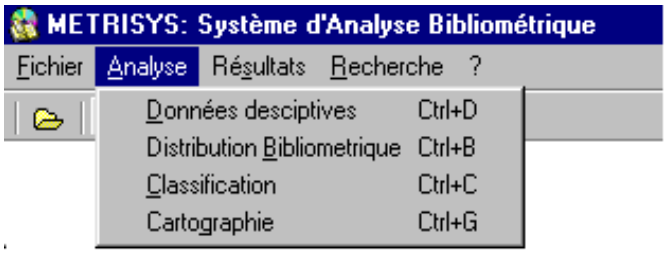



**- Analyse**

Ce menu permet d'effectuer les traitements  suivants:

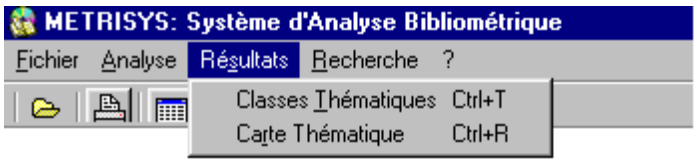
- Génération de données descriptives : classement des formes par ordre de fréquence,
- Application des distributions bibliométriques pour le découpage des données,
- Classification thématique des données,
- Représentation des classes thématiques sur une carte.

Il se présente comme suit :



**- Résultat**

Ce menu permet de visualiser les résultats de la classification thématique et la cartographie des données, autrement dit, la description détaillée des classes thématiques et leur représentation sur une carte.

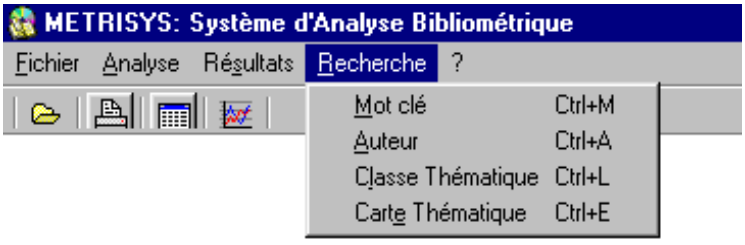



**- Recherche**

Ce menu comprend les fonctionnalités suivantes:

- Recherche par mot clé,
- Recherche par auteur,
- Recherche par classe thématique,
- Recherche sur la base de la carte thématique.

Il se présente comme suit :

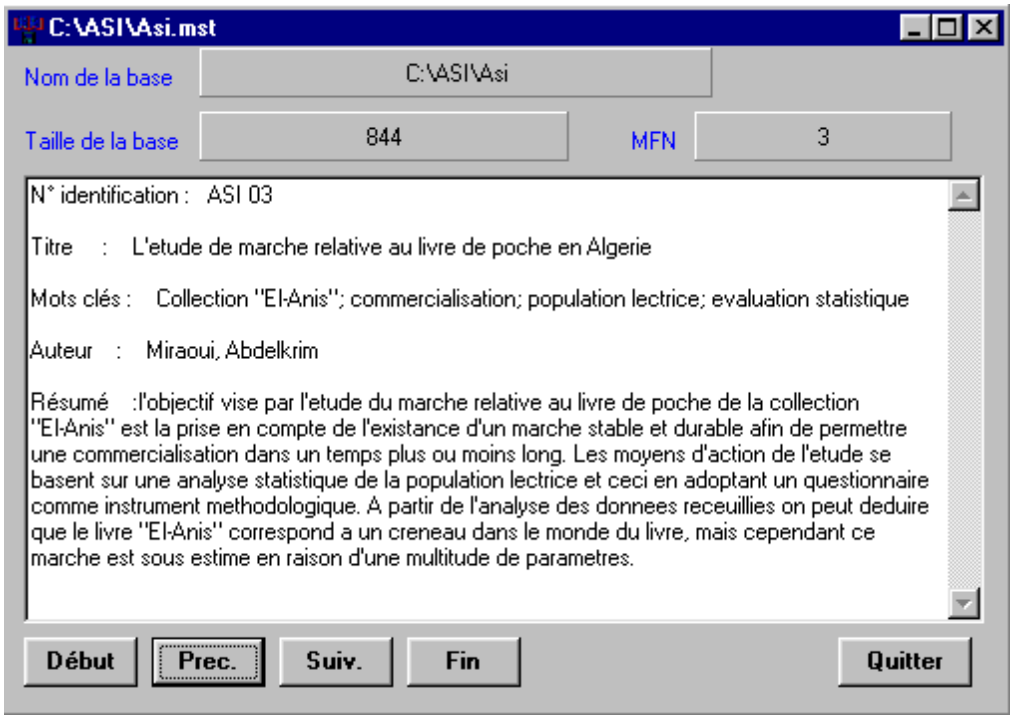



**- ?**

Comprend le module aide du logiciel.

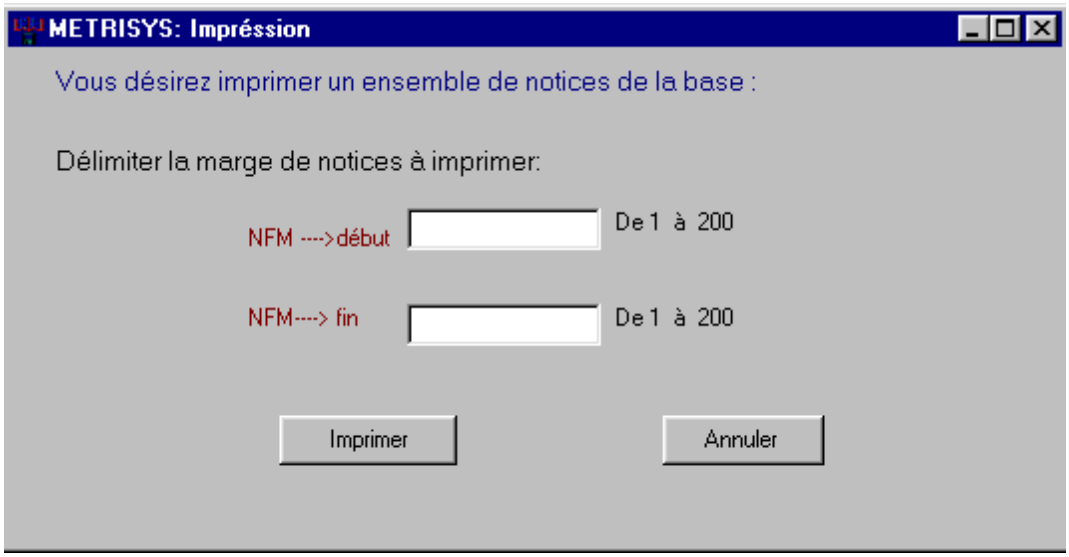
## 2.1- Balayage des données

Le système METRISYS permet de charger des bases de données bibliographiques ( ayant un format ISO 2709 ) et de faire un balayage de ces dernières.




## 2.2- Impression des données

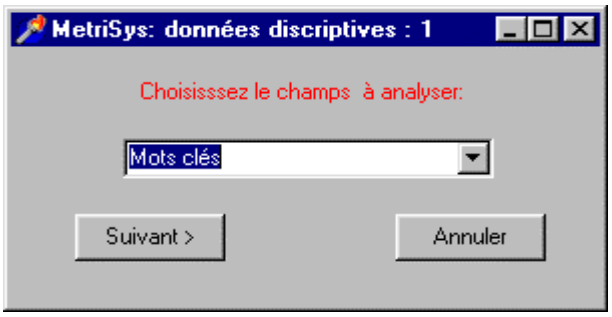
Le système permet d'imprimer un ensemble de références de la base bibliographique.



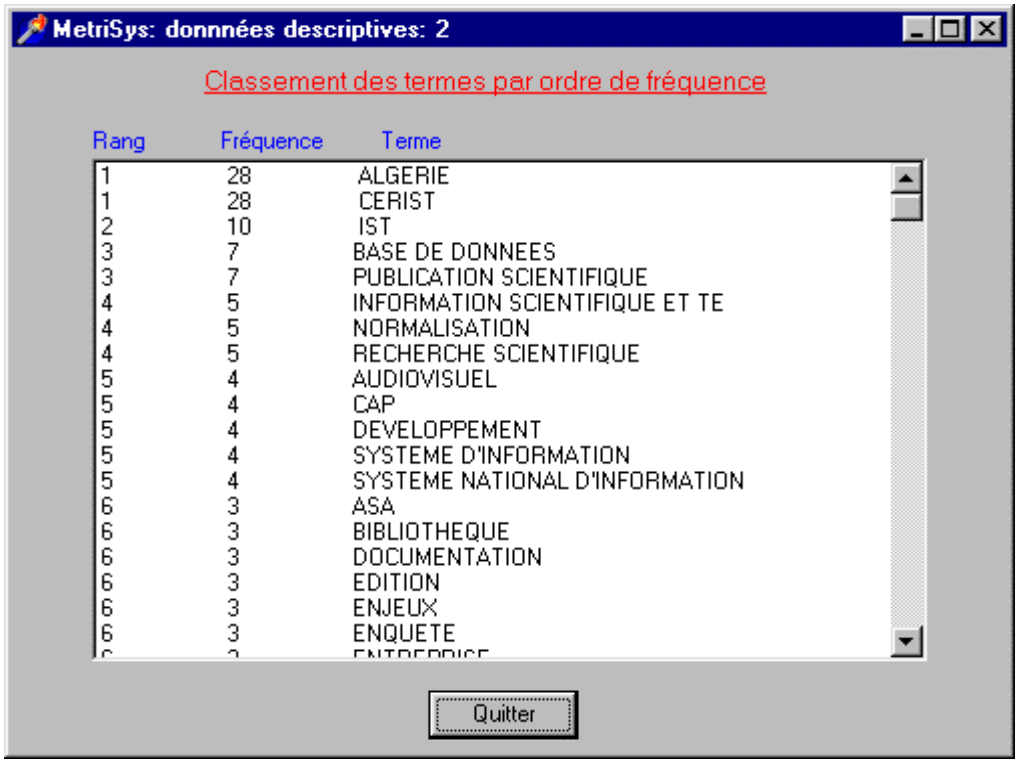

### 2.3- Génération de données descriptives

Cette fonctionnalité comporte l'extraction des unités d'information (mots clés, auteurs,..) à partir de la base de données bibliographiques puis le tri de ces derniers en fonction de leur fréquence dans la base.

L'utilisateur peut effectuer le choix du champ à traiter (mots clés, auteurs…etc.).

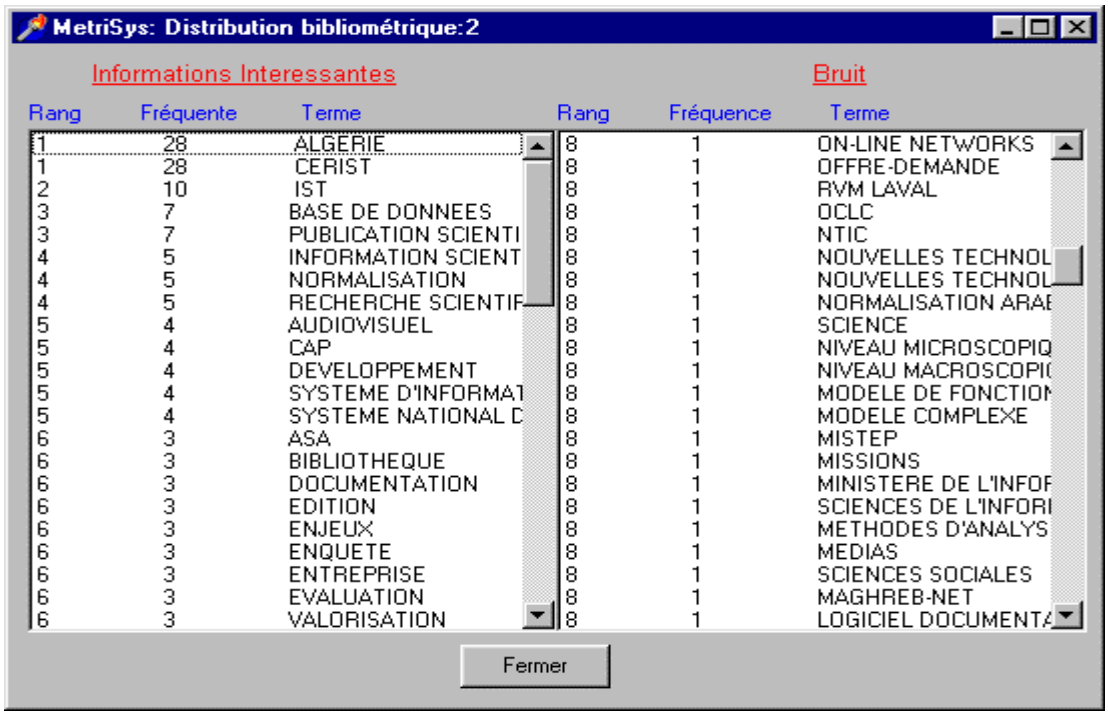



Par la suite l'ensemble des unités d'information ou des formes sont classées par ordre de fréquence.



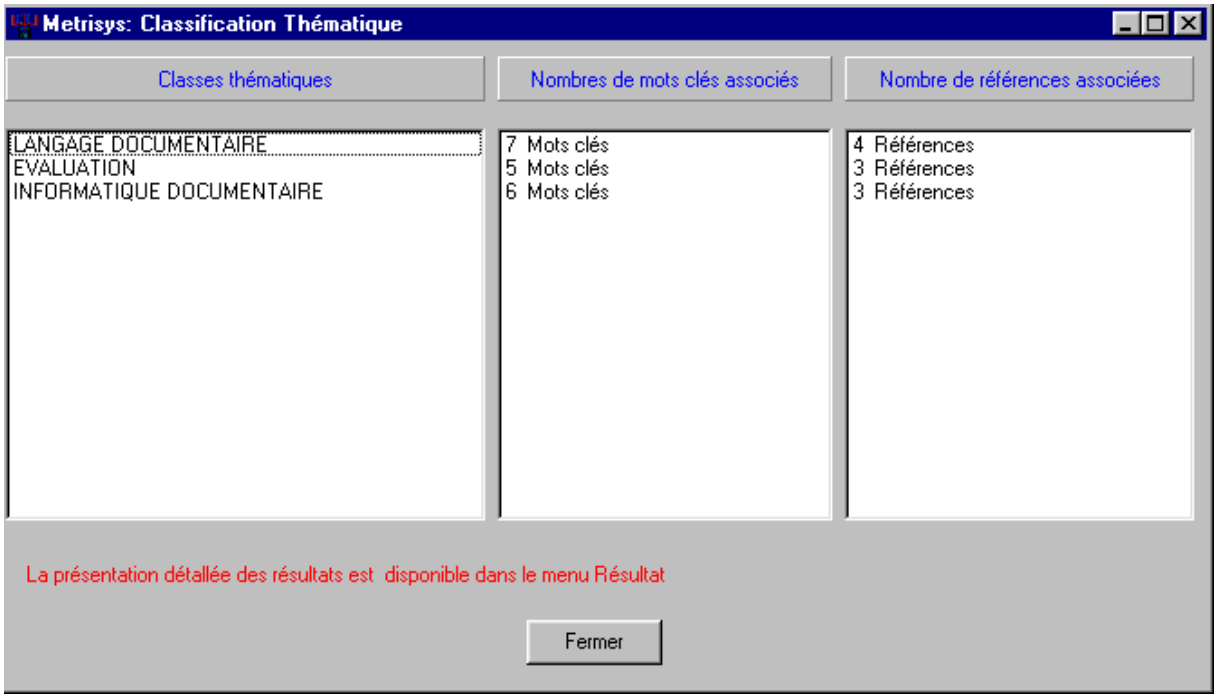

## 2.4- Découpage des données

Cette fonctionnalité permet de déterminer l'ensemble des informations pertinentes parmi toute la masse d'unités d'information. L'utilisateur peut choisir le champ sur lequel il veut effectuer ce découpage. Par la suite, les données pertinentes et le bruit sont automatiquement séparés.



## 2.5- Classification des données

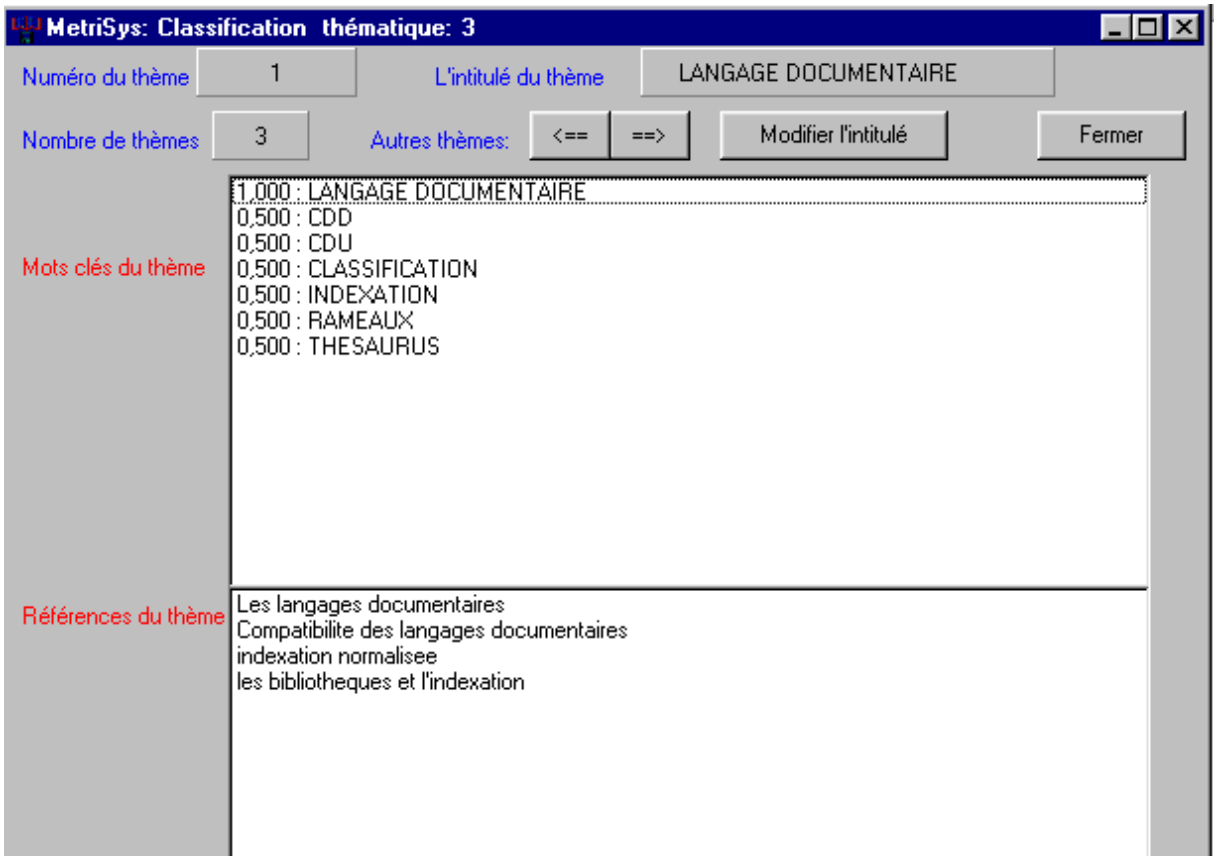
L'option classification du menu analyse permet de dégager les classes thématiques abordées dans la base de données. Chaque classe est associée à un ensemble de références bibliographiques et un ensemble de mots clés.



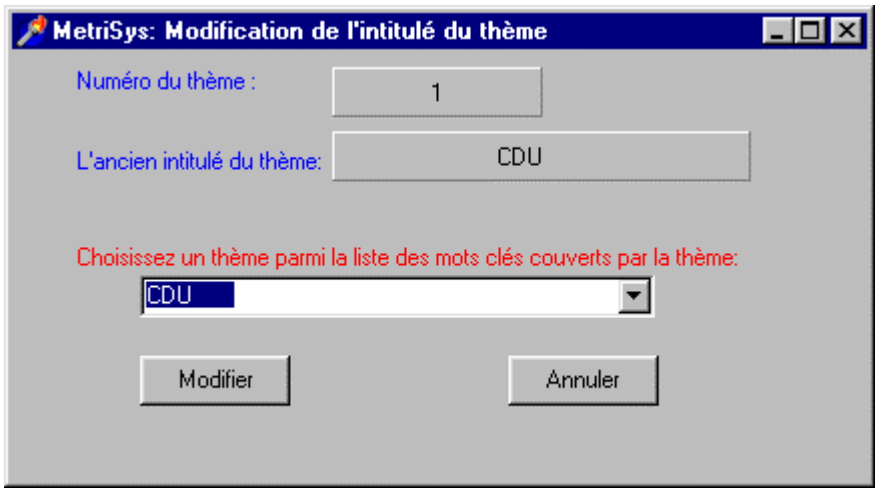

La présentation détaillée de chaque classe thématique et obtenue grâce au menu Résulat.

Cette présentation comprend pour chaque classe :

- Numéro du thème,
- L'intitulé du thème,
- Les mots clés associés au thème : munis d'indices reflétant leur importance.
- Les références associées au thème.

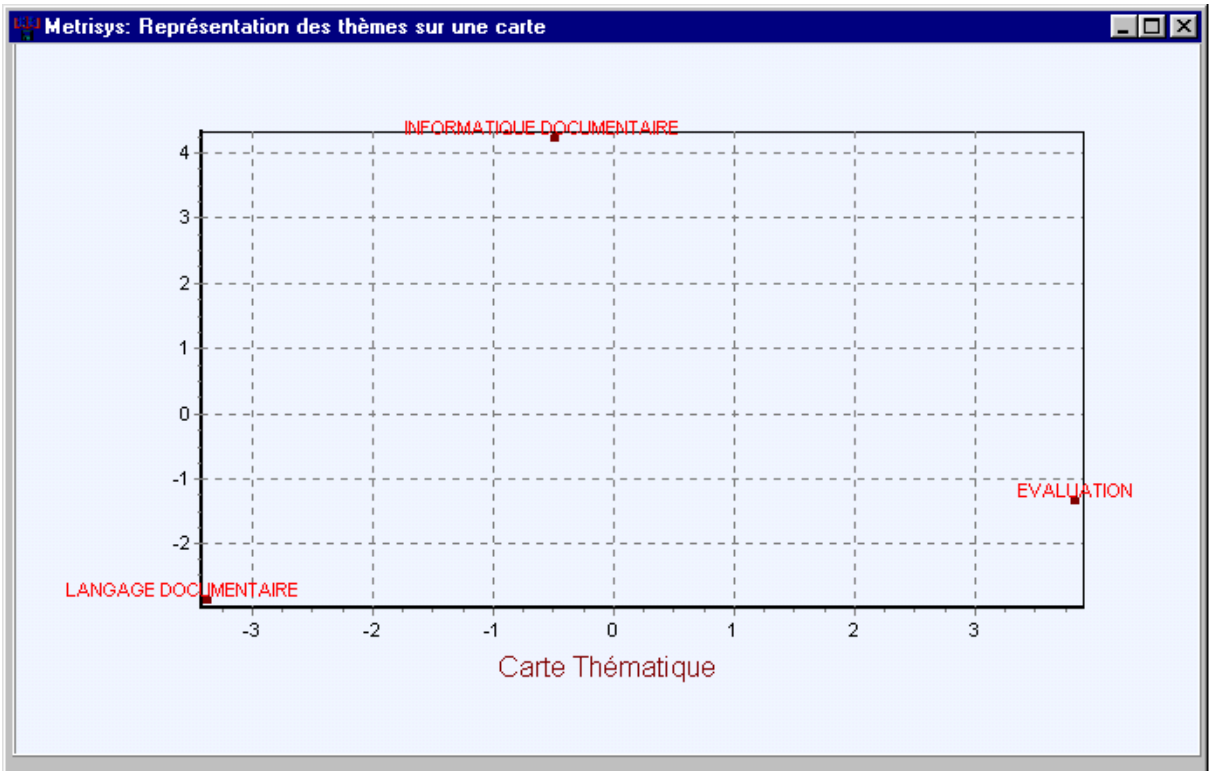



L'intitulé du thème est choisi parmi la liste des mots clés en fonction de l'indice suscité. Etant donné que plusieurs mots clés peuvent avoir le même indice, le système MetriSys offre à l'utilisateur la possibilité de modifier l'intitulé de la classe thématique.



## 2.6- Cartographie des classes thématiques

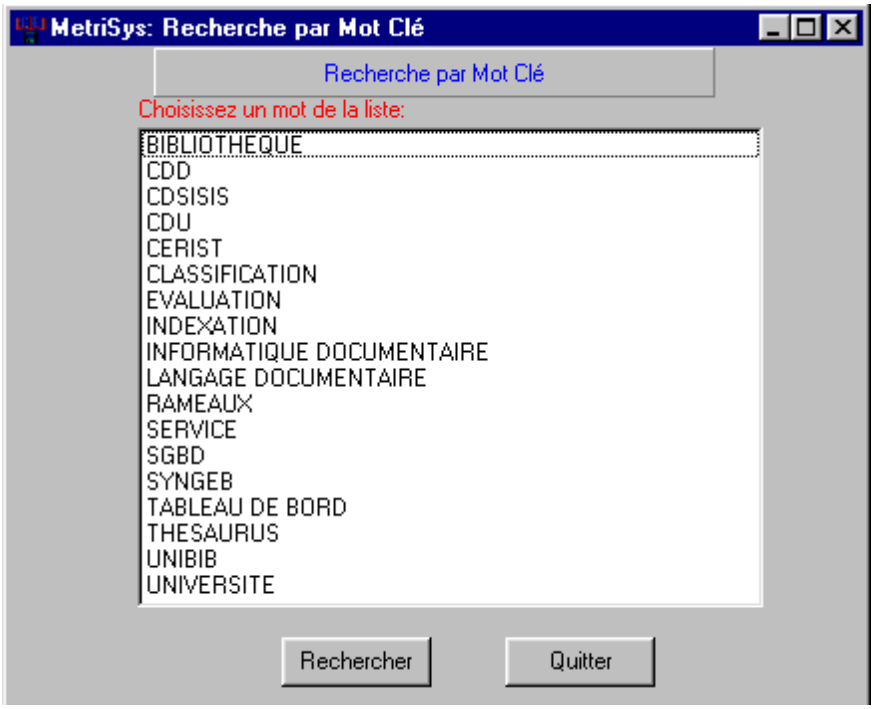
Comprend la représentation des classes thématiques sur un plan en se basant sur les résultats obtenus suite à l'application de l'ACP.



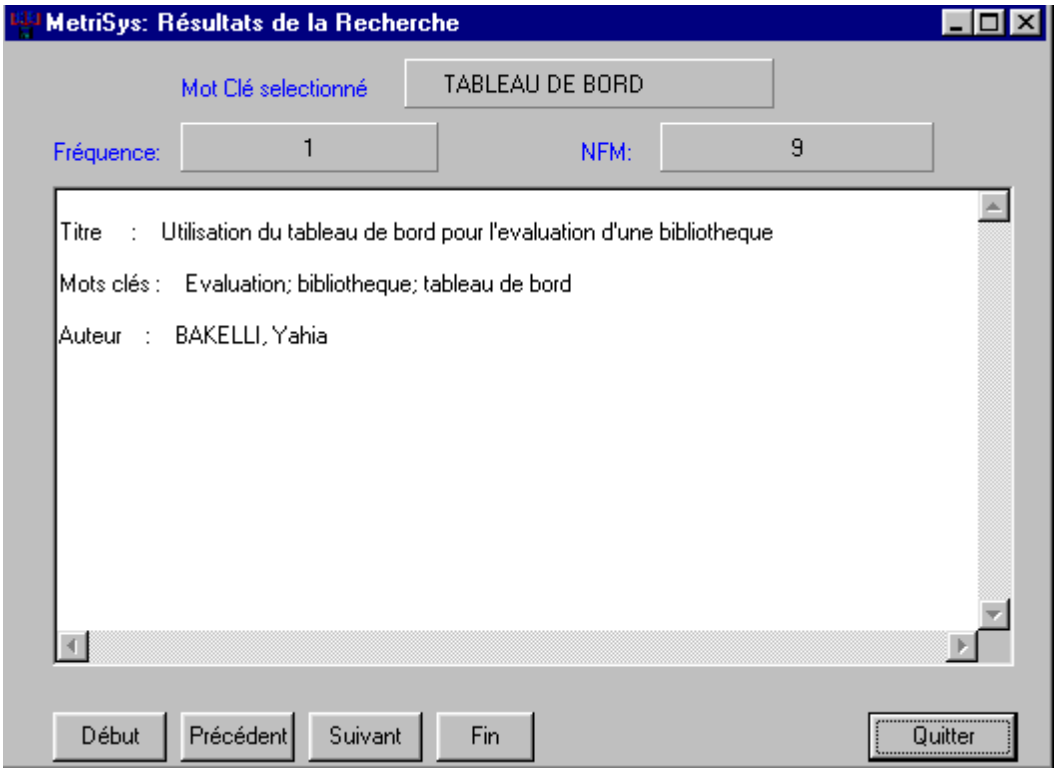

## 2.7- Recherche bibliographique

La recherche peut se faire selon deux approches:

a- En démarrant d'un index de mots clés ou d'auteurs:



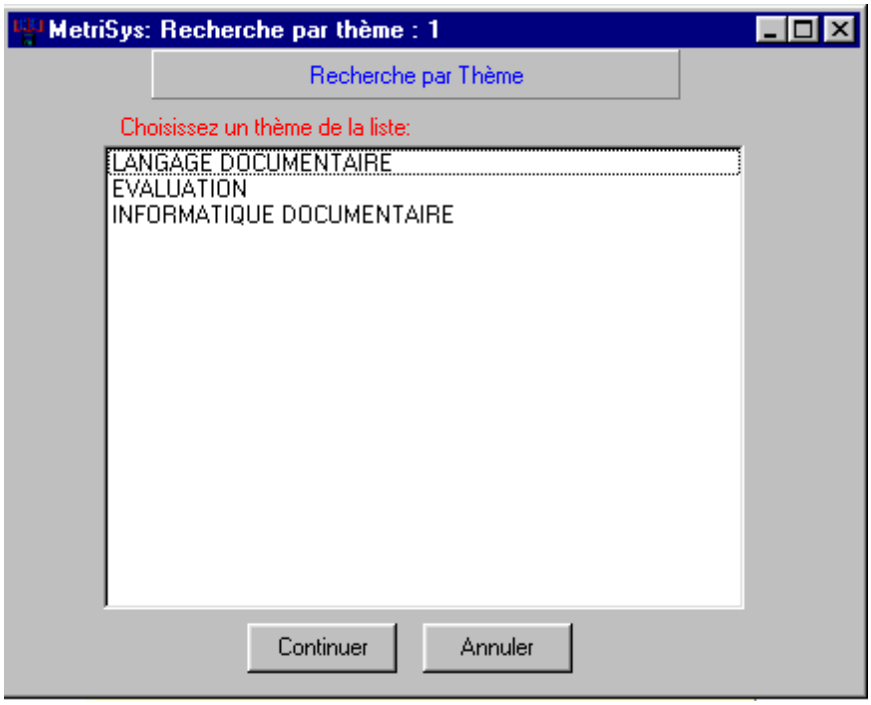

Après le choix d'un élément de l'index, l'utilisateur accède aux résultats de la recherche.



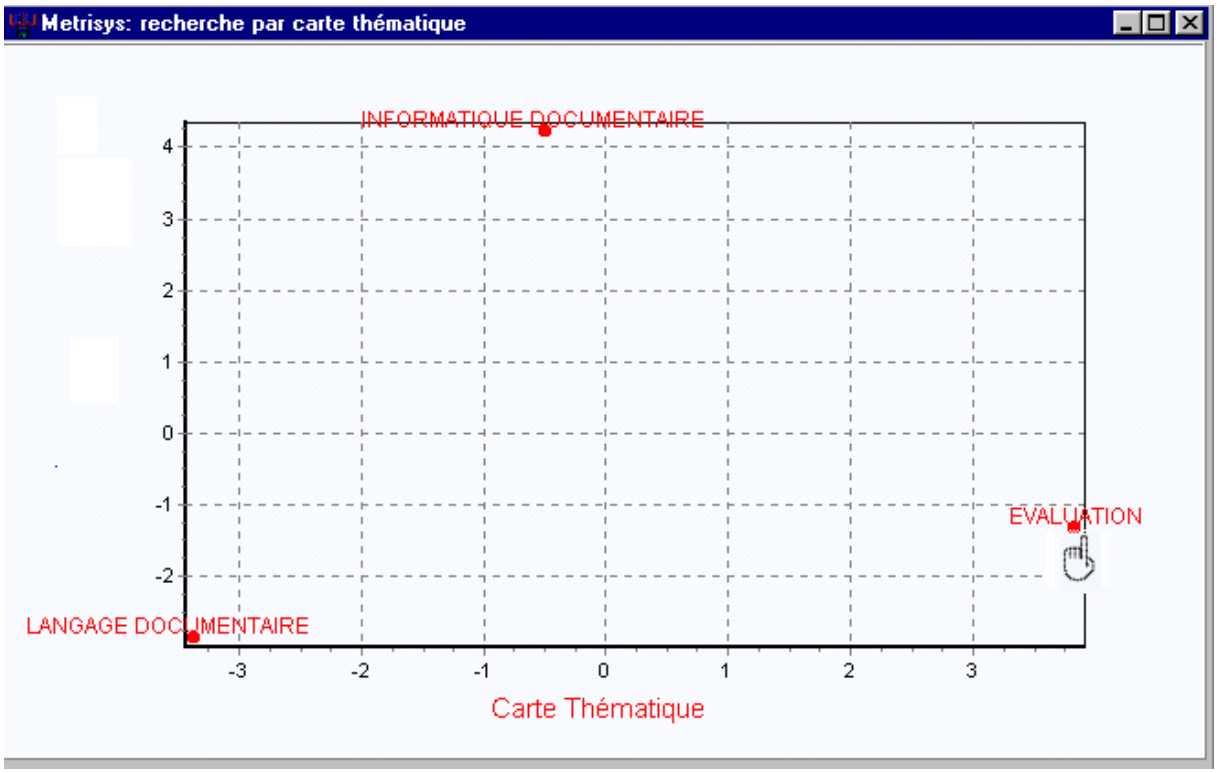
b- Pour la deuxième approche, l'utilisateur dispose de l'ensemble des classes thématiques abordées dans la base. Elle passe par deux principales étapes :

- La première étape consiste en le choix d'une classe thématique, grâce à :

- l'option Classe thématique du menu recherche, qui offre la liste des classes thématiques sous forme d'un index.

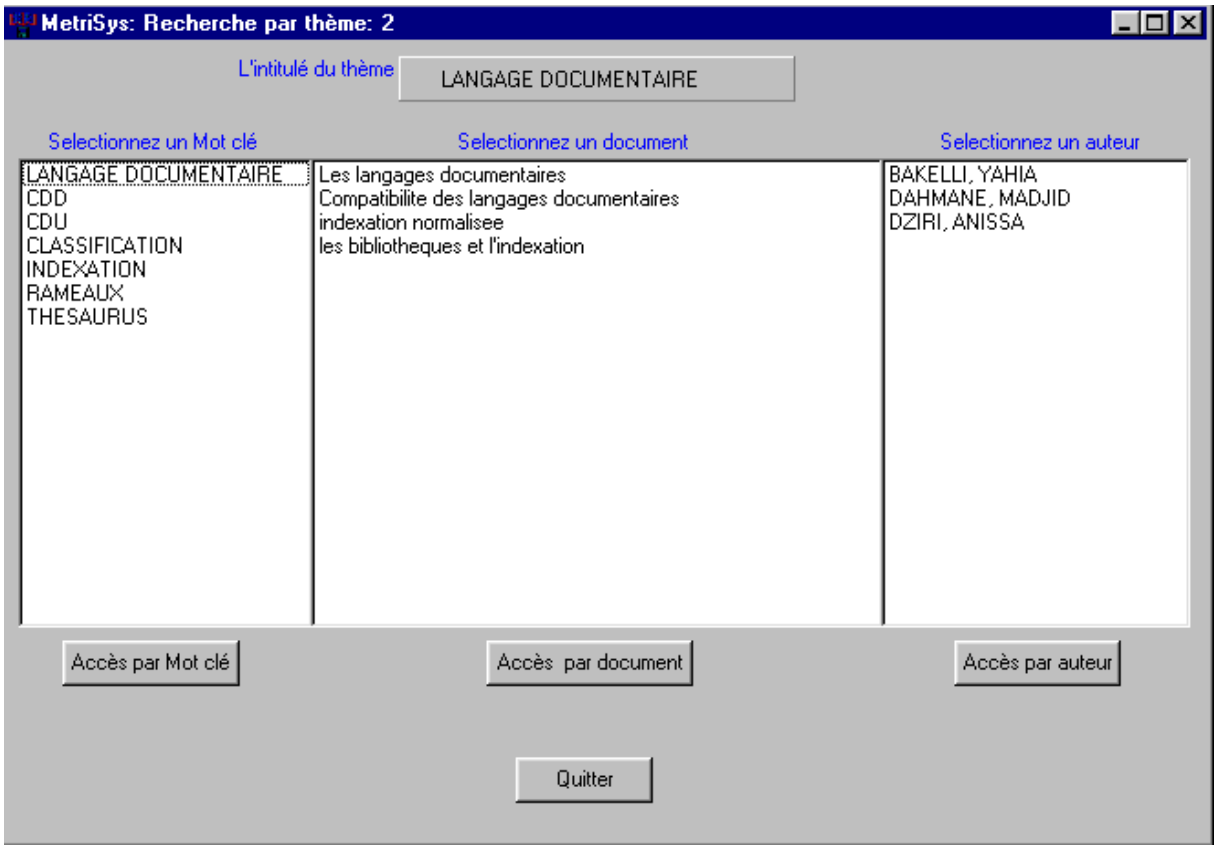



- Ou l'option carte thématique du menu recherche qui propose un accès grâce à la carte thématique. Chaque classe correspond à un point sensible de la carte (lien hypertexte). Le choix se fait par un simple click sur le point correspondant à la classe sélectionnée.




- Suite à la sélection de la classe thématique, la deuxième étape de cette approche propose trois modes d'accès aux données:

  - Accès à partir des mots clés de la classe thématique,
  - Accès à partir des références de la classe thématique,
  - Accès à partir des auteurs de la classe thématique choisie.




Les résultats de la recherche se présentent de la même manière que dans la première approche de recherche.

**3- Perspectives**

Ceci représente une première version du logiciel METRISYS. Ainsi tous les menus et toutes les fonctionnalités déjà présentés sont susceptibles d'être modifiés et améliorés.

Aussi, la méthode de classification appliquée sur les mots clés en les considérant comme étant les meilleurs indicateurs de contenu pour obtenir les classes thématiques peut être appliquée, en plus, sur le champ auteur. Ce qui permet de dégager des groupes d'auteurs (chercheurs ) travaillant sur les mêmes thématiques et de donner des idées de constitution de réseaux de chercheurs.

## VIII- Contexte d'application

Certes, il est clair que nous cherchons à concevoir un système permettant d'exploiter et traiter la masse d'information contenue dans les bases de données bibliographiques dans le but d'induire des informations élaborées pour faciliter la recherche bibliographique aux utilisateurs. En effet, la présentation des thématiques, couvertes par la base de données, sur une carte accessible via une interface permettant la navigation dans la base de données évite à l'utilisateur d'introduire des mots clés qui peuvent bien limiter sa recherche (oubli d'un mot clé important, erreur dans le choix des opérateurs bouléens séparant les mots clés dans l'équation de recherche... etc. ).

Le système peut être conçu, autrement, dans un but strictement de veille. La compilation de plusieurs cartes thématiques, par exemple, peut servir pour avoir des courbes d'évolution des domaines de recherche ou des mots clés. De telles courbes peuvent servir comme indicateurs stratégiques de la dynamique des sciences à travers les publications scientifiques des chercheurs. De plus, même les statistiques bibliométriques simples (descriptives) que peut induire un système d'analyse bibliométrique (fréquence des formes, distribution par forme... ) peuvent servir pour faire de la veille à travers des études comparatives.

# Bibliographie


**Thèses et mémoires de DEA**

**[1]** N. BEN ABDALLAH. « Analyse et structuration de documents scientifiques pour un accès personnalisé à l'information utile : vers un système d'information évolué ». Thèse de doctorat en Sciences de l'Information et de la Communication. Université Claude Bernard, Lyon 1, 1997.

**[2]** V. HENRY. « Le processus et les outils de veille technologique dans un centre de recherche développement ». Mémoire de D.E.A en Science de l'information et la communication. Université Lumière Lyon2, 1998

**[3]** E. KLOMAYER. « Contribution à l'analyse des processus cognitifs mis en jeu dans l'interrogation d'une base de données documentaires ». Thèse de doctorat en psychologie Université René Descartes, Paris5, 1997

**[4]** A. LELU. « Modèles neuronaux pour l'analyse de données documentaires et textuelles ». Thèse de doctorat en mathématiques. Université PARIS IV, 1993

**[5]** C. MICHEL. « Evaluation de systèmes de recherche d'information comportant une fonctionnalité de filtrage, par des mesures endogènes ». Thèse de doctorat en sciences de l'information et de la communication. Université Lumière Lyon 2, 1999

**[6]** J. MOTHE . « Un modèle connexionniste pour la recherche d'information : reformulation de requêtes et apprentissage ». Thèse de doctorat en informatique Université P. Sabatier, 7 octobre 1994

**[7]** Y. OUSSAR. « Réseaux d'ondelettes et réseaux de neurones pour la modélisation statique et dynamique de processus ». Thèse de doctorat en Robotique. Université Paris IV, 1998

**[8]** P. PRENON. « Réalisation d'un prototype de Système de Recherche d'information scientifique ». Mémoire de D.E.A. Laboratoire RECODOC, Université Claude Bernard, Lyon 1, 2000

**[9]** L. TAMINE. « Les systèmes de recherche d'information: reformulation de la requête et apprentissage basés sur les algorithmes génétiques ». Thèse de Magister en Informatique. Université Mouloud Mammeri, TiziOuzou, 1998

**Monographies**

**[10]** S. BIREME. « ISIS Application Program Interface : ISIS_DLL User's Manuel ». Preliminary version.UNESCO, August 1997

**[11]** J. CHAUMIER. « Travail et méthodes du /de la documentaliste ». Les éditions ESF, 1982

**[12]** G. DEUTSCH, M. GROSS, K. RICHTER. «  Le grand livre : Borloand Delphi » MICRO Application, 1998

**[13]** Y. DODGE. « Statistique : dictionnaire encyclopédique ». Dunod, Paris, 1993

**[14]** Ecole des Sciences de l'Information. « L'indexation automatique ». ESI, juin 1998

**[15]** C. GUINCHAT, Y. SKOURI. « Guide pratique des techniques documentaires : traitement et gestion des documents».  EDICEF, 1989

**[16]** T. LAFOUGE. «  Mathématique du document et de l'information : bibliométrie distributionnelle ». .Habilitation à diriger des recherches, Volume 1. Laboratoire RECODOC, Université Lyon 1, 1998

**[17]** Y. LE COADIC. « La science de l'information ». Collection : que sais-je ? 1994

**[18]** H. ROSTAING. « La bibliométrie et ses techniques ». Collection «outils et méthodes » CRRM, 1996

**[19]** B.SAPORTA. « Analyse des données et statistiques ». Edition Technip, 1990

**[20]** UNISIST. « principes d'indexation ». UNESCO, Paris,  1975

**Articles de périodiques**

**[21]** A. ARAB. « Techniques et lois bibliométriques ». RIST , Vol.4, N°1, 1994, pp. 22-27

**[22]** P.  BELBENOIT-AVICH. " Les bases de données plein-texte biomédicales et la fourniture de documents". Bull.Bibl.France, Paris, t.37, n°6, 1992, pp. 14-19

**[23]** M. BUCKLAND. « What is a document ». Journal of american society for information sciences 48(9), 1997, pp. 804-809

**[24]** P. COTE. « Modelisation de l'utilisateur dans une interface de recherche 'information ». Documentation et bibliothèques, Avril-Juin 1991, pp 65-70

**[25]** S. LAINE-CRUZEL. « PROFILDOC : Filtrer une information exploitable ». BBF , T.44 n°5 , paris 1999, pp.143-147

**[26]** H.P. LUHN. "A statistical approach to mechanized encoding and searching of literary information". IBM journal of research and development, vol. 1, n°4 October 1957

**[27]** L. QUONIAM, P. HASSANALY, P. BALDIT,  H. ROSTAING, H. DOU.« Bibliometric analysis of patent documents for R&D management ». Research evaluation, vol. 3, N° 1, April 1993, pp. 13-18

**[28]** L. QUONIAM, H. ROSTAING,E. BOUTIN, H. DOU. « Treating bibiliometric indicators with caution : their dependence on the source database ». Research evaluation, vol. 5, N° 3, December 1995, pp.177-181

**[29]** M.G.SURAUD, L. QUONIAM, H. ROSTAING, H. DOU. «On the signification of data bases keywords for a large scale bibliometric investigation in fundamental physics ». Scientometrics, Vol. 33, N° 1 , 1995, 41-63

**[30]** G. TEASDALE. « L'hypertexte: historique et applications en bibliothéconomie ». Cursus vol.1 no 1, octobre 1995, périodique électronique

**Communications**

**[31]** M. BUCKLAND . « Forme, signification et structure des systèmes de sélection du savoir ». ISKO99, Lyon ,France , Oct 21-22 , 1999

**[32]** C. FLUHR. « Problèmes d'optimisation de l'accès à l'information dans les bases de données textuelles ». Journées d'études : applications informatiques conversationnelles et le langage naturel, Juin 1984

**[33]** S. KOUICI. « Classification et cartographie des données : outils de recherche bibliographique et de décision ». 2eme édition du séminaire national sur : « Le système national d'information : état actuel et perspectives », 21-22 juin 1999

**[34]** T. LAFOUGE. « Les lois de l'information : une réalité ». Séminaire de l'ADEST, Mai 1997

**[35]** S. LAINÉ CRUZEL . « SRI incluant un filtrage personnalisé : pourquoi ? Comment ? » Laboratoire Recodoc, Université Claude Bernard Lyon1. Séminaire ADEST, 24/03/00

**[36]** C. MICHEL, T. LAFOUGE. « Profil-doc : un système personnalisé de requête à des bases de données en texte intégral ». Actes de congrés SFBA « les systèmes d'information élaborée », Ile Rousse , 12-16 Mai 1997

**[37]** C. MICHEL, S. LAINE-CRUZEL. « Profil-Doc : Un prototype de système de recherche d'information personnalisé selon le profil des utilisateurs ». Ateliers A2, 11ème conférence francophone Interaction Homme Machine IHM'99 "L'interaction pour tous", Montpellier, 22-26 novembre 1999

**[38]** C. MICHEL. « Le prototype Profil-Doc ». Laboratoire Recodoc, Université Claude Bernard Lyon1. Séminaire ADEST 24/03/00

**[39]** C. MICHEL. « Diagnostic evaluation of personalized filtering information retrieval system : Methodology and experimental results ». Congrès RIAO 2000, 6eme conférence internationale sur la Recherche d'Information Assistée par Ordinateur : "Content based multimedia information access" . Collège de France, Paris, 12-14 avril 2000

**[40]** Josiane MOTHE, Malika ABCHICHE. « SYRENE : Un système de recherche d'information basé sur un modèle de réseaux de neurones ».Veille stratégique, scientifique et technologique, VSST, Toulouse, octobre 1995.

**[41]** X. POLANCO, J. ROYAUTÉ, L. GRIVEL, A. COURGEY. « Infométrie et linguistique informatique : une approche linguistique au service de la veille scientifique". Journées sur les systèmes d'information élaborée, Ile Rousse (Corse), 30 mai-2 juin 1995

**Pages Web**

**[42]** M. CORSINI. « Réseaux de neurones artificiels : une introduction »
www.scico.u-bordeaux2.fr/~corsini/pedagogie/ANN/main/node14.html
3/11/1998

**[43]** P. MALAN. « Définition de l'information »
19/12/1998
http://www.olats.org/schoffer/definfo.htm

**[44]** PMSI. « Guide du concepteur de réseaux de neurones »
1991
www.geocities.com/CapeCanaveral/Launchpad/7651/guide.htm

**[45]** C. ROISIN. « Qu'est ce qu'un document »
Habilitation à diriger des recherches en INFORMATIQUE
L'Institut National Polytechnique de Grenoble
www.inrialpes.fr/opera/people/Cecile.Roisin/habilitation.html

**[46]** P. TRIGANO. « Indexation automatique et sauvegarde des connaissances de l'entreprise»
Texte sélectionné, 1994
http://infoweb.magi.com/~godbout/Kbase/trigano.htm