École Nationale Supérieure des Sciences
de l'Information et des Bibliothèques

DEA

Sciences de l'information et de la Communication

option : Systèmes d'Information Documentaire

MEMOIRE DE DEA

STRUCTURATION ET NAVIGATION DANS UN SYSTEME D'INFORMATIONS TEXTUELLES GÉRÉES AVEC SPIRIT

Houria OUICI

Sous la direction de

MME SYLVIE LAINÉ-CRUZEL

1995

Université Lumière
Lyon 2

École Nationale Supérieure des Sciences
de l'Information et des Bibliothèques

Université Jean-Moulin
Lyon 3

École Nationale Supérieure des Sciences
de l'Information et des Bibliothèques

DEA

Sciences de l'information et de la Communication

option : Systèmes d'Information Documentaire

MEMOIRE DE DEA

STRUCTURATION ET NAVIGATION DANS UN SYSTEME D'INFORMATIONS TEXTUELLES GÉRÉES AVEC SPIRIT

Houria OUICI

Sous la direction de

MME SYLVIE LAINÉ-CRUZEL

1995

Université Lumière
Lyon 2

École Nationale Supérieure des Sciences
de l'Information et des Bibliothèques

Université Jean-Moulin
Lyon 3

DEDICACES

♥ *A la mémoire de ma très chère et regéttée mère* ♥

♥ *A mon* très cher père ♥

♥ *A mes adorables neveux et tous les enfants du monde* ♥

♣ *A ma famille* ♣

♣ *A mes ami(e)s* ♣

REMERCIEMENTS

Je ne manquerai pas de souligner que mon mémoire a vu le jour grâce à la collaboration effective de l'E.S.S.I.B toute entière. Je tiens à remercier le personnel de la bibliothèque, en particulier Jacqueline, Mari-joe, et David pour leur gentillesse et tous les services rendus. Je remercie également l'équipe du centre de calcul qui n'a ménagé aucun effort pour m'aider à accomplir ma tâche, ainsi que pour leur sympathie et leur concours technique.

Je remercie Mme Sylvie LAINE-CRUZEL d'avoir dérigé mon travail et accepté d'examiner mon présent travail pour me permettre de présenter ce mémoire. Je la remercie aussi pour toutes ses incitations qui m'ont permis de voir plus clair.

Je tiens à remercier M^{RS} R. BOUCHE et M. HASSOUN pour leur bonne volonté et leur compréhension car elles ont permis à plusieurs étudiants de continuer sans entrave, et de progresser dans leur travail durant les vacances.

Je remercie toute personne qui a contribué de prés ou de loin à ce travail, et je rmercie vivement mes amies qui m'ont beaucoup soutenue.

Résumé:

Ce mémoire aborde le problème de la structuration de données et l'organisation d'enchaînements de traitements dans le cadre du logiciel d'indexation et de recherche d'informations textuelles "SPIRIT".
Notre travail porte tout d'abord sur l'organisation des relations entre unités documentaires pour ensuite définir les mécanismes de navigation relatifs aux types de chaînages établis. Ainsi, des liens statiques sont préétablis entre les unités documentaires afin de guider au mieux l'utilisateur.

MOTS CLES : HYPERTEXTE; STRUCTURATION; NAVIGATION; SPIRIT; LIENS; FILTRAGE; ANTE-SEVEUR.

Abstract :

This study treat the problem of data structuring and the treatment chaining organisation in the context of the indexation and textual information retrieval software "SPIRIT".
Our work concerns, in the first place, the organisation of relations betveen documentary units define later the navigation mecanism relative to established chaining types. Therefore, statics ties are preestabished betveen the unities of documents in order to guid in a much better way the user.

KEY WORDS : HYPERTEXT; STRUCTURING; NAVIGATION; LINKS; GATEWAY; FILTERING.

I) INTRODUCTION

Les techniques classiques de recherche d'information [DACH 90], sont qualifiées de traditionnelles en opposition aux systèmes plus «modernes» que sont les systèmes de recherche d'information conceptuels (SRI), à la base de traitement de langage naturel (TLN) et autres hybrides (voir Glossaire). En effet, contrairement à ces autres écoles, les critères mis en jeu dans ces techniques classiques sont éloignés de la sphère conceptuelle, puisqu'ils relèvent au mieux de l'approche statistique/probabiliste. En revanche, leur étude est toujours, et pour longtemps, un point de passage obligatoire pour les concepts de SRI. Les systèmes de recherche d'information ont pour rôle de servir de médiation entre l'utilisateur et la base d'information, en lieu et place du traditionnel «broker en information» . En conséquence, leur problématique peut se résumer ([HALI 89] p.8, [CHIG 86] p.373, cités par COND 94) en l'adoption d'un couple de modèle, l'un adapté à la prise en compte du besoin formulé par l'utilisateur, l'autre à même de caractériser les informations de la base, ce tout étant chapeauté par une fonction de correspondance dont la finalité est de permettre de savoir dans quelle mesure une information satisfait à la demande de l'utilisateur. Dans ce cadre, la recherche d'information est une activité envisagée sous l'angle d'un processus consistant à identifier, extraire, puis ordonner un ensemble d'information du sein d'une collection, de manière à répondre du mieux possible au besoin formulé par l'utilisateur. Cette définition de la RI implique la prise en compte du mode de représentation de l'information à des fins de comparaison entre la requête et la base mais n'englobe pas la phase de génération des représentations, c'est à dire la phase d'indexation [DACH 90]. IL est à mentionner aussi, que le terme "Information" n'est pas très significatif, car le cadre restreint de la recherche documentaire est sur le point d'imploser, ainsi que le sous-tend l'apparition des systèmes d'information répartis. Ces systèmes ([LANG 93], [OBRA 93], cités par COND 94) ont été produits par l'afflux constant de nouvelles sources d'informations hétérogènes éparpillées tout au long du réseau INTERNET. Ces systèmes s'orientent vers la problématique de la RI, en s'inspirant de l'hypertexte pour WWW, et en assimilant le bouclage de pertinence dans le cas de WAIS (voir Glossaire), Alors qu'au départ, ils étaient spécialisés dans l'aspect protocole de communication et partage de ressource.

Nous abordons ainsi les systèmes hypertexte qui reposent sur l'idée d'associer, grâce à divers liens, des informations stockées dans une BD à des représentations graphiques, l'utilisateur

pouvant se déplacer le long de butineurs (voir Glossaire). L'utilisateur étant souvent dans l'incapacité de décrire ce qu'il cherche, mais à même de le reconnaître instantanément[1] un tel butinage peut être une approche de solution.

De même la tâche de l'utilisateur est loin d'être triviale dans certains domaines, et il parait nécessaire de lui adjoindre l'assistance d'un système informatique. En effet, "*s'il est clair que le rôle de l'être humain est central dans un tel secteur d'activité, il est également visible que de nombreux aspects de la tâche qui lui incombe, on a atteint, voire dépassé, le seuil de ce que l'on peut légitiment exiger de l'esprit humain*" [COND 94].

On est confronté non seulement à un besoin de filtrage de l'information en temps réel mais à des recherches rétrospectives. Les solutions répondant à ces problèmes reposent sur la modélisation de l'utilisateur, la diffusion sélective de l'information (DSI), l'interface en langage naturel, les systèmes coopératifs, qui sont présentés en Annexe.

Le travail que nous allons présenter, rentre dans le cadre des SRI plein texte utilisant intensivement le traitement du langage naturel TLN, lors de la phase d'indexation, ainsi que lors de la phase de la reformulation de la requête. En plus de ces deux points, la gestion de données textuelles et factuelles, le mutilinguisme, l'indexation automatique des descripteurs, facilité d'usage pour l'utilisateur sont pris en considération. Dans ce cadre, le projet Profil-doc (qui sera présenté ultérieurement) a choisi le logiciel "SPIRIT" (Syntactic and Probalistic Indexing and Retrieval of Information in Texts) de SYSTEX pour développer une chaîne du traitement du document textuel permettant d'observer les effets des différents paramètres sur l'efficacité du système pour traiter le texte intégral.

[1] citation *"Je ne sais pas ce qu'est l'art, mais lorsque j'en vois je le reconnais"*

II) Filtrage Des Informations Textuelles

II-1 Problèmes posés au niveau des bases de données hétérogènes

Les différents acteurs intervenant dans une recherche documentaire sont, d'une part, les acteurs humains, d'une autre part, les sources d'information documentaire et les données. Les acteurs humains sont représentés par l'utilisateur final et l'intermédiaire. Le but du premier (indépendamment de son genre et de son évolution) est d'être satisfait par le résultat : conséquence directe d'une formulation d'un besoin d'information . Le second, professionnel de la recherche, généralement présent avec l'utilisateur doit aider ce dernier à effectuer une bonne recherche en :

- précisant la formulation
- choisissant les bons termes
- localisant les bonnes sources d'information
- etc ...

En automatisant le processus, l'intermédiaire serait un ensemble de tâches permettant le dialogue entre l'utilisateur et le système tout en constituant une interface souple et conviviale ainsi que la prise en considération de la notion de transparence d'accès du fait que le dialogue s'établit entre l'utilisateur et une (ou plusieurs) base(s) de données. Cependant, dans la recherche documentaire un grand problème se pose. C'est celui de l'hétérogénéité des bases de données. Ce dernier se traduit par deux formes :

- hétérogénéité des données pour une même base de données
- hétérogénéité d'accès lorsqu'il s'agit de plusieurs bases de données

II-1-1 Hétérogénéité des données dans une même base de données

Cette hétérogénéité est traduite à ce niveau par :

- nature des informations contenues ;
- structuration de ces informations ;
- technique de représentation de ces informations .

On peut dire que le dernier point est lié aux règles internes des participants et aux règles internes des participants et aux différences individuelles des indexeurs malgré un vocabulaire commun.

II-1-2 Hétérogénéité d'accès aux différentes bases de données

D'une façon générale, il s'agit d'hétérogénéité du point de vue pratiques d'indexation variant d'une base à une autre. Cela repose sur les points suivants :

- techniques de recherche documentaire et stratégies de recherche ;
- principes selon lesquels une base de données est alimentée
- structuration d'une base de données ;
- connaissances sur les vocabulaires (lexique, thesauri, classification).

D'une façon particulière, on peut définir certains points déterminant les bases de données et personnalisant l'accès :

- modèle conceptuel de la base de données
- niveau sémantique
- données relatives à des éléments de surface (données signalétiques, descriptives et à caractère objectif)
- données analytiques, représentant une entité dont l'existence est conséquence d'un processus d'inférence. La référence est indirecte, les descripteurs

représentent des concepts où la notion de points de vue joue un rôle important.

II-2 Difficultés d'Utilisation Des Bases De Données En Ligne

Les utilisateurs de systèmes d'information se trouvent confrontés à des difficultés de plus en plus contraignantes devant la richesse et la variété des sources disponibles sur le marché de l'information en ligne. On estime, en effet qu'il existe aujourd'hui plus de 4000 banques de données accessibles à travers les réseaux de télécommunications nationaux et internationaux. A la complexité due à la nature, à la variété et au nombre de sources, il faut ajouter, la multiplicité des accès (réseaux et serveurs, langages d'interrogation) et la diversité des systèmes de facturation.

L'expérience a montré que l'interrogation des banques de données en ligne, fait appel, en plus d'une compétence en matière de langages documentaires, à la connaissance du domaine concerné et à une maîtrise minimale de la langue anglaise pour la formulation d'équation de recherches pertinentes.

La croissance continue de l'offre en matière de BDD en linge, l'absence d'harmonisation des services offerts par les producteurs et les serveurs placent les utilisateurs devant de telles barrières, qu'ils renoncent souvent à les franchir, en se limitant, au quotidien, à l'interrogation de quelques BDD sur un ou deux serveurs.

Cette situation engendre une certaine frustration devant les richesses représentées par les fonds documentaires en ligne devenus paradoxalement inaccessibles [PAOLI 92]

II-3 Besoins d'aides à l'utilisateur

L'idée de se libérer des difficultés et des contraintes citées ci-dessus, a donné depuis quelques années naissance à de nombreuses études, projets et réalisation en vue de créer des systèmes plus ou moins élaborés ayant pour fonctions principales:

- un accès unique à plusieurs serveurs
- un seul langage d'interrogation
- un seul mode de facturation

Différentes appellations existent pour désigner les systèmes décrits dans la littérature depuis quelques années tels que : (antéserveurs, gateways, passerelles "intelligentes", passerelles de communication...). Le concept d'Interface Intelligente" est le plus adopté pour désigner de telles applications. Cependant le terme "Gateways" reste plus général et recouvre divers types d'interfaces.

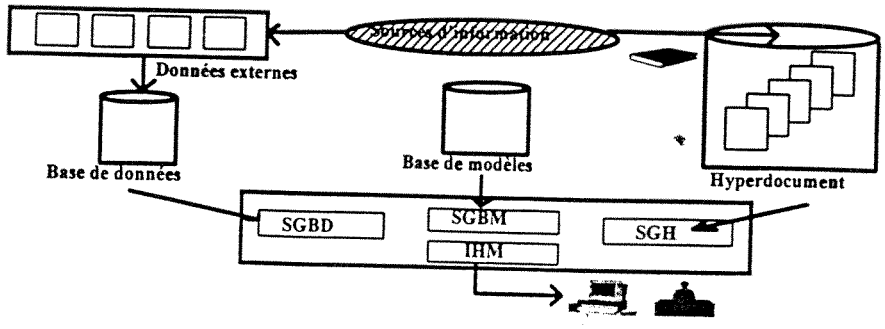
II-4 Les Systèmes Hypertextes d'Aide à la Décision (SHAD)

Nous appelons *hyperdocument* un ensemble de documents ou fragments de documents appelés *noeuds* et reliés entre eux par des *liens* (voir [PIN 90] et [BIE 90]).La *navigation* est le passage d'un noeud à un autre en fonction des liens qui existent entre ces noeuds, lors de la consultation. La classe des noeuds peut être spécialisée en deux sous-classes : les *noeuds élémentaires* et les *hyper-noeuds* constitués par un ensemble de noeuds et de liens (voir [PIN 91]. Un *système hypertexte* est un système permettant de gérer une base de documents organisée en hyperdocuments, c'est à dire possédant une gestion des références intra et inter-documents.

La vocation première du système qui va être présenté est d'apporter des informations de décision, mais il peut également être utilisé en temps que système documentaire classique, ses fonctionnalités le rendant particulièrement efficace pour la recherche de documents.

II-4-1 Architecture Générale

L'architecture du SHAD proposé prend en compte le besoin exprimé précédemment, c'est à dire, des données textuelles non structurables et intègre un hyperdocument (ou base de documents) géré par un modèle sémantique, afin d'en extraire de manière synthétique toute l'information intéressant l'utilisateur et rien que cette information. Cette architecture se présente comme suit:



Architecture du système hypertexte d'aide à la décision présenté

II-4-2 Indexation par la méthode vectorielle analysée par Foltz

Dans cette méthode, la base de données textuelle est représentée par une matrice qui croise les mots significatifs (lignes) et les documents (colonnes).

Les mots utilisés très fréquemment sont considérés comme étant non significatifs et sont éliminés de l'analyse. Chaque valeur de la matrice représente le nombre a*ij* d'apparitions du mot i dans le document j. Ainsi, les documents peuvent être perçus comme étant des vecteurs dans un espace multidimensionnel, les dimensions étant les mots utilisés pour représenter les textes, et les valeurs de la matrice étant les coordonnées de chaque document. Ces vecteurs sont également connus sous le nom de "listes inversées" (voir [TOM 93]) et ce principe de représentation multidimensionnelle est aussi utilisé pour la recherche d'images, afin de trouver dans une base d'images celles qui possèdent des configurations similaires (voir [GAR 93] cité par CHABBAT 94).

Afin de prendre en compte les documents longs et les documents courts de la même façon, les valeurs de la matrice peuvent être déterminées par la fréquence d'apparition (et non le nombre d'apparitions) d'un terme dans un document. Le problème est comment transformer cette matrice afin de gérer la sémantique des documents et de retrouver les informations désirées. (la matrice originale étant calculée par attribution plus ou moins complexe de poids aux termes selon des méthodes de calcul de poids pour les termes).

La similarité entre deus documents est obtenue en calculant le cosinus de l'angle formé par les deux vecteurs correspondants dans cet espace vectoriel, ou leur produit scalaire

	doc1	doc2	.....	docm
mot1	a_{11}	a_{12}		a_{1m}
mot2	a_{21}	a_{22}		a_{2m}
.....				
motn	a_{n1}	a_{n2}		a_{nm}

aij est le nombre d'apparitions du mot i dans le document j

Les requêtes peuvent alors être modélisées par des vecteurs de mots et être comparées aux différentes colonnes de la matrice. Les documents correspondant le mieux à ce vecteur d'entrée seront ainsi sélectionnes.

Le calcul étant fait en supposant les mots indépendants, cette approximation reste au niveau lexical puisqu'elle ne prend pas en compte la sémantique des mots.

La méthode LSI (Latent Semantic Indexing) se révèle plus efficace que les méthodes d'indexation classiques et nous allons présenter se niveaux sémantiques.

II-4-3 Les niveaux sémantiques de la méthode vectorielle

La méthode LSI est une extension de la méthode vectorielle. Elle se base sur le fait qu'il existe une structure latente inhérente à la fréquence d'utilisation des mots dans un document et que des techniques statistiques peuvent être utilisées pour appréhender cette structure (voir [DEE 90] cité par CHABBAT 94).
Un modèle de termes (mots), documents et de requêtes utilisateur, basé sur cette structure sémantique, est utilisé pour représenter et extraire l'information disponible.
LSI (Latent Semantic Indexing) décompose la matrice décrite précédemment en 100 à 300 facteurs orthogonaux (indépendants). Pour cela, elle utilise la méthode de décomposition SVD (Singular Value Décomposition ou décomposition en valeurs propres) qui permet de constituer les facteurs à partir desquels on pourra évaluer les termes et les documents (voir [GEL 84] cité par Foltz). Les facteurs sont déterminés par l'analyse des textes que l'on désire indexer. Ces facteurs sont indépendants, ce qui permet de constituer les deux matrices finales qui serviront à modéliser et extraire l'information.
Dans les deux matrices, les lignes représentent des facteurs orthogonaux : ce sont les dimensions de notre espace vectoriel E. On obtient d'une part une représentation LSI des documents dans cet espace où chaque document est un vecteur et d'autre part une représentation LSI de chaque mot. Dans cet espace vectoriel caractérisé par k facteurs orthogonaux, chaque mot et chaque document y ont leurs coordonnées.
La similarité entre deux mots ou deux documents est alors obtenue en calculant le cosinus de l'angle formé par les deux vecteurs correspondants ou leur produit scalaire

	mot1	mot2	.....	motn
fac1	m_{11}	m_{12}		m_{1n}
fac2	m_{21}	m_{22}		m_{2n}
.....				
fack	m_{k1}	m_{k2}		m_{kn}

pour les mots

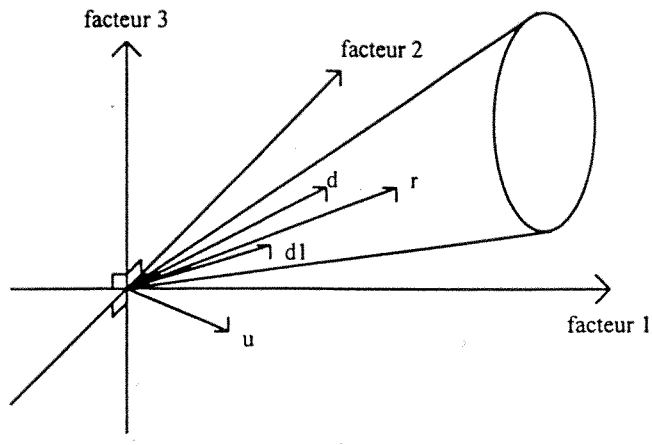
	doc1	doc2	.....	docm
fac1	d_{11}	d_{12}		d_{1m}
fac2	d_{21}	d_{22}		d_{2m}
.....				
fack	d_{k1}	d_{k2}		d_{km}

pour les documents

Avec (fac1, fac2, ..., fack) représentant la base de l'espace vectoriel de dimension k, m_{ij} représentant la coordonnée du mot j sur l'axe fac_i, et d_{ij} représentant la coordonnée du document j sur l'axe fac_i.

Une requête de l'utilisateur est transformée en pseudo-document (vecteur) en calculant sa représentation approchée dans notre espace vectoriel, puis à l'aide de la seconde matrice, on peut donc calculer la similarité entre les documents de la base et ce vecteur d'entrée. On peut donc obtenir par ordre décroissant de similarité, la liste des documents correspondants à la requête de l'utilisateur.

de la même façon, si l'on désire indexer de nouveaux documents dans notre base alors qu'ils ne figuraient pas dans l'analyse LSI de départ, et si l'on ne veut pas recalculer l'indexation générale, il est possible de les considérer de la même façon que les requêtes (pseudo-documents), et de calculer leurs représentations approchées dans notre espace vectoriel. Toutefois; il est préférable d'indexer l'ensemble des documents de la même manière et de recalculer l'indexation de l'ensemble des documents de la base lorsque trop de documents ont été ainsi indexés.



Exemple de représentation des documents; requête et utilisateur dans un espace vectoriel à trois dimensions

Pour accéder aux informations, on construit un hyperdocument de manière dynamique en suivant les directions de recherche de l'utilisateur. Dans ce but, on utilise la méthode d'indexation LSI décrite précédemment et on se situe dans l'espace vectoriel E des documents. Les axes sont les facteurs indépendants (orthogonaux). D'après ce qui précède, les comparaisons que nous ferons seront du type "comparaison entre deux documents" (la requête est considérée comme un "pseudo-document". L'espace considéré sera donc celui des documents et non des termes. les noeuds élémentaires de notre hypernoeud peuvent être typés(voir [PIN 91]): ils peuvent correspondre à un document, mais aussi à un chapitre, un paragraphe ou une phrase comme nous le verrons par la suite.
Pour l'instant nous appellerons documents les noeuds de notre hyperdocument.

II-4-3-1 Interprétation sémantique de l'espace vectoriel

Dans la méthode LSI, aucune tentative n'est faite pour interpréter d'une quelconque manière les k facteurs calculés. Toutefois, il apparaît que pour comparer plusieurs documents entre eux, on calcule leur "proximité" dans cet espace, c'est à dire les produits scalaires des vecteurs les représentant ou le cosinus de leur angle. Cette proximité est interprétée de manière sémantique dans cette méthode LSI. Les k facteurs représentant les k dimensions de l'espace vectoriel des documents peuvent donc être perçus comme des "directions

sémantiques" sans pour autant que l'on puisse exprimer en général de manière claire l'essence même de cette sémantique. C'est ainsi que nous considérons l'espace vectoriel caractérisé par ses k dimensions.

II-4-3-2 Personnalisation de l'accès

Lors de l'accès aux informations, il est essentiel de tenir compte du profil de l'utilisateur. L'idée est d'essayer d'automatiser une tâche normalement dévolue aux collaborateurs d'un décideur : la recherche des documents pouvant intéresser celui-ci. Lors d'une demande de documents de la part de ce décideur, ses collaborateurs prennent connaissance de sa direction de recherche (modélisée ici par un vecteur requête r), mais tiennent également compte de l'intérêt global qu'il a manifesté jusqu'ici pour certains domaines d'information. Pour essayer de rendre compte au mieux cet intérêt global, l'utilisateur est modélisé par un vecteur qui sera mis à jour à la fin de chaque accès. Ce vecteur représente l'intérêt général de l'utilisateur pour un certain champ sémantique (si l'on interprète les facteurs k comme étant des directions sémantiques). La requête étant également représentée par un vecteur r, u peut être ajouté à r afin de personnaliser accès et de rendre celui-ci plus proche des préoccupations de l'utilisateur.

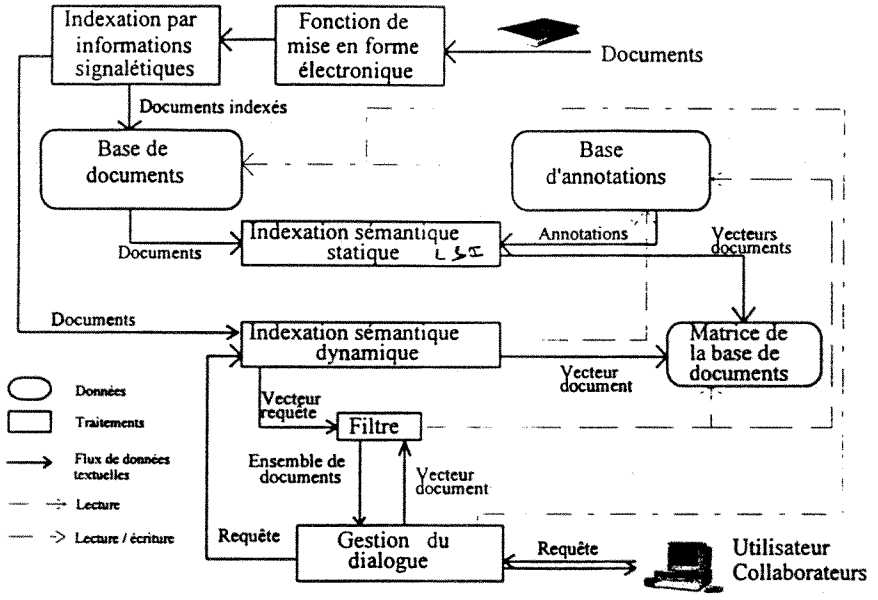
II-4-3-3 Navigation globale entre documents

On commence par déterminer le vecteur d le plus proche de r, c'est à dire celui qui vérifie : $\cos(r,d) = \max[\cos(r,d_m)]$, avec dm, ensemble des documents de la base. S'il y'a trop de documents d_i tels que $\cos(d,d_i) = 1$, alors on demande à l'utilisateur de compléter sa requête. La limite L de documents à proposer est à déterminer éventuellement avec l'utilisateur.
Ce dernier dispose des fonctionnalités suivantes :

- formulation de la requête,
- tenir plus ou moins compte des précédents accès,
- modifier complètement une requête,
- indiquer un document (le plus proche de son centre d'intérêt),
- sélectionner un document,
- consulter un document,
- consulter les documents sélectionnés,
- sortir du système,

- revenir en arrière.

Cette dernière fonction est particulièrement importante : l'utilisateur ne doit pas se perdre dans l'hyperdocument. Il est donc nécessaire de toujours conserver une "trace" permettant le retour en arrière lors de la navigation.



Flux de données lors de la navigation globale

- **fonction de mise en forme électronique** : permet de transformer un document "papier" ou éventuellement une information donnée oralement en document électronique, c'est à dire pouvant être traité avec des moyens informatiques.
- **indexation par informations signalétiques** : fonction utilisée couramment dans les systèmes documentaires classiques et permettant de classer les documents préalablement mis sous forme électronique d'après leur titre, leur auteur, ...
- **base de documents** : contient l'ensemble des documents servant à l'établissement de la matrice LSI.
- **base d'annotations** : contient les annotations relatives à chaque document de la base de documents; avec cette dernière, elle constitue l'hyperdocument.

- **indexation sémantique statique** : fonction transformant la base de documents et la base d'annotations en matrice LSI;
- **indexation sémantique dynamique** : fonction similaire transformant chaque requête de l'utilisateur en vecteur utilisable par le filtre, ou bien transformant les documents non répertoriés initialement dans la base de documents en vecteurs pouvant être ajoutés à la matrice LSI;
- **matrice de la base de documents** : matrice contenant les coordonnées de chaque document dans l'espace vectoriel considéré
- **filtre** : compare un vecteur initial (vecteur requête ou vecteur document) choisi par l'utilisateur aux vecteurs documents de la matrice et propose à l'utilisateur l'ensemble des documents les plus proches de ce vecteur initial;

gestion du dialogue : interface homme-machine permettant le dialogue entre SHAD et l'utilisateur.

II-4-3-4 Navigation particulière à l'intérieur de plusieurs documents

Si aucun document ne convient à l'utilisateur, cela signifie que le point d'entrée de l'hyperdocument est mal choisi (ou que les documents qu'il cherche ne figurent pas dans la base). Dans ce cas, il y'a trois solutions.

- on propose d'autres documents en élargissant la tolérance t de Δt. Ces documents *di* vérifiant alors : $\cos(r,d) \in [1-t, 1-t-\Delta t]$, $\Delta t>0$ étant choisi de manière à ne proposer que L documents au plus,
- on demande à l'utilisateur de reformuler complètement sa requête et le processus est repris au début,
- on demande à l'utilisateur de reformuler partiellement sa requête et on ajoute alors cette nouvelle requête r' à l'ancienne : $r \leftarrow r+r'$.

II-4-3-5 Navigation à l'intérieur des documents sélectionnés

L'utilisateur peut choisir de consulter les documents sélectionnés sans aide, c'est à dire qu'avec les touches d'édition, il naviguera à l'intérieur des documents de la façon dont il le désire.

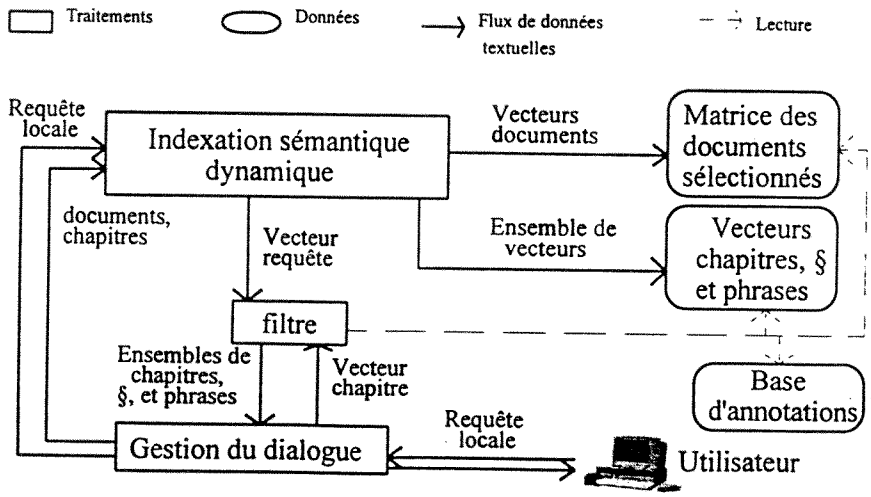
Si l'utilisateur exprime la volonté que le système l'aide dans cette phase, le processus suivra le même principe que précédemment, mais au lieu d'agir au niveau des documents, il va se situer tout d'abord au niveau des chapitres, puis des paragraphes, puis des phrases, en fonction de ce qu'il désire.

Une indexation des niveaux de structuration est effectuée. Pour avoir la meilleure efficacité, il faut rappliquer la méthode LSI aux seuls documents sélectionnés, de façon à avoir une modélisation sémantique la plus proche de la sémantique réelle de ces documents. En effet, les différentes significations possibles d'un terme peuvent n'avoir aucune utilité en regard d'une seule qui est utilisée dans ceux-ci.

Il faut donc suivre l'utilisateur dans sa direction de recherche d'informations en limitant la modélisation des champs sémantiques aux seuls documents sélectionnés.

Il faudrait se limiter au même niveau de consultation pour l'ensemble des documents sélectionnés ou n'accepter que phrases et paragraphes simultanément du fait que l'affichage simultané de titres d'un document et de phrases d'un autre document peut paraître déconcertant.

II-4-3-6 consultation des documents sélectionnés



Flux de données lors de la consultation des documents sélectionnés

Au niveau de la consultation particulière des documents seul le module d'indexation sémantique dynamique fonctionne; sa particularité sera le traitement des flux des autre données textuelles (documents, chapitres, paragraphes, et phrases).

II-5 La communication par ordinateurs ou télématique

La communication par ordinateurs ou télématique est fondée sur deux techniques qui d'une part, sont les ordinateurs, et d'autre part, les réseaux de télécommunications.

II-5-1 Le Cyberespace

"Le cyberespace qui est un mot forgé par William Gibson dans son fameux roman de science-fiction *Neuromancier*, est le nom que certains donnent à cet espace conceptuel où des mots, des liens affectifs, des données, de l'information et du pouvoir sont produits par ceux qui utilisent la télématique." [Howard Rheingold 95].
Le cyberespace est vu comme une espèce de bouillon de culture social, le réseau étant le milieu nourricier de cette culture et les communautés virtuelles, dans toute leur diversité, les colonies de micro-organismes qui s'y développent.

II-5-2 Les forums électroniques

Les forums électroniques naquirent, de manière inattendue, parce qu'ils permettaient d'utiliser les possibilités de communication des réseaux pour établir des relations de type social sans contraintes. la façon dont le commun des mortels a pu détourner des techniques de leur fonction originelle pour les mettre au service de ses besoins en communication est d'ailleurs un thème récurrent de l'histoire de la télématique.
" D'ailleurs, les mutations techniques les plus profondes ont en général été initiées par le marginaux de l'informatique, et non par l'*establishment* de cette discipline." [Howard Rheingold 95]

II-5-3 les ordinateurs "passerelles" dans les réseaux

Les ordinateurs qui servent de passerelles relient un réseau à un autre en adaptant les caractéristiques «protocoles» de l'un à celles de l'autre sans aucune perte, de manière «transparente».

Les responsables d'Internet et de Fidonet ont collaboré ces dernières années, pour relier les quelque dix mille micro-ordinateurs composant Fidonet aux millions d'utilisateurs et aux dizaines de milliers d'ordinateurs d'Internet. Fidonet étant le réseau indépendant de BBS du monde entier qui s'est progressivement constitué.
Le BBS, est l'infrastructure télématique la plus simple et la moins chère. Le magazine américain *Boardwatch* estimait à soixante mille le nombre de BBS en fonction, en Amérique en 1993.
Chaque BBS sert une population de quelques dizaines à quelques centaines de participants voire quelques milliers. L'effectif des BBS pourrait occuper des dizaines de pages (liste des différents BBS pourrait occuper des dizaines de pages) vue que les BBS sont consacrés à diverses disciplines (politique, religion, BBS pour les handicapés, les enseignants, les enfants etc...).
Cette culture du BBS s'est propagée des Etats Unis au Japon, en Europe, Amérique centrale et en Amérique du Sud.

II-6 Les Anté-serveurs

II-6-1 Définition

Les anté-serveurs sont des logiciels frontaux multiserveurs, multibases, ce sont ceux qui se rapprochent le plus de la définition des " gateways " et constituent ainsi une interface intelligente avec l'univers documentaire.

II-6-2 L'Anté-serveur Triel

II-6-2-1 Description

La conception de l'anté-serveur a été centrée autour des trois problèmes suivants :

- sélection des banques
- interprétation de la demande de l'utilisateur
- aide à la poursuite de la recherche

résultant des caractéristiques des banques de données documentaires d'aujourd'hui à savoir leur multiplicité et leur diversité.

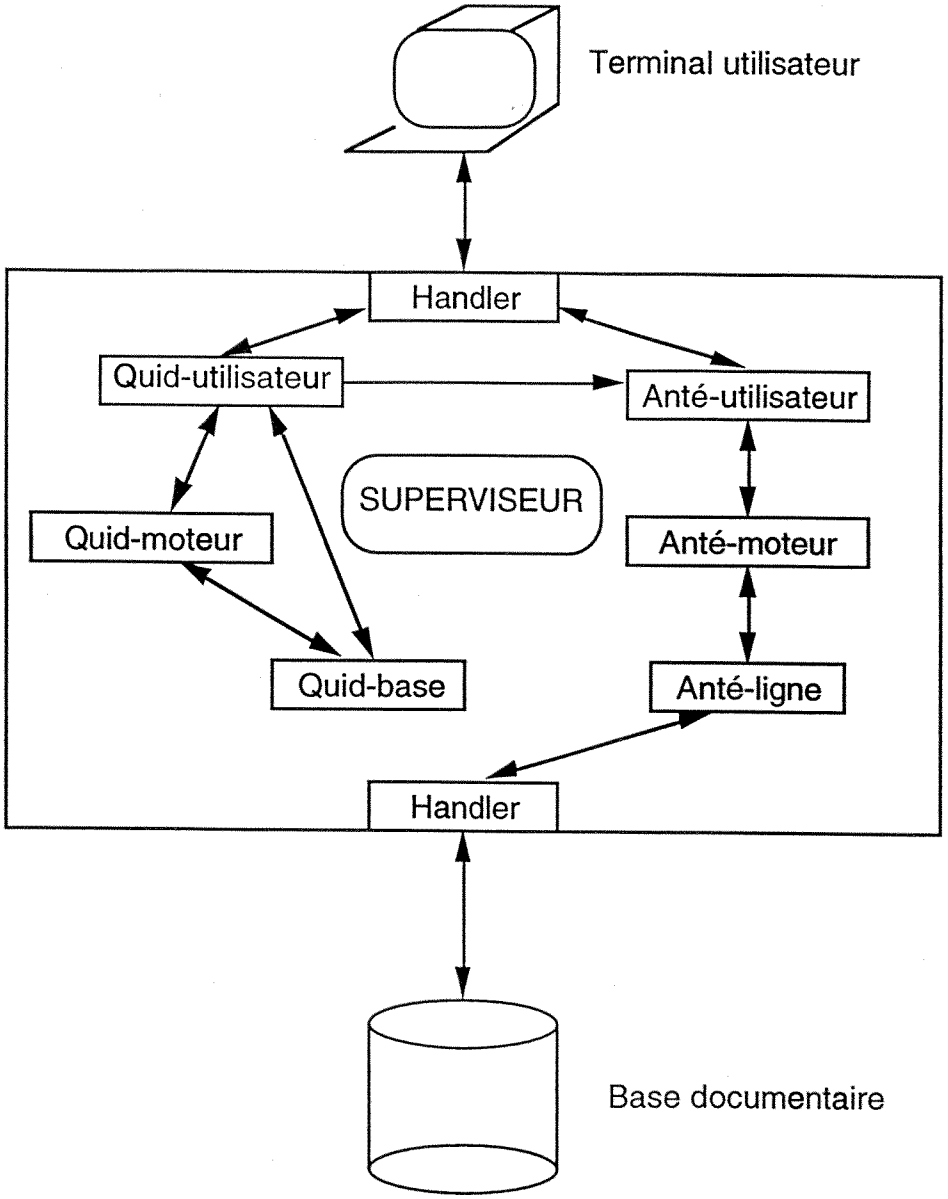
L'anté-serveur Triel a été développé par une équipe pluridisciplinaire de chercheurs de l'université de Caen, comportant, informaticiens, linguistes, spécialistes de l'information documentaire et ergonomes. Ce projet de recherche et développement est issu d'une collaboration étroite entre des laboratoires de recherche (LIUC et ELSAP) et une entreprise (TRIEL). Il a été soutenu d'une part par la région Languedoc-Roussillon (en documentation avec le CNUSC) et d'autre part le MRT.
L'esprit qui a animé l'équipe lors de la conception et la réalisation de l'anté-serveur présenté a consisté à la placer résolument du côté de l'utilisateur "profane" et occasionnel, qui cherche à obtenir des informations de banques de données documentaires.

Cet anté-serveur s'appuie sur le principe d'une "interface coopérative" ([Fisher, 1990] cité par Victorri). Ce principe permet à l'utilisateur de formuler ses besoins dans un langage naturel, en lui fournissant au fur et à mesure des informations lui permettant de comprendre la structure des systèmes d'information auxquels il s'adresse, afin de l'aider à contrôler le dialogue avec ces systèmes. L'anté-serveur développé est destiné à servir d'interface entre les utilisateurs et les banques de données documentaires présentant ainsi les fonctionnalités suivantes :

- la sélection des banques documentaires pertinentes à interroger (plusieurs banques peuvent être interrogées simultanément), grâce à une "banque des banques" faisant partie du système.
- la connexion automatique aux serveurs hébergeant ces banques, et le dialogue automatique dans les langages de commande des serveurs.
- l'analyse linguistique des requêtes des utilisateurs exprimées en formulation libre dans des champs de saisie.

l'aide à la recherche par le mise en oeuvre d'une stratégie d'interrogation et de reformulation des requêtes permettant de classer les documents par degré d'adéquation à la requête.

II-6-2-2 Architecture Générale De l'Anté-serveur



Architecture de l'Anté-serveur

Les six types de modules internes sont les suivants :

- QUID-BASE : ce module est une "banque des données". Il contient toutes les informations sur les banques documentaires accessibles, aussi bien les informations

destinées à l'utilisateur que celles nécessaires aux autres modules (mode de connexion, etc...).

- QUID-UTILISATEUR : ces modules sont chargés de l'interface utilisateur de sélection des banques documentaires. Il existe autant de modules de ce type que de types de terminaux (Minitels, PC, etc...).
- QUID-MOTEUR : ce module permet la recherche des banques pertinentes à partir d'une interrogation dans le champ DOMAINE : analyse linguistique de la demande, calcul du degré de pertinence des banques, calcul des pistes.
- ANTE-UTILISATEUR : ces modules réalisent l'interface utilisateur de traitement des requêtes aux banques sélectionnées (saisie des requêtes et autres commandes, affichage des résultats). Comme pour QUID-UTILISATEUR, ils sont spécialisés chacun pour un type de terminal.
- ANTE-MOTEUR : ce module est responsable du traitement "intelligent" de la recherche : analyse linguistique des requêtes, stratégie d'interrogation des banques, classement des réponses, prise en compte des sélections et rejets de pistes ou de documents.

ANTE-LIGNE : ces modules sont chargés du dialogue avec les serveurs : traduction des questions et autres demandes dans le langage de commande du serveur et récupération des informations renvoyées par le serveur. Il en existe autant que de serveurs différents.

II-6-2-3 Les Fonctionnalités de l'Anté-serveur

a) Choix des banques

La première phase d'une session avec l'anté-serveur consiste à choisir les banques documentaires que l'on veut interroger ([Morris et al., 1988] cités par Victorri). L'utilisateur peut sélectionner bien sûr des banques par leur nom (s'il le connaît).Sinon une étape de recherche de banques commence : l'utilisateur remplit librement un champ de saisie DOMAINE (il peut écrire par exemple *Chimie organique pour l'industrie alimentaire*). Il recevra en réponse deux listes :

- une liste de banques, triée par ordre décroissant de degré de satisfaction de la demande formulée. Pour chacune de ces banques, l'utilisateur peut accéder à un descriptif sommaire de son contenu et du serveur qui l'héberge (tarifs, etc...)

- une liste de pistes (par exemple : *biochimie, économie, articles scientifiques, thèses*, etc...) : cette liste, calculée automatiquement à partir de la demande et des banques existantes, permet à l'utilisateur de préciser sa demande, s'il n'est pas satisfait des premières banques proposées. Dans ce cas, il sélectionne une ou plusieurs pistes, et reçoit en retour une nouvelle liste triée de banques ainsi que de nouvelles pistes.
- cette étape de recherche de banques se termine quand l'utilisateur sélectionne la ou les banques qu'il veut interroger : l'anté-serveur permet en effet l'interrogation simultanée de plusieurs banques.

b) La première requête

L'utilisateur se voit alors proposer un écran de saisie dans lequel il va pouvoir exprimer son besoin documentaire ([Ogden et Sorkenes, 1987 ; Larsen, 1988] cités par Victorri) en remplissant librement des champs typés, tels que SUJET, AUTEUR, DATE, etc... . Ainsi il pourra entrer dans le champ SUJET des expressions aussi diverses que :

(1) Méthodes de détermination d'éléments traces, notamment de métaux lourds par spectroscopie d'absorption atomique dans les corps gras (beurre ou margarine)

(2) études générales sur l'influence du vide ou d'une atmosphère modifiée sur la conservation en emballage du poisson frais ou cru

(3) Articles sur le fructane dans les plantes

et dans le champ DATE

(4) d'oct. 90 à juin 91

(5) dans les trois dernières années

c) Documents, pistes et questions

Une fois la requête validée, le système va générer une série de questions booléennes à chaque banque sélectionnée. Ainsi, une requête telle que (1) générera des questions telles que :

(Q1) (*éléments* AV traces) OU (*métaux* AV *lourds*)

(Q2) *spectroscopie* ET (*absorption* AV *atomique*)

(Q3) Q1 ET Q2

etc.

L'utilisateur recevra en réponse trois types d'information :

- Une liste de titres de documents, dont les premiers correspondent aux réponses aux questions les plus "proches" de la requête utilisateur. L'utilisateur peut sélectionner dans cette liste les documents qui l'intéressent, et rejeter ceux qui ne sont pas du tout pertinents.
- Une liste de pistes, obtenues à partir d'un tri fréquentiel sur les descripteurs dans les banques des documents obtenus. Là aussi, l'utilisateur peut sélectionner ou rejeter des pistes.
- Une liste des questions posées aux banques, avec le nombre de réponses correspondants, pour chaque question et chaque banque. Ceci permet à l'utilisateur confirmé de contrôler le travail du système, et aussi éventuellement d'abandonner une banque qui s'avérerait moins pertinente pour la recherche en cours.

d) La poursuite de la recherche

A tout moment, l'utilisateur peut relancer la recherche (Salton et al., 1985 ; Radasoa, 1988 ; Bates, 1990 cités par Victorri) : le système utilise alors les informations données par l'utilisateur (sélection et rejet de pistes et de documents) pour générer de nouvelles questions aux banques. Les titres des nouveaux documents obtenus viennent alors s'ajouter aux titres déjà vus par l'utilisateur, de nouvelles pistes lui sont présentées, ainsi que la liste des nouvelles questions générées avec le nombre de réponses de chaque banque.

Cette stratégie d'interaction correspond au caractère probabiliste de la recherche documentaire, liée à la variabilité du langage, tant dans la formulation de la question que dans l'indexation des documents (Blair, 1990 cité par Victorri).

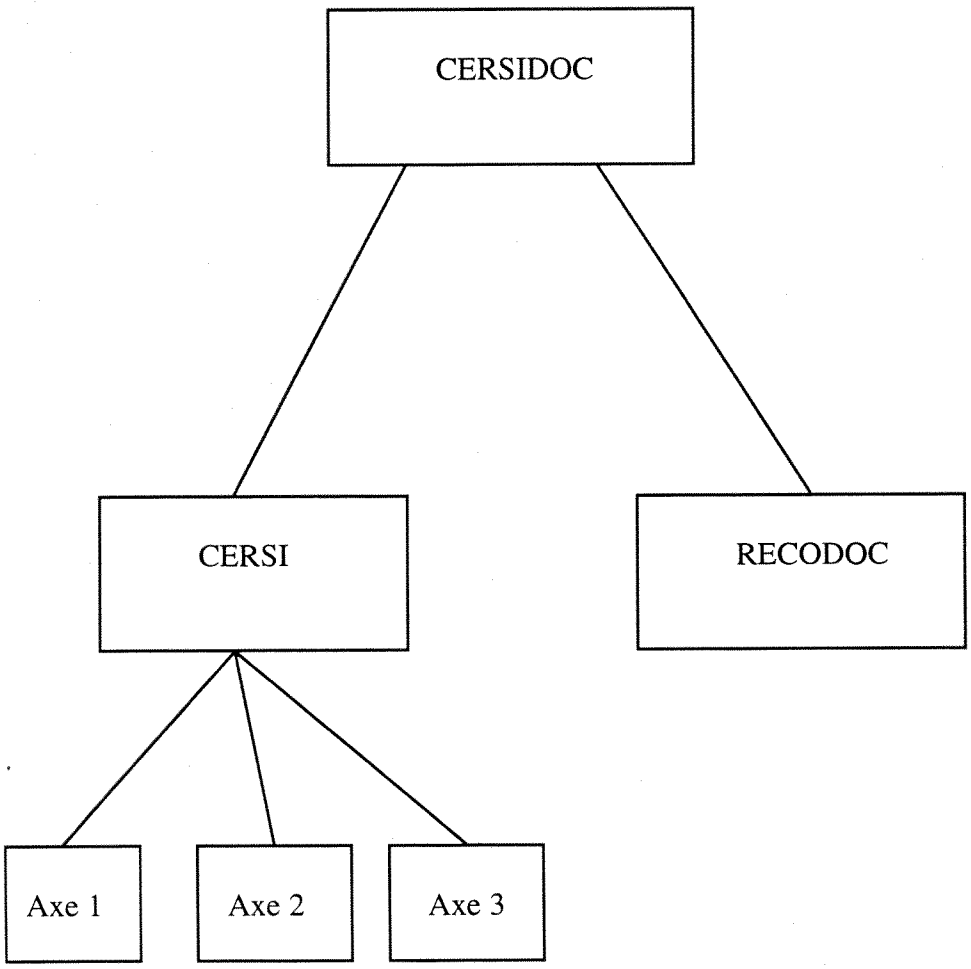
e) Services complémentaires

A tout moment, l'utilisateur peut demander à visualiser la notice complète d'un document, et utiliser les différents services fournis par les serveurs : commande de documents, téléchargement, etc... . D'autre part l'anté-serveur retient toute l'information nécessaire à la facturation : temps sur chaque serveur, nombre de documents visualisés, etc...

III) Présentation des structures constituant l'environnement de recherche et Profil-doc

Le travail à réaliser, étant un projet de mémoire se déroulant dans le cadre de Profil-doc, nous allons brièvement présenter ce dernier à travers sa création puis son intégration dans la structure globale tout en établissant un organigramme regroupant d'autres structures.

Profil-doc est un projet tournant autour de Mme Sylvie Lainé-Cruzel et de Mr Thierry Lafouge. Ce projet a été mené dans le cadre du CERSI (Centre d'Etudes de Recherches en Sciences de l'Information) en 1994, Profil-doc dépendait de l'E.N.S.S.I.B (Ecole Nationale Supérieure des Systèmes d'Informations et des Bibliothécaires). En 1995, il rejoint la structure CERSIDOC qui regroupe le CERSI de l'ENSSIB et RECODOC de Lyon I. Nous allons présenter un organigramme traduisant les relations entre les différentes structures et projets.



Axe 1 : Système d'Information et Interface , conception, organisation et représentation, sous la direction de Mr Richard Bouché.

Axe 2 : Histoire et Conservation de L'Ecrit, sous la direction de Mr Dominique Varry.

Axe 3 : Economie Management et Sociologie des services d'information, sous la direction de Mme Anne Mayère.

CERSIDOC : Fédération associant le CERSI et RECODOC

CERSI : Centre d'etudes et de recherches en sciences de l'information, est le département de l'ENSSIB. Il rassemble des universitaires et des conservateurs, dans une démarche pluridisciplinaire. Il compte une vingtaine de membres et accueille autant de doctorants. Le CERSI croise les travaux d'historiens du livre, d'informaticiens, d'économistes, de sociologues, de linguistes, de psychologues, etc ... pour faire avancer la recherche dans trois directions principales, à savoir :

- la conception, l'organisation et les représentations des systèmes d'information et de leurs interfaces,
- L'histoire et la mise en valeur du patrimoine graphique,
- l'économie, le management et la sociologie des services d'information.

L'activité de recherche est organisée sous forme de projets ou de contrats de recherche. Parallèlement, des rencontres et conférences, des séminaires d'études, des colloques internationaux témoignent de la volanté d'ouverture et d'échange du CERSI. De nombreux articles et plusieurs ouvrages de référence ont été publiés par ses membres. Les orientations du centre sont discutées au sein du conseil scientifique de l'ENSSIB.

RECODOC : **RE**présentation des **CO**nnaissances et **DOC**umentation de l'université Claude Bernard Lyon 1, rassemble des universitaires d'origine scientifique (informatique, physique, pharmacie, mathématiques). Les principaux centres d'intérêt sont : l'information scientifique et technique. L'équipe RECODOC s'intéresse aux domaines de recherche suivants :

- structuration,caractérisation, diffusion et évaluation de l'information scientifique et technique dans un contexte de mise en réseau (INTERNET),
- les nouveaux systèmes d'information : modélisation et conception,
- besoins et pratiques informationnels des chercheurs et professionnels scientifiques.

L'objectif essentiel du projet Profil-doc de l'équipe RECODOC est la minimisation du bruit. En effet, ce dernier est un facteur très gênant pour l'utilisateur lorsque le volume des réponses dépasse un certain seuil "tolérable". La réalisation de cet objectif, s'appuie sur l'établissement d'un pré-aiguillage vers un corpus d'informations restreintes aux informations pertinentes pour l'utilisateur selon des critères d'usages. Ce système de pré-aiguillage s'appuie sur trois composantes fondamentales , à savoir :

- un découpage des textes à traiter en fonction d'une structuration établie sur des fonctions d'usage, c'est à dire caractérisant des parties du texte en fonction des usages qui pourront en être faits et donc des types d'utilisateurs qu'elles peuvent intéresser,
- une caractérisation du profil de l'utilisateur, qui détermine à la fois ses compétences et pratiques usuelles, mais aussi des informations liées aux objectifs d'une recherche spécifique,
- un système d'aiguillage qui établira le lien entre une description de profil d'utilisateur et des propriétés caractérisant des unités documentaires, pour définir un corpus "personnalisé".

Cette chaîne du traitement du document textuel, se situe en amont, et permettra une pré-sélection du matériau textuel à exploiter par le logiciel documentaire.

Les motivations de recherche d'un utilisateur, ses besoins, etc ... étant pris en considération, un profil utilisateur peut être établi, ce profil sert d'aiguillage vers un corpus personnalisé d'informations de documents par apport à un type d'utilisateur déterminé.

Profil-doc progresse en s'appuyant bien sûr, sur les performances du système Spirit (qui sera décrit ultérieurement), tout en lui attribuant des techniques permettant de réaliser l'objectif visé. Ainsi, nous pouvons citer les travaux de :

- Nabil BEN ABDALLAH (thèse à soutenir en 96) : consistant en le découpage de documents en unités documentaires (un paragraphe est consacré plus loin à cette partie), ainsi que la détermination des profils utilisateur, détermination des différentes étapes du processus de recherche etc...,

- Christine.MICHEL (D.E.A juillet 95) : consistant en l'étude de l'influence d'un découpage logique (par chapitres, paragraphes, etc) ou physique (page par page) sur la recherche documentaire. Il traite plus particulièrement, de l'influence qu'il a sur les pondérations des termes descripteurs. Différentes méthodes statistiques ont été introduites, la pondération et la mise en place d'indicateurs dévaluation de différents paramètres de la chaîne de traitement,

- Marc.JOUINEAU (D.E.A à soutenir en 96) : sur la prise en compte de données sémantiques dans une structure SGML

- Ce présent travail : consistant à structurer de l'information dans Spirit pour l'utiliser sous sa meilleure forme, établir des liens entre les différentes unités documentaires, réaliser une navigation entre différents corpus qui seront présentés plus tard afin de pouvoir faciliter la tâche à l'utilisateur et de le guider dans sa recherche.

Les détails seront exposés au fur et à mesure de l'avancement dans ce présent mémoire.

IV) Présentation du sujet

IV-1 Objectif du travail

L'objectif du projet profil-doc est de restreindre le champ de recherche documentaire d'un utilisateur en ne lui fournissant que ceux qui répondent à son besoin. Mon travail dans ce cadre, consiste à structure de manière optimale les données internes dans le système Spirit, et d'organiser l'enchainement des traitements . Les questions de l'utilisateur portant sur des sous-ensembles de documents (comme nous le verrons plus loin) constitués d'unités documentaire, il faudra établir des stratégies de conception et d'organisation afin de pouvoir manipuler ces unités documentaires lors d'une recherche en les liant (les reliant éventuellement), c'est à dire les combiner d'une façon rigoureuse et logique.
Ainsi, il faut réussir à gérer efficacement les étapes séparant la formulation et la satisfaction, qui pour l'utilisateur constituent successivement les portes d'entrée et de sortie du système Spirit.

IV-2 Principales caractéristiques de Spirit

IV-2-1 Qu'est ce que SPIRIT, structure générale:

SPIRIT est un système d'indexation et de recherche d'informations textuelles entièrement automatisé. L'interrogation se fait en Langage Libre et les documents réponses sont présentés par ordre de pertinence. SPIRIT utilise pour cela des traitements linguistiques et statistiques portant sur le texte intégral.

Les TRAITEMENTS LINGUISTIQUES jouent un rôle essentiel car ils permettent de rapprocher automatiquement les formes conjuguées et la forme infinitive ("mettre" et "mis"), de distinguer les homographes ("livre" nom ou verbe, "car" nom ou conjonction), de détecter les locutions ("chemin de fer"). En effet, le rôle du traitement linguistique dans la recherche documentaire est d'identifier les entités que l'on manipule et de déterminer leurs propriétés linguistiques sachant que les documents à traiter sont de nature très diverse, et qu'une question peut être formulée de plusieurs manières et en langage naturel, sans aucune contrainte syntaxique, et sans aucune limitation dans sa longueur. " *Le traitement linguistique, doit assurer la transformation de la question en des entités identiques à celles utilisées dans les documents afin que la comparaison puisse se dérouler dans de bonnes conditions* " [Radasoa 88]. Le traitement linguistique de Spirit est indépendant de la taille de

l'information (documents, question), de la typographie utilisée (riche ou pauvre), et du domaine entrepris. Ce traitement regroupe plusieurs opérations élémentaires assurant des tâches et assurées chacune par un module. Un tel traitement modulaire, réalise le bon fonctionnement global et permet d'éviter toute perturbation de ce dernier dans le cas d'une défaillance de l'un des modules. En plus de ces points importants relatifs au fonctionnement, cette structure de modules indépendants offre la possibilité d'améliorer l'un d'eux voire le remplacer en cas de besoin. L'organisation de ce traitement permet la séparation de la programmation informatique, des données linguistiques, ce qui permet de manipuler des bases dans différentes langues comme on le verra ultérieurement dans le paragraphe "structure de données".
Une analogie, peut être faite ici, confirmant cet avantage, avec les systèmes experts dont le moteur d'inférence est indépendant des règles de production. Nous citons enfin les fonctions linguistiques réalisées par le système Spirit à savoir :

- le découpage,
- l'analyse morphologique,
- la reconnaissance des expressions idiomatiques (l'exemple ci-dessus "chemin de fer" en est une),
- l'analyse syntaxique,
- l'élimination des mots vides
- la normalisation

Rem: Le découpage est pris dans le sens Spirit comme consistant à reconnaître les mots constituant un document ou une question grâce à l'utilisation d'un automate d'état fini qui permet de prendre en compte les séparateurs spéciaux ou ambigus. Le découpage de documents au sein de Profil-doc est une toute autre tâche qui sera exposée dans le paragraphe "Méthodologie" du IV).

Les TRAITEMENTS STATISTIQUES donnent aux mots un poids informationnel d'autant plus important qu'ils sont plus rares dans la base. Cette notion de poids informationnel, rejoint le calcul des fonctions de poids à partir de l'entropie du mot (ou du concept) calculée sur le corpus. En effet, cette fonction de poids permet d'évaluer le pouvoir informationnel du concept, plutôt que la précision sémantique de ce dernier pour la discrimination des

documents en recherche documentaire. La quantité d'information véhiculée par un concept donné dans le corpus est d'autant plus faible que le concept est uniformément réparti sur le corpus. Ce fait peut être mis en évidence par l'entropie du concept qui varie à l'inverse de la quantité d'information véhiculée. Si l'on considère un fond documentaire contenant N documents $d_1, d_2, d_3, ..., d_i, ..., d_n$ avec :

L : valeur limite de la fonction de poids sémantique. L est ajustée expérimentalement afin que l'on obtienne une bonne hiérarchisation des réponses aux questions documentaires qui n'est autre que le facteur "pertinence". Cette fonction de poids est d'autant plus grande que le concept est précis ou informationnel.

$\mathbf{C_i}$: concept donné dans le corpus.

On a :

$$H(d/ci) = -\sum p(dj/ci)Log2(dj/ci)$$

- Si la répartition est uniforme sur l'ensemble du corpus,

  on aura $H = -1/N \sum Log2\ 1/N = Log2N$

  car $p(dj/ci) = p(dj) = 1/N$

- Si le mot n'existe que dans un seul document dk, on aura :

  $P(d_j \backslash c_i) = 0 \quad \forall j \neq k$

  $P(d_k \backslash c_i) = 1 \Rightarrow Log_2 (P(d_k \backslash c_i)) = 0$

  $\Rightarrow H = 0$

La fonction de poids sémantique mesurant le pouvoir informationnel du concept, crée à partir de H est :

$$f(c) = L - (H(d/c))/Log_2(N.P_N(c_i))$$

après une suite de calcul et en posant $PN(c_i)$ = fréquence moyenne du concept ci sur l'ensemble du corpus, on obtient :

$$H(d/c_i) = (H((c_i/d))/N.P_N(c_i)) + Log_2(N.P_N(c_i))$$

Rem: On peut calculer aussi des fonctions de poids à partir des champs sémantiques élaborés sur le corpus. Cette méthode ainsi que la précédente sont plus fiables car le calcul s'effectue à partir du corpus et la signification de la fonction de poids dépend en particulier de la dimension du corpus et de sa représentativité du domaine. D'autres méthodes sont

utilisées pour le calcul des fonctions de poids mais faisant appel à des outils externes au corpus tels que les dictionnaires de spécialités, les thesaurus contenant les relations génériques-spécifiques, qui ne font pas l'objet de notre étude mais qui sont intéressants et détaillées dans la thèse de C.FLUHR 88. Il est quand même utile de mentionner que ces méthodes permettent d'avoir une statistique plus juste des concepts relativement à d'autres méthodes telles que, grammaticales etc... , puisqu'elles introduisent une analyse statistique utilisant un moyen de faire apparaître la sémantique dans les fréquences.

La taille de la base consultée et la longueur de la question ont peu d'influence sur le temps de réponse de SPIRIT. Plus une question est détaillée, meilleure est la réponse. Les pages d'un document sont présentées par ordre de pertinence[2] .

SPIRIT se distingue des autres systèmes ne comportant pas de traitement linguistique et dans lesquels tous les mots indexés ont la même importance, ce qui provoque d'importants phénomènes de bruit[3] et de silence[4] . Dans ces systèmes, l'utilisateur doit poser des questions successives avec des opérations plus ou moins complexes. Tous les documents-réponses sont présentés au même niveau, non classés et les documents longs doivent être entièrement parcourus.

Grâce à SPIRIT, le travail de recherche des informations se fait automatiquement et reste transparent pour l'utilisateur.

Le COMPARATEUR agit sur deux niveaux :

- sur les documents, afin de sélectionner ceux qui sont pertinents,
- sur les pages des documents proposés, afin de localiser les pages les plus pertinentes.

---

[2] *Le mot "pertinence" n'est pas pris au sens propre philosophique mais au sens du modèle statistique de* SPIRIT

[3] *Proportion de documents qui sont soumis par le système bien que ne répondant pas à la question.*

[4] *Proportion de documents qui ne sont pas fournis par le système bien que répondant à la question*

Le rôle du comparateur est de rechercher les intersections entre les documents de la base et la question, en manipulant des entités de même nature (c'est à dire, comme il a été cité antérieurement, normalisés) en utilisant un modèle statistique.
Le comparateur élabore une description de l'intersection entre les documents de la base et la question à l'aide d'une liste de mots (simples ou composés) reliés par les opérateurs booléens ET AVEC UN LIEN DE DEPENDANCE ou/et ET. Quand la comparaison se termine, les documents proposés qui ont la même intersection, sont regroupés dans une même classe. Cette intersection est en fait la meilleure question booléenne qu'on aurait pu poser pour obtenir cette classe avec les opérateurs autorisés.

IV-2-2 Structure de données

IV-2-2-1 Structure de la base

a) structure logique

La structure logique de la base est définie par :

- la langue : Comme il a été cité dans le paragraphe précédent, SPIRIT est actuellement opérationnel sous trois langues: Français, Anglais et Allemand. Il est cependant conçu pour intégrer toute nouvelle langue.

- Le nombre de champs

- Les noms de champs

- Les types de champs : Il existe trois types de champs :

- complémentaires du texte, telles que les noms d'auteurs, les titres, les dates, les références, etc... ; ils servent généralement à faire des recherches Les champs factuels : Ils contiennent les informations de type booléen.

- Les champs textuels : Ils contiennent du texte sur lequel SPIRIT exécute ses algorithmes linguistiques et statistiques. Les questions en langage libre s'appliquent à ces champs.

- Les champs non-interrogeables : Ils peuvent apparaître à la visualisation mais leur contenu ne sera pas un sujet de recherche. Ceci est dû à la limite de la grille du système Spirit en nombre de zones.

Rem: Tous les champs sont de longueur variable, et aucun n'est obligatoire, à l'exception du champ référence du document.

b) Découpage : principes généraux

Il s'agit de l'alimentation de la base. L'information contenue dans une base de données, est découpée en unités documentaires (ou documents). A une question posée, SPIRIT propose un certain nombre de documents que l'utilisateur peut visualiser. la notion de document est variable d'une base de données à l'autre. Toutefois, l'administrateur de la base doit effectuer le découpage en tenant compte des règles générales suivantes :

- Chaque document doit avoir un contenu homogène.

- Les documents trop courts dispersent l'information dans la base et la rendent difficile à retrouver.

- Les documents trop longs ne permettent pas d'utiliser au mieux les mécanismes d'optimisation de la recherche.

IV-2-2-2 Structure d'un document

Un document peut couvrir des réalités diverses, et les informations qu'il contient sont de nature différente. Ces informations sont rassemblées dans des unités appelées champs (définis antérieurement). Les documents que l'on souhaite rajouter à la base doivent se

présenter, soit sous format WINWORD, soit sous un format DOS texte. Avant d'être insérés, les documents doivent être préparés dans un format SPIRIT inclus dans l'éditeur de texte. Chacun des champs définis précédemment structure le document pour la recherche, des sauts de pages peuvent être définis pour faciliter la consultation et enfin des renvois et signets peuvent être insérés par l'utilisateur pour permettre une navigation de type hypertexte, entre les divers documents (d'une même base), ou bien à l'intérieur d'un même document. Une fois le document traité, il ne reste plus qu'à lancer la "fonction d'intégration de données". Le système va alors faire l'analyse linguistique et statistique du document et engendrer les liens sur les fichiers inverses.

IV-2-3 Consultation :

IV-2-3-1 Définition

Spirit Consultation permet de consulter facilement des Bases de Données Spirit en langage naturel. Aucun langage d'interrogation n'est nécessaire.
Cette tâche consiste en la recherche de documents dans une base et s'effectue selon diverses possibilités à savoir :

- par référence, ou identification unique, du document ,
- par le contenu des champs factuels et/ou textuels des documents, en utilisant des grilles d'interrogation,
- par le contenu strictement textuel des documents, à partir d'une question en langage naturel indépendamment de la forme et de la longueur.

IV-2-3-2 Sélection de la base

L'accès à une base, nécessite sa sélection préalable. Cette sélection se fait par le biais de l'affichage de la boite de dialogue "sélection de la base" qui suit automatiquement l'étape d'identification (elle même réalisée par l'affichage de la boite de dialogue "Identification").
Cette boite de sélection contient la liste de toutes les bases disponibles à la consultation, ainsi il suffira de sélectionner la base voulue et de valider cette opération.

Le système Spirit affiche, suite à cela, une barre d'outils constituée de boutons correspondant au menu principal. Les boutons ainsi que les tâches relatives sont présentés comme suit :

Bouton	**Tâche à réaliser**
Base	permet de sélectionner la base active
Infos	permet d'afficher des informations sur la base active, telles que la langue utilisée, le nombre de documents contenus, le nombre de champs, la date de la denière mise à jour, et la liste des champs.
Quest	permet d'interroger la base active par son contenu textuel en utilisant la question en langage naturel.
Grille	permet d'interroger la base active par les contenus des champs factuels et textuels.
Select	permet de retourner le classement de documents après avoir sélectionné des classes à visualiser.
Visu	permet de visualiser un document en accès direct
Aide	permet d'obtenir des informations sur le système d'aide
Fin	permet de quitter le système le système Spirit.

IV-2-3-3 Recherche de documents dans une base : elle se fait selon trois méthodes possibles.

- Accès direct à un document : Cette opération n'est réalisée que si la référence exacte du document est connue sachant en plus que la différence entre minuscule et majuscule est significative et que la taille de la référence ne peut dépasser 50 caractères. Comme il a été mentionné précédemment, la visualisation du document se fera par l'intermédiaire du bouton "Visu", ainsi la boite de dialogue "Accès à un document" s'affiche, la référence étant saisie, l'opération validée, le document sera visualisé. Dans le cas où ce dernier n'existe pas, Spirit, affiche le message suivant "Le document n'existe pas".

- Interrogation par la Grille : La recherche de documents s'effectue à partir des contenus factuels et textuels de ces derniers. Spirit propose par défaut une grille contenant sept zones, où chaque zone est un champ ou un ensemble de champs ayant les mêmes propriétés (exemple : type factuel). Elles contiennent éventuellement des valeurs prédéfinies. Le nom de la grille est par défaut celui de la base active. Il existe une correspondance entre chaque zone de la grille et un champ de la base. Cependant une seule zone "texte" regroupe tous les champs textuels de la base.

Rem: L'utilisateur peut créer sa propre grille, elle remplacera alors la grille par défaut et devient la grille courante.

La tâche "interrogation par la Grille" est réalisée grâce au bouton "Grille" (comme il a été mentionné précédemment). Ainsi, une fenêtre s'affiche avec une barre menu offrant la possibilité d'effectuer une recherche, de modifier ou de créer des grilles et enfin de retourner au menu précédent.
Cette tâche, pour être réalisée, doit se baser sur des critères de recherche, à savoir :

la spécification du contenu d'au-moins une zone,

- le genre d'information à entrer dépendant du type de la zone.

Pour une zone factuelle, il y'a possibilité d'entrer une série de valeurs et de préciser si ces valeurs doivent être toutes présentes dans un champ, ou seulement quelques unes.

Une entrée de mots séparés par des blancs, déclenche une recherche sur tous les documents dont le champ contient tous les mots de la série, ce qui correspond au "ET" logique.

Une entrée de mots séparés par des virgules, déclenche une recherche sur au moins un parmi les mots de la série, ce qui correspond au "OU" logique.
Une zone factuelle peut contenir jusqu'à 3050 caractères.

Rem:
si la recherche ne s'appuie que sur les champs factuels, et qu'au moins un document corresponde à la recherche, Spirit indiquera le nombre de documents-réponses.

si la recherche implique aussi la zone textuelle, et s'il y'a au-moins un document-réponse, Spirit affichera un classement de réponses (par degré de "pertinence" comme il a déjà été mentionné au paragraphe 1-1 du III).

si la recherche n'aboutit pas, Spirit affichera le message suivant "Aucun document ne répond à la question". Dans ce cas, la recherche peut être élargie en jouant de nouveau sur les critères de sélection.

- Interrogation en langage libre: La recherche de documents s'effectue par leurs contenus textuels. La question est posée en langage naturel, Spirit en extrait les mots qui lui sont significatifs (informationnels) ainsi, la recherche portera sur les mots significatifs dont les normalisations[5] se trouvent dans les dictionnaires de la langue.
  Le système Spirit permet, pour éviter d'éventuelles erreurs introduites dans les mots de les visualiser.

Rem :

[5] *un mot est normalisé lorsqu'il est ramené à sa forme de base (singulier ou infinitif). Les mots significatifs dont les normalisations figurent dans les dictionnaires courants, s'affichent dans la zone "mots clés"*

Les questions seront d'autant mieux analysées qu'elles seront posées en caractères minuscules avec les accents. Les majuscules étant conseillées pour les noms propres et à chaque début de phrase. (Il s'agit de l'écriture de la question selon les règles de la langue prise en compte)

IV-2-4 Caractéristiques des questions Spirit

IV-2-4-1 Question en Langage Libre sur les champs textuels :

L'utilisateur peut poser une question sans aucune contrainte. Le système SPIRIT se charge alors de la détection des éventuelles erreurs typographiques et d'accentuation, de la normalisation des mots, de la recherche des mots composés. Après ce travail, SPIRIT recherche dans la base les documents répondant à la question et les classe en fonction du poids informationnel des mots composés et des mots qui ont provoqué leur sélection. la réponse à une telle question est une liste de documents classés du plus pertinent au moins pertinent.

IV-2-4-2 Hypertexte dynamique :

SPIRIT offre la possibilité de remplacer une question en langage libre par n'importe quelle partie d'un document affiché.

IV- 2-4-3 Grille d'écran multi-critères :

Cette option de SPIRIT permet de définir une grille d'écran sur laquelle figurent les différents critères d'interrogation. Chaque critère peut recouvrir un ou plusieurs champs de la base de données. Si le critère est de type textuel, la question sera traitée comme une question en langage libre. S'il est de type factuel, la question sera traitée comme une question booléenne.

IV-2-5 Définition des liens Spirit

Les liens sont définis par les renvois et signets, ils permettent une navigation hypertexte à la consultation.

IV-2-5-1 Définition d'un signet :

a) Syntaxe

$DE{nom du signet}FE

- le signet est placé n'importe où dans le document,
- {nom du signet} est une chaîne alpha-numérique de 1 à 130 caractères,
  le nom du signet doit être unique dans la base.

b) Fonction

L'insertion d'un signet dans un document Spirit, permet de marquer et donner un nom à un emplacement spécifique. Cela veut dire que ce marqueur déclare une étiquette de point d'arrivée dans un document de la base.

L'association de ce signet à un mot ou à un ensemble de mots, crée un *renvoi texte*.

Rem: Lors de la consultation, les renvois d'un signet d'un document s'affichent en bleu. En cliquant sur un renvoi, l'utilisateur peut passer directement à l'emplacement du signet associé, qu'il soit dans les mêmes documents, ou dans un autre document de la même base.

IV-2-5-2 Définition d'un renvoi

Un renvoi est l'association d'un mot ou d'un ensemble de mots, et d'un signet ou objet qui permet de faire un saut entre les deux lors de la consultation de Spirit.

a) Renvoi vers un signet interne

a1. syntaxe :

$DR{texte du renvoi}$FRI{nom du signet}$FF

C'est un marqueur placé n'importe où dans le document, avec :

{texte du renvoi} : Chaîne alpha-numérique de 1 à 254 caractères représentant le libellé du renvoi.

{Nom du signet} : désignation de l'endroit de branchement du renvoi.

a2. fonction :

Le marqueur ainsi défini, permet de déclarer un renvoi qui constitue une étiquette de point de départ depuis un document de la base vers un point d'un document quelconque de la base.

REM :

- plusieurs renvois peuvent pointer vers le même signet,
  le marqueur renvoi est couplé avec le marqueur signet qui définit le point d'arrivée du renvoi

b) Renvoi vers un objet externe

b1. syntaxe :

$DR{texte de renvoi}$FRE{type}{référence objet}$FF

C'est un marqueur placé n'importe où dans le document, avec :

{texte de renvoi} : chaîne alpha-numérique de 1 à 254 caractères représentant le libellé du renvoi.

{Type} : chaîne de 3 caractères représentant l'extension du fichier contenant l'objet.

{Référence objet} : désigne le nom logique de l'objet .

Ce dernier paramètre est un nombre compris entre 1 et 99999999, unique dans le document, et doit préalablement avoir été déclaré par un marqueur de "définition et insertion d'un objet externe", permettant de définir d'une part un objet externe à insérer dans un document de la base, et d'autre part le fichier contenant cet objet.

Cette séquence de marqueur est couplée au marqueur considéré (c'est à dire, marqueur "définition d'un renvoi à un objet externe") qui définit, dans le document, le point d'insertion de l'objet.

b2. fonction :

Le marqueur "Définition d'un renvoi à un objet externe" permet de déclarer un renvoi constituant un point de départ depuis un document de la base vers un objet externe à la base

IV-2-6 Navigation

IV-2-6-1 Types de navigation

a) navigation dans les documents d'une classe

Cette tâche est réalisée par Spirit permettant ainsi de naviguer de document en document, ou de page en page.
Spirit distingue deux types de pages :

- les pages physiques correspondant aux pages réelles
- les pages informationnelles correspondant aux pages classées par ordre de "pertinence".

La navigation entre documents d'une même classe s'effectue par le biais de la fenêtre "Actions", offrant les actions suivantes :

Docs : déplacement de document en document

Pages : déplacement de page physique en page physique

Info : déplacement de page informationnelle en page informationnelle

b) Déplacement de document en document

Spirit offre la possibilité à l'utilisateur de se déplacer entre les documents, les pages physiques, ainsi que les pages informationnelles. Pour chaque action Spirit dispose de boutons réalisant des tâches bien déterminées, tels que

- aller sur le dernier document
- aller sur le premier document
- aller sur le document précédent
- aller sur le document suivant

De même, par analogie, le même principe fonctionne pour les pages physiques et les pages informationnelles.

IV-2-6-2 Principes de navigation

La navigation dans Spirit, ou moyen d'accès à une destination déterminée à partir d'un point de départ est réalisée par l'hypertexte. Ce dernier est un mot relativement récent dans la recherche documentaire (apparu dans [CON 87] cité par Radasoa 88). Cependant son rôle est de plus en plus important et son utilisation de plus en plus répandue dans le domaine pour les raisons suivantes :

- la navigation est facilitée dans la base en passant rapidement d'une partie d'un document à un autre, ou d'un document à un autre,

- l'utilisateur est guidé grâce aux liens établis dans l'hypertexte. Ces liens peuvent être de divers types permettant ainsi de référencer différentes catégories d'information (texte, résumé, bibliographie, etc ...).

La consultation d'un hypertexte est facilitée par la structuration modulaire de l'information, ainsi il trouve son application dans tous les types de documentation structurés.

Spirit dispose de deux sortes d'hypertexte, ou renvois, et permet après avoir navigué de retourner sur le document de départ.

a) L'hypertexte statique

C'est un renvoi permettant de se positionner directement à un signet (constituant un point d'arrivée), dans une autre partie du même document ou d'un autre document. Les liens sont définis et péétablis lors de la saisie ou la mise à jour du document.

Rem :

- Si le renvoi est associé à un signet, Spirit affiche le document contenant le signet, à l'emplacement de ce dernier, le signet restant invisible.

- Si le renvoi est associé à un objet, Spirit lance l'application gestionnaire de l'objet et ouvre le fichier lui correspondant.

b) L'hypertexte dynamique

Lors de la consultation d'un texte en cours de visualisation, une nouvelle question à partir de ce dernier peut être exprimée et générée, et ce, grâce à l'hypertexte dynamique.
La navigation s'effectue donc de document en document à partir de texte contenu dans chaque document.

IV-3 Méthodologie

IV-3-1 Formalisme

La base de données est formée d'un ensemble de documents D.
D_j est un document de la base. Chaque document est découpé en unités documentaires. U_{ij} est l'unité documentaire appartenant au document D_j.
C est l'ensemble des unités documentaires dans la base.
Des propriétés sont associées aux documents (nom et domaine scientifique de l'auteur, son affiliation, domaine d'intérêts scientifiques et techniques du journal, titre du document, etc...).
Des propriétés sont associées aux unités documentaires. Elles se réfèrent :

- au style du texte (argumentatif, descriptif, ...)
- aux différentes formes de représentation d'information (Schémas, équations mathématiques, ...) etc ...

On note P_{ij} les propriétés appliquées à U_{ij}, issues des propriétés de D_j et des propriétés spécifiques à U_{ij}.

IV-3-2 Découpage de documents

Le découpage des documents en unités documentaires, conduit à la production de sous-ensembles de documents sur lesquels porteront les questions. Ce découpage est traité en fonction d'une structuration établie sur des fonctions d'usage. Les parties du texte ainsi obtenues, seront mises en relation avec des types d'utilisateurs intéressés par l'intermédiaire d'une fonction d'aiguillage. Le découpage dans l'exploitation de la structure logique du document, s'appuie sur deux niveaux :

- niveau 1 : découpage en chapitres,sections, paragraphes, sous-paragraphes, etc..., c'est à dire, en unités documentaires.
- niveau 2 : contenu, c'est à dire le type du texte. Il consiste en l'attribution de propriétés à l'unité documentaire.

"*L'éclatement du document en unités documentaires nous permet tout en préservant l'unité globale du document (le lien entre l'unité documentaire et le document auquel elle appartient) de présenter à* l'utilisateur *une information plus affinée, facile à saisir*" [LAINE 94].

Le découpage en unités est basé sur la fonction remplie par chaque unité documentaire et non sur son contenu. Bien que plusieurs points soient flous, des structures éclaircissant les différents types d'unités d'un document ainsi que les différents liens s'établissant entre-elles seront présentées dans la partie Développement. Nous allons, avant d'entamer la pertie développaement déterminer les paramètres carctérisant :

- l'utilisateur (caractéristiques),
- les documents.
- les unités documentaires (propriétés).

Ces paramètres sont entrepris par l'équipe du projet profil-doc et extraits de l'article à paraître dans "Information Processing & Management 1996"

1) les champs référenciels

PA : référence du document

PB : référence de l'UD dans le document

2) Champs contenant des propriétés liées au document

PC : titre du document

PD : auteur du document

PE : co-auteurs (libellé)

PF : affiliation de l'auteur (libellé)

PG : code nationalité auteur

PH : année

PI : type de l'environnement éditorial (actuellement, les valeurs possibles répertoriées sont :

M3 pour les mémoires 3e cycle,

PPROF pour les articles publiés dans des revues professionnelles,

PGP pour les articles de presse grans public,

PFOND pour les articles extraits d'une revue internationale à comite de lecture (fondamentale, recherche ou théorique)

DIVERS

Liste d'autorité en cours d'élaboration

PJ : profession de l'auteur (dans le corpus actuel, les valeurs possibles sont :

E3 pour les étudiants,

SPE pour les spécialistes du secteur concentré,

J pour les journalistes,

DIVERS pour les autres)

Liste d'autorité en cours d'élaboration

PK : champ disciplinaire de l'auteur (dans le corpus actuel :

SIC pour la recherche en sciences de l'information,

BIBL pour les bibliothécaires et documentalistes,

INFO pour les informaticiens,

MATHS pour les mathématiciens,

ECO pour les économistes et gestionnaires,

SOCIO pour les sociologues,

DIVERS

Liste d'autorité en cours d'élaboration

PL : Communauté à laquelle appartient l'auteur (codage affiliation)

ETUD pour les étudiants,

UNIV pour les universitaires,

INDUS pour les grands groupes industriels,

PME pour les PME-PMI,

PUB pour les scteurs public et para-public,

INDIV pour les auteurs qui ne font pas apparaitre l'appartenance à une communauté)

Liste d'autorité en cours d'élaboration

3) Champs contenant des propriétés liées à l'unité documentaire

PM : type d'unité documentaire

Résumé et éventuels mots-clés

Table des matières (index, sommaire)

introduction,

description du contexte (générale),

description du thème,

environnement (outils disponibles),

développement,

expérimentation (mesures, enquêtes...)

résultats

discussion,

conclusion

bibliographie

annexe)

PN : forme discursive de l'unité documentaire. Valeurs possibles actuellement :

descriptif

narratif

argumentatif

PO : langage (style) de l'unité documentaire : valeurs possibles

LITT : littéraire pur,

QUANT : parties contenant surtout des données numériques,

CALC : formules de calcul ou equations,

SCHEMA : schéma ou figures,

REPR : formulation, représentation, symboles, programmes.

V) Développement

V-1 Profil utilisateur

Chaque utilisateur du système est caractérisé par un profil. Ce dernier est décrit par un ensemble de propriétés associées à l'utilisateur. Nous pouvons distinguer deux types de caractéristiques, référencées par Ps et Pr.

Ps, est l'ensemble de caractéristiques utilisées pour sélectionner le corpus Cs. Cs est le corpus "personnalisé" et restreint sur lequel la question sera appliquée. C'est le corpus qui répond le mieux aux caractéristiques de l'utilisateur. Ps sera principalement constitué d'informations stables, décrivant le niveau pédagogique de l'utilisateur, son domaine disciplinaire, le genre de publication qui l'intéresse etc ...

Pr est l'ensemble des caractéristiques utilisées pour générer le corpus résultant Cr. Cr est constitué des unités demandées de l'ensemble de documents pertinents Dr.

V-2 Processus de recherche dans Profil-doc

Le processus est constitué de trois étapes, nous supposons avoir un profil utilisateur Ps, et une question Q, qui ont été définis.

V-2-1 Première étape

Le système sélectionne un corpus personnalisé Cs à partir de l'ensemble initial des unités documentaires, cela se fera par des règles telles que : Si Ps alors P_{ij} (dans notre expérimentation ces règles sont directement écrites sous forme d'équation booléenne, établissant une relation entre Ps et P_{ij}). Nous appelons cette étape : **l'étape de présélection**.

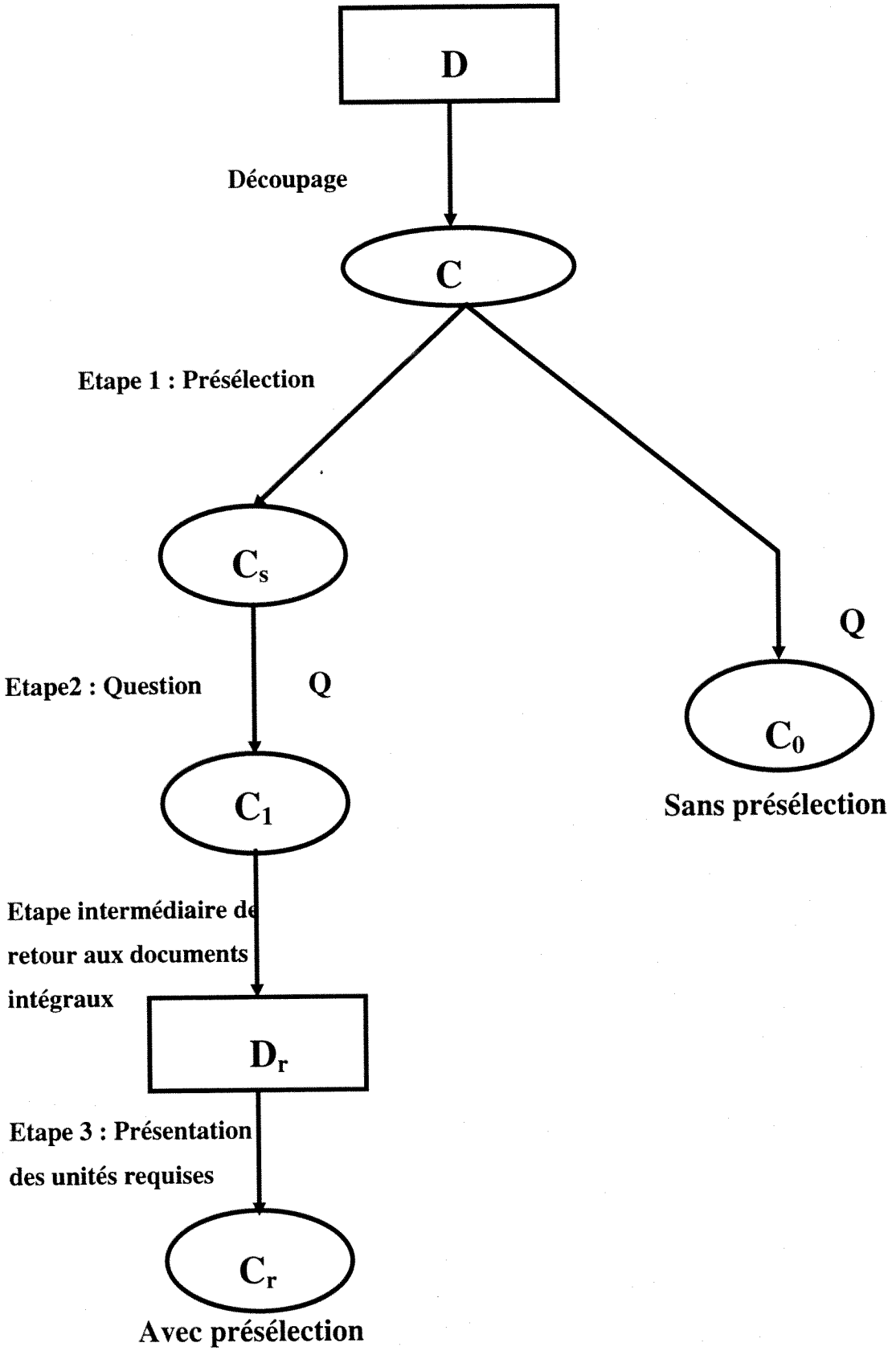
V-2-2 Seconde étape :

Le système applique la question Q de l'utilisateur sur le corpus Cs et alors génère le corpus C1. Le corpus C1 est inclus dans le corpus Cs.

V-2-3 Dernière étape

Le système construit le corpus résultant Cr par une sélection d'unités documentaires. Cela se fera par des règles telles que : si Pr alors P_{ij} (dans notre expérimentation, ces règles sont directement écrites sous forme d'équation booléenne, établissant une relation entre Pr et P_{ji}). Les unités documentaires extraites, résultent de l'ensemble de documents Dr.

V-2-4 Schéma récapitulatif



V-3 Organisation des relations entre unités documentaires

Les types d'organisation qui s'imposent dans le cadre de ce travail, sont au nombre de trois :

- Une structure logique propre à un type de document, (tous les documents devant être pris en considération),

- une séquentialité (précédent-suivant) à l'intérieur d'un document,permettant d'enchaîner des unités documentaires pour reconstituer l'ensemble du document dans l'ordre initial,

- Une navigation dans le corpus Cr construit dynamiquement, cette navigation s'effectuant d'une unité documentaire à une unité documentaire, à l'intérieur d'un même document, ou d'un document à un autre.

V-3-1 Structure logique

Cette structure logique, pour sa fonctionnalité, peut être décrite par le langage SGML. Cette description est la détermination des éléments logiques qui composent le document (chapitre, paragraphe, ...). Pour distinguer ces éléments, éventuellement les combiner, nous utiliserons un schéma contenant toutes les informations de cette structure qui est : la "Définition de Type de Document" ou "DTD". Cette structure logique, est conçue à travers les étapes suivantes :

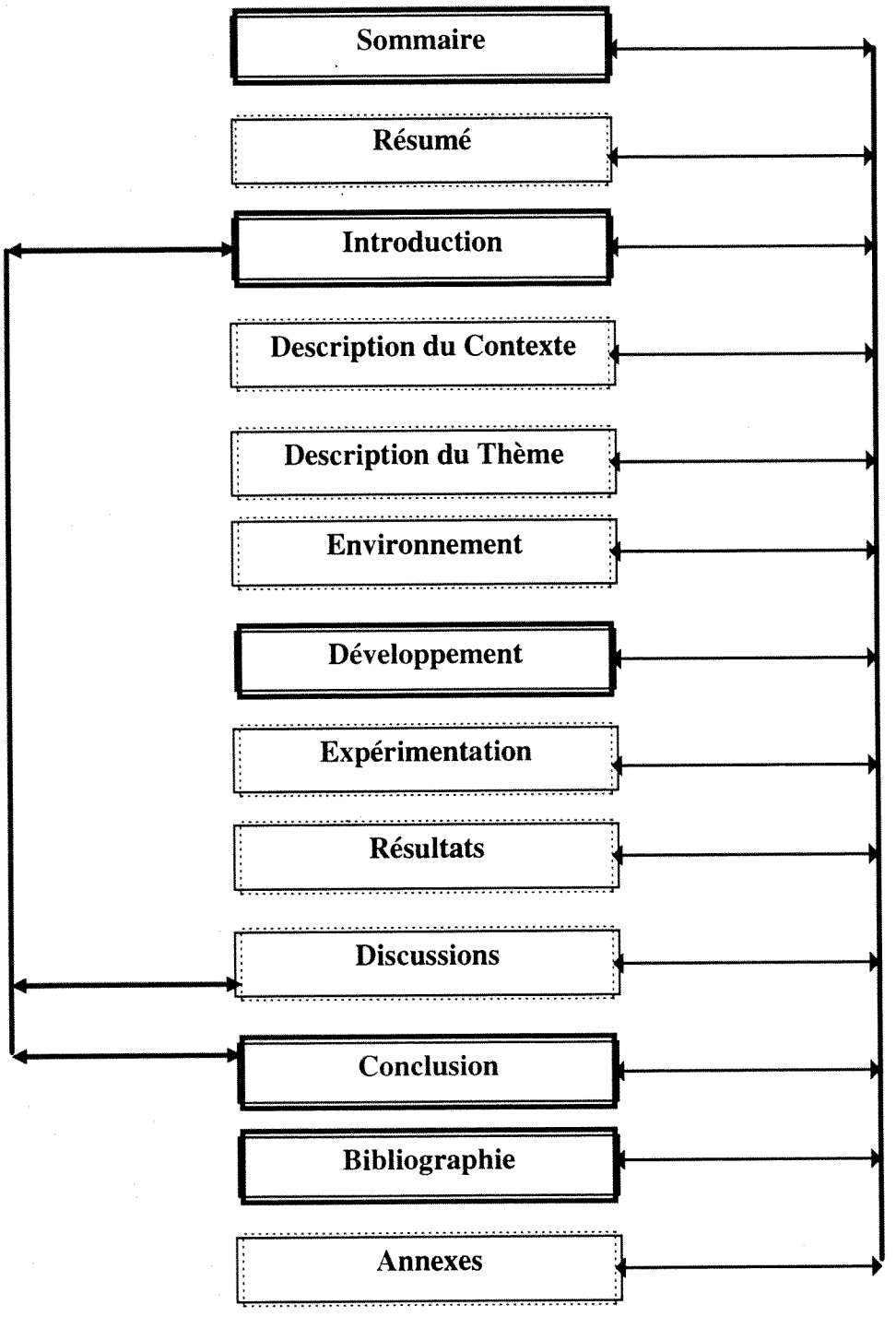
- Considérer un bloc cible que l'on construira à partir d'UD de même type. Chaque unité documentaire aura son numéro dans le bloc de même type. Ainsi, un bloc de type donné, sera délimité par un "UD début" et une "UD fin".

- Considérer un document cible de forme arborescente dont la racine est le bloc "sommaire" auquel viendront s'attacher les différents autres blocs.

Cette structure s'avère souple dans le sens où l'on considère des documents de différents types, et qui ne présentent donc pas la même structure physique. Cependant, nous allons introduire des liens qui existent entre différents blocs d'un document définissant des blocs

indispensables et des blocs facultatifs (Ceux-ci varient selon le type de document).

V-3-2 Liens hypertextes à l'intérieur d'un document



bloc indispensable

bloc facultatif

Nous allons à présent illustrer cette structure logique sous sa forme arborescente en considérant l'exemple suivant. Considérons le document doc qui se présente en réalité comme suit :

doc

sommaire

résumé

envir1

envir2

thème

dével1

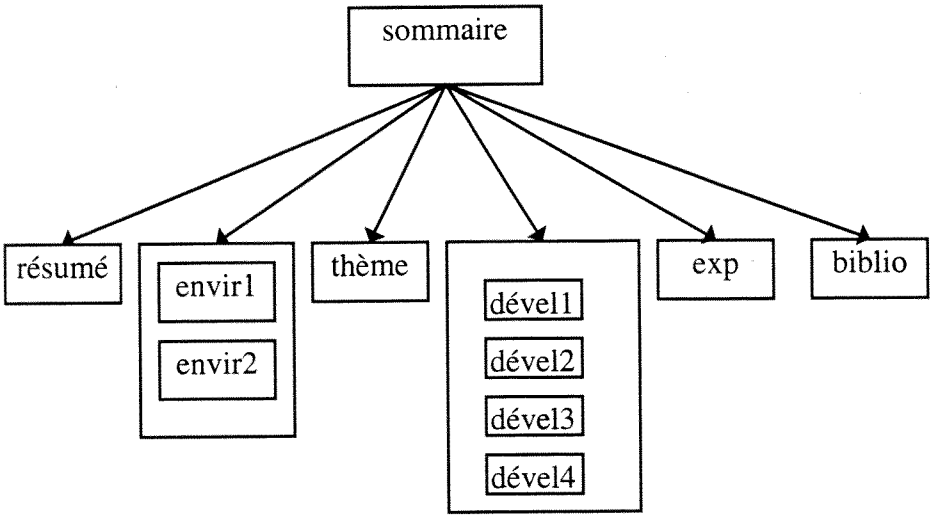
dével2

dével3

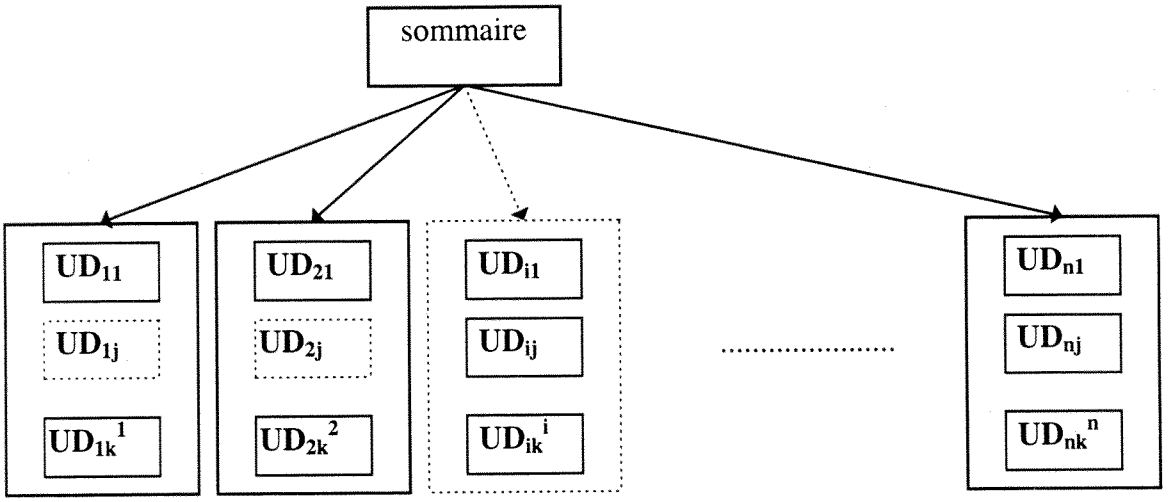
exp

devel4

biblio



Nous pouvons généraliser l'exemple précédent, nous obtenons la structure logique généralisée suivante:



avec : $n >= 1$ $k^n >= 1$

V-3-3 Séquentialité précédent-suivant

La tâche à réaliser consiste à passer d'une unité documntaire à autre séquentiellement. C'est un parcours séquentiel à l'intérieur du document tel qu'il se présente réellement. On utilisera le champ référence de l'unité documentaire dans le document, c'est à dire PB. On pourra incrémenter PB (ou le décrémenter) au fur et à mesure que l'on avance (ou l'on recule) dans le document.

V-4 Navigation

V-4-1 Les principes de navigation

Les principes de navigation sont au nombre de deux; Ils s'appuient sur les relations suivantes:

a) Relation lier à : Permet de passer à une unité documentaire de destination.

b) Relation lier avec : Permet de passer à une unité documentaire de destination à partir de laquelle on peut revenir à l'unité documentaire initiale.

V-4-2 Les mécanismes de navigation

Ils sont définis en fonction des types de chaînage :

a) concernant la structure logique,

* la relation lier à : Elle sera établie en fonction du champ type de l'UD par l'intermédiaire du sommaire, de la façon suivante :

- aller de l'unité documentaire initiale vers le sommaire,
- aller du sommaire vers l'unité documentaire de destination.

* La relation lier avec : Elle sera établie par l'intermédiaire du sommaire de la façon suivante :

- aller de l'unité documentaire initiale vers le sommaire,
- aller du sommaire vers l'unité documentaire de destination,
- revenir de l'unité documentaire de destination vers le sommaire,

aller du sommaire vers l'unité documentaire initiale.

Rem :

Il suffira pour aller d'un document d'un certain type vers une UD d'un autre type de pointer le bloc de type visé (en utilisant le mécanisme précédent) et atteindre l'UD voulue dans le bloc considéré par son numéro d'ordre dans le bloc.

b) concernant la séquentialité (précédent-suivant),

* la relation lier à : Elle sera établie par l'intermédiaire d'un bouton renvoi de l'UD initiale vers l'UD de déstination.

* la relation lier avec : Elle sera établie par l'intermédiaire d'un bouton "renvoi" de l'UD initiale vers l'UD de déstination à partir de laquelle on revient à l'UD initiale par un bouton "retour". Il suffira pour cela de considérer le champ référence PB d'une UD dans le document, la valeur de PB sera incrémentée (ou décrémentée) selon que l'on avance (ou l'on recule) dans le document.

V-4-3 Les types de navigation

Afin de pouvoir accéder directement aux unités documentaires, nous établissons un champ référence PB' qui est la concaténation du champ référence PA document et la référence de l'unité documentaire PB dans le document. Ainsi, nous aurons : **PB' = PA.PB**

V-4-3-1 Navigation dans Cr

a) Navigation entre les unités d'un même document :

On utilise la référence PA du document, la navigation s'effectuera entre les différentes unités documentaires en utilisant le sommaire. Cette navigation obéira aux principes et mécanismes définis antérieurement.

b) Navigation entre les unités documentaires de documents différents :

Il s'agira de documents différents dont les unités documentaire considérées ont la même valeur PM, (par exemple : naviguer à travers les conclusions, à travers les résumés, etc...). La clé PB' sera utilisée, et la relation dans ce cas, sera une relation lier avec .

c) Navigation entre les unités documentaires de Cr et les unités documentaires de C1 :

C'est le cas où l'utilisateur, ayant sélectionné à partir de C1 le corpus Cr, (en ayant déterminé différentes valeurs de PM des unités documentaires dans C1), souhaiterait, visualiser une (ou plusieurs) unités doumentaires n'appartenant pas au Cr. Pour cela , l'unité documentaire à visualiser doit appartenir au document contenant l'unité documentaire courante dans le Cr. Il suffira de remonter à l'ensemble des documents Dr grâce à PA, puis pinter l'unité documentaire voulue.

V-5 Gestion des Grilles

V-5-1 Création d'une grille sous Spirit

Spirit permet de créer et de modifier des grilles, en les personnalisant en fonction des besoins de l'utilisateur. Le format par défaut affiche la grille avec une zone par rang. L'utilisateur peut accepter ce format, ou créer des formats différents. Une nouvelle grille doit contenir au moins une zone, avec un nom et au moins un champ d'application., sachant qu'une zone peut correspondre à un champ, ou à un ensemble de champs.

Rem : Après création de la grille par l'utilisateur, ce dernier aura à l'enregistrer (éventuellement) en utilisant la boite de dialogue "Sauvegarde d'un fichier Grille" et l'extension du nom du fichier est ".grd". La validité de la grille est bien sûr vérifiée par Spirit, l'extension proposée est conseillée cependant pour toutes les grilles nouvellement crées.

V-5-1-1 Création, définition des valeurs prédéfinies pour les zones

Le système Spirit offre la possibilité de pré-définir des valeurs dans les zones d'une grille, afin de faciliter l'entrée. Ceci, consiste en la construction d'une liste de valeurs *disponibles par défaut* pour une zone. L'utilisateur, devra avant l'interrogation, sélectionner une valeur de la liste, ou bien en saisir une autre.
Il existe deux moyens pour pré-définir les valeurs pour chaque grille, à savoir :

- L'*initialisation par zone*, permettant de définir des valeurs associées à chaque champ

- L'*initialisation par dépendance,* permettant d'établir une relation entre les valeurs de différentes zones (maximum 3)

V-5-1-2 Initialisation par zone

L'initialisation par zone s'effectue lors de la construction ou la mise à jour d'une grille. L'utilisateur peut initialiser jusqu'à 1000 valeurs prédéfinies (dont la valeur de chacune ne dépasse 80 caractères) par zone. Cette initialisation s'effectue par le biais de la rubrique "Initialisatios", les valeurs sont saisies successivement à l'aide du bouton "Ajoute". Toute l'opération se déroule dans la boite de dialogue "Propriétés de la zone".

V-5-1-3 Initialisation par dépendance

Les valeurs inscrites dans les zones peuvent être définies comme dépendantes. Il y'a au plus, comme il a été mentionné plus haut, trois zones dont les valeurs sont dépendantes.

L'utilisateur peut définir jusqu'à 1000 relations de dépendance par grille, une boite de dialogue permettant la création de ces relations s'affiche et se présentant en deux parties.

Une partie supérieure définissant les relations entre les zones,

Une partie inférieure affichant les relations déjà spécifiées.

Rem :

- L'utilisation de cette fonctionnalité et la déclaration de valeurs sont exclusives,
- Le choix d'une valeur dépendante pour une zone fournit automatiquement les valeurs des autres zones,
- Cette technique définit un mode de recherche très sélectif.

V-5-1-4 Interdépendance des zones précisées

L'idée surgissant à ce niveau a déjà été mentionné ci-dessus en remarque. En effet, les valeurs précisées dans une relation sont *interdépendantes*, c'est à dire, dés que l'utilisateur sélectionne une valeur qui figure dans une relation en vigueur, les valeurs précisées pour les autres zones de la relation s'ajoutent automatiquement aux listes des valeurs disponibles. De même, lorsque l'utilisateur précise plusieurs relations pour la grille, elles s'appliquent simultanément.

V-5-1-5 Spécification d'une relation de dépendance entre trois zones d'une grille

Cette opération est réalisée sous l'éditeur de grille par le biais de la boite de dialogue "Zones Dépendantes". L'utilisateur aura ainsi à suivre les étapes suivantes :

1. L'utilisateur saisit le numéro correspondant à une zone, dans la deuxième colonne de la grille,

2. L'utilisateur répète l'étape précédente pour les autre zones,

3. En actionnant le bouton "Ajouter", la relation de dépendance apparaît dans la partie inférieure de la grille.

Rem : Pour supprimer une relation de dépendance, l'utilisateur aura à la sélectionner de la liste, puis à actionner le bouton "Supp".

V-5-2 Analogie entre Grille SPIRIT et le processus de recherche Profil-Doc

La grille, outil de SPIRIT est composée de sept (07) zones. Les valeurs des six premières zones (représentant des champs factuels) correspondent à des propriétés (éventuellement combinées par des "OU") que l'utilisateur définira. Le système SPIRIT permet de créer une correspondance entre ces zones et des champs de la base en instanciant ces derniers par les propriétés. La septième zone contient la question posée en Langage Naturel. SPIRIT effectue la recherche selon la question, et en concaténant les valeurs des différentes zones (Z1, Z2, Z3, Z4, Z5 et Z6). En procédant par un raisonnement analogique, le système effectue simultanément les étapes qui au niveau de Profil-Doc constituent les suivantes et utilisent deux fois la Grille.

Nous allons, avant de présenter ces deux étapes, simuler un exemple extrait de l'article de S.LAINE (cet article à paraître dans ******). L'énoncé de cet exemple est le suivant :

Exemple :

Soit un corpus initial constitué de :

- 30 mémoires d'étudiants en sciences de l'information, segmentés en 305 unités documentaires,
- 21 articles de périodiques professionnels en sciences de l'information, segmentés en 170 unités documentaires;
- 27 articles de presse grand public, segmentés en 145 unités documentaires.

Le corpus initial contient donc 620 unités documentaires.

Chacune des unités gérées par le logiciel Spirit se présente sous forme suivante : une suite de 17 champs factuels, suivis d'un champ de type textuel qui contient le texte de l'unité documentaire. Les champs factuels ne comportent pas de sous-champs.

Le profil de l'utilisateur contient les informations suivantes :

- données statiques : l'utilisateur est un professionnel des sciences de l'information (gestionnaire de centre de documentation dans une petite entreprise du secteur para-chimique). Son niveau d'étude est Bac + 4. Il a une formation en gestion et documentation
- données dynamiques (liées à la recherche en cours) : L'utilisateur souhaite recueillir des comptes-rendus d'expérimentation de mise en place de serveurs WWW, réalisées par des centres de documentation du secteur privé (méthodologie, aspects techniques, coûts). Il souhaite recueillir un nombre assez restreint d'exemples qui fourniront un point de départ à son étude, sans viser l'exhaustivité.

1) Spécification de Ps (pré-sélection) :

Restrictions relatives au document :

PI = PPROF ou M3

PK = SIC ou BIBL ou INFO ou ECO

PL = PME

Restrictions relatives aux unités documentaires sur lesquelles portera la question :

PM = description du thème OU expérimentation

PN = descriptif OU narratif

Po = LITT ou QUANT ou REPR

Cette pré-sélection permet d'extraire 218 unités documentaires sur les 620 unités initiales. Ces unités constituent le corpus Cs.

2) Question : Création d'un serveur WWW. Installation sur Internet d'un système hypertexte. Création de pages HTML.

Résultat : 18 unités documentaires qui constituent le corpus C1. Ces unités appartiennent à 12 documents qui constituent le corpus Dr.

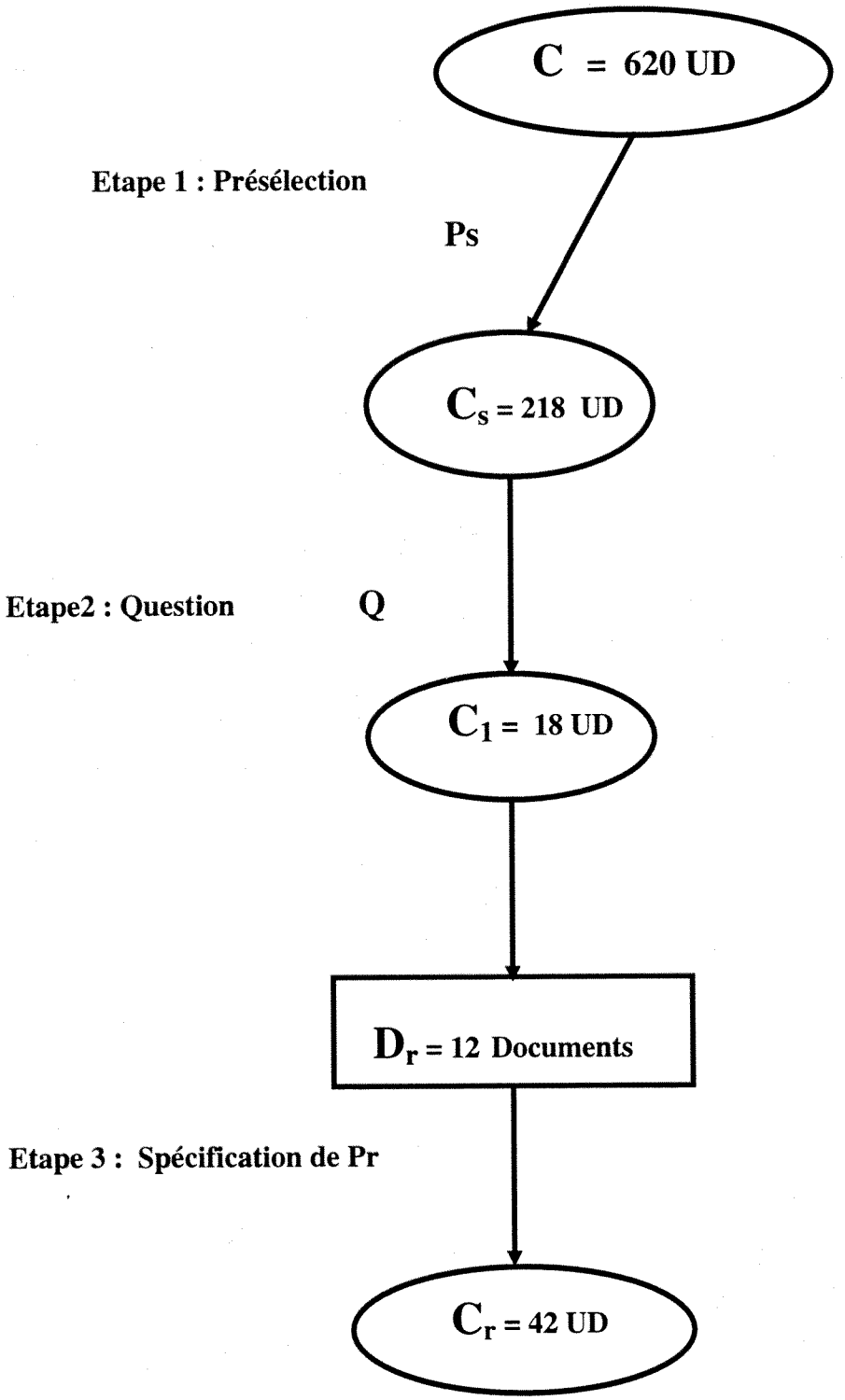
3) Spécification de Pr (éléments visualisables)

PM = environnement OU expérimentation OU résultats

PN = descriptif OU narratif

L'utilisateur obtient alors un ensemble de 42 unités documentaires qui constituent le corpus Cr.

Représentation des trois étapes du processus de recherche pour l'exemple



V-5-2-1 Première étape

L'étape de présélection qui détermine le PS avec la question Q posée sur Cs représentent l'utilisation de la Grille avec les propriétés de PS correspondant aux champs factuels et la question Q posée au niveau de la Grille.

Exemple :

PI = PPROF **OU** M3
PK= SIC **OU** BIBL **OU** INFO **OU** ECO
PL = PME
PM = description du texte **OU** expérimentation
PN = descriptif **OU** narratif
PO = LITT **OU** QUANT **OU** REPR
Q : Création d'un serveur WWW. Installation sur Internet d'un système hypertexte. Création de pages HTLM.

V-5-2-2 Seconde étape

La Grille est de nouveau utilisée pour l'extraction des champs références PA de documents, la récupération des PA se faisant actuellement manuellement, cette seconde étape regroupe au fait dans les champs factuels d'une part la combinaison des PA et d'autres part, les propriétés correspondant au PR.

Exemple : La première ligne contient les références des 12 documents correspondant aux 18 UD. Ils sont numérotés dans l'exemple ci-dessous de 1 à 12 mais cela ne veut pas dire qu'ils sont consécutifs, c'est à dire qu'il ne s'agit pas des 12 premiers documents (même s'il y'a une probabilité que ce cas se produise).

PA = D1 **OU** D2 **OU** D3 **OU** .....**OU** D12
PM = environnement **OU** expérimentation **OU** résultats
PN = descriptif **OU** narratif

V-5-3 Solution proposée (ou adaptée)

L'utilisateur interrogeant la base, exprime les informations de son profil. Le système doit lui fournir une réponse "satisfaisante", "affinée" et d'une façon transparente. Cela veut dire que l'utilisateur ne se rendra pas compte des actions produites lors des étapes intermédiaires entre la formulation de sa question et la réponse finale constituant ainsi tout le processus de recherche suivant, s'effectuant en deux étapes et ne devant pas provoquer un temps d'attente gênant pour l'utilisateur.

V-5-3-1 Première étape

L'utilisateur utilise la Grille SPIRIT en remplissant les zones par les propriétés correspondant au PS, et en posant sa question Q en Langage Naturel. La recherche est lancée normalement, c'est à dire qu'elle est effectuée par le système SPIRIT .

V-5-3-2 Seconde étape

Elle sera transparente pour l'utilisateur et rapide. Il s'agira de récupérer tous les champs références PA de documents pertinents dans un vecteur Vect, (on peut prévoir une allocation mémoire dynamique pour optimiser). Le vecteur Vect subira un tri et réalisera les accès aux documents dans la BDD, en pointant par PA les documents qui répondront aux unités documentaires déterminées par le PR.

VI) CONCLUSION

Il est clair qu'aprés avoir présenté les différents chainages concernant les unités documentaires ainsi que les mécanismes relatifs, nous pensons ainsi permettre à l'utilisateur une navigation facile et souple. Cependant, ce qui a été présenté, n'englobe que les aspects techniques du système lors d'une requête utilisateur et la chaîne de traitement en amont de l'action citée. Cette partie étant importante pour diminuer le bruit afin d'obtenir les documents pertinents pour un utilisateur déterminé, il est à remarquer que la prise en compte de la phase se situant en aval du résultat obtenu s'avère importante. Pour cela, nous dégageons un ensemble de propriétés qui sont directement impliquées dans la description des traits caractéristiques de tout utilisareur, à savoir :

- la signalétique,
- les droits,
- les préférences,
- les statistiques,
- l'environnement.

Dans le cadre de notre travail, le dernier point constitue la propriété la plus importante. En effet, la dimension environnement qui rassemble toutes les données propres à l'utilisateur (requêtes, résultats nouveaux, résultats archivés, table tâche à faire etc ...) est à réaliser.

Il faudra définir les concepts qui permettront de restructurer dynamiquement l'espace utilisateur (construit dynamiquement auparavant). Pour cela, l'interface à développer en aval doit être conviviale et permettre à l'utilisteur de naviguer facilement, de ne pas subir la "Surcharge Intellectuelle" (voir Glossaire).

Nous pouvons imaginer plusieurs actions dont la fonctionnalité s'appuiera à priori sur des "boutons" en développant des scripts (la continuation de ce présent travail dans un futur proche définira plus clairement cela) par l'intermédiaire d'un logiciel tel que Tool-book, ou alors réaliser cette tâche directement sous Spirit si cela est possible, ou encore sous un langage de programmation riche en Macro et en routines tels que par exemple turbo c, c++ etc ...

Les actions regrouperont en effet diverses tâches telles que l'historique des sessions récentes, la visualisation des résultats nouveaux, la possibilité de récupérer une unité documentaire déterminée, proposer quelques directives qui faciliteront la navigation, etc ...

Un autre point trés important qui n'a pas été développé dans ce présent travail est le lien dynamique constituant un renvoi spontané vers une unité documentaire, nous citons par exemple un renvoi vers une annexe ou un renvoi vers une bibliographie, etc ... Ce cas rentre pleinement dans la structure dynamique qui s'avère délicate à travers les liens dynamiques dans le domaine du traitement du texte intégral, mais nous pensons pouvoir résoudre ce problème dans la suite de ce travail en s'inspirant des évolutions des techniques dynamiques hypertexte. Nous espérons avoir apporté dans ce qui a été sugéré et conçu une solution positive qui enrichira le travail de toute l'équipe du projet Profil-doc et que les perspectives citées développeront les solutions présentées, et contribueront à l'évolution du projet dans sa totalité.

VII) Bibliographie

BEN ABDALLAH, Nabil.
Amélioration de la recherche documentaire par découpage de documents et utilisation du profil de l'usager
document interne à l' E.N.S.S.I.B. Octobre 94.

BIENNIER, Frédérique.
Modélisation d'une base d'hyperdocuments et méthode connexioniste d'aide à la navigation.
Thèse de doctorat. Edition interne INSA de Lyon. 1990.

BORNES, Christian.
Interfaces Intelligentes Dans l'Information Scientifique et Technique.
Cours INRIA,1992. ISBN : 2-7261-0726-5. 191 p.

BOUCHE, Richard., LAINE-CRUZEL, Sylvie., METZGER, Jean Paul.
Extraction de connaissances à partir d'une collection de documents.
In Tools of Knowledge Organization and Human Interface,
Congrés organisé par l' ISKO (International Society for Knowledge Organisation). 14-17 Aôut 1990. Darmstadt, England.

CARTHY, G.
Intelligent Front-ends for the idéal environmental database.
in online information, 1993. ISBN: 0-904933-85-7. 7-9 December 1993. London, ENGLAND.

CHABBAT, Bertrand., PINON, Jean-Marie., OU-HALIMA, Mohamed.
SHAD : Système Hypertexte pour l'Aide à la Décision.
Congrès Systèmes d'Information, Systèmes à base de connaissances. 17-20 mai 1994, Aix-en -Provence. INFORSID. pp 209-227.

CLAVEL,G., BIONDI, J
Introduction à la Programmation Tome 2, Structure de Données.
1984. Masson. ISBN : 2-225-80324-2

CONDUCTIER, Bruno.
Recherche conceptuelle d'informations dans des banques de données en ligne : application au projet INDUSCOPE.
Thèse. Université d'Aix-Marseille III. 1994 - 184 p.

DALBIN, Sylvie.
Interfaces dans les systèmes d'aujourd'hui.
in cours INRIA, Interfaces Intelligentes Dans l'Information Scientifique et Technique, 1992.
ISBN.2-7261-0726-5.

FLUHR, C
Algorithmes à apprentissage et traitement automatique des langues.
Thèse. Université Paris Sud. Centre d' Orsay. 1977. 274 p

FOLTZ, W., DUMAIS, T.
Personalized information delivery : an analysis of information filtering methods.
in communication of the ACM, December 1992, Vol 35, N° 12. pp 51-60.

GIROLLET, D., VICTORRI, B.
The Linguistic Analyser of Smart Gateway.
Conférence 16-18 Septembre, 1992. Grenade, Espagne.
Publié par SEPLN, Grenade, Espagne.

Guide d'utilisation de TOOLBOOK
in manuel de construction et d'utilisation. 1989-1991 par Asymtrix Coporation

HOCINE, A.
Conception d'une base de connaissances explicatives : utilistion d'un système hypertexte intelligent.
in Expersys-91, IITT-International. 1991. pp 355-365.

Intoduction à Spirit
in manuel utilisateur. T-GID Société. 1994

LAINE-CRUZEL, Sylvie.
Vers de nouveaux systèmes d'information prenant compte le profil des utilisateurs.
in Documentaliste. Sciences de l'information, 1994, vol31, N° 3. Pp143-147.

LAINE-CRUZEL, Sylvie., LAFOUGE, Thierry., LARDY, J.P., BEN ABDALLAH, Nabil.
Improving information retreval by combining yser profile and document segmentation.
A paraitre dans Information Processing and Management. 1996.

LE CROSNIER Hervé
Systèmes d'accès à des ressources documentaires : vers des anté-serveurs intelligents.
Thèse. Université Aix-Marseille III. 1990. 355 p.

Logiciel de recherche documentaire en langage naturel SPIRIT
in Plaquette d'information T-GID Société. 1993.

PAOLI, C.
Les passerelles de communication.
dans cours INRIA. 1992. ISBN : 2-7261-0726-5. 191 p.

PINON, Jean-.Marie., LAURINI, R.
La documentation multimédia dans les organisations.
Collection Technologies de Pointe.1990. HERMES, Paris. ISBN : 2-86601-217-8

PINON, J.ean-Marie., BIENNIER Frédérique.
Base de documents hypermédias.
Actes du Congrés Bases de Données BD'91. Alger, Algérie. 1991. pp 162-215.

RADASOA, H
Méthodes d'amélioration de la pertinence des réponses dans un système de bases de données texruelles.
Thèse. Université Paris Sud. Centre d'Orsay. 1977. 156 p.

RHEINGOLD, Howard.
Les communautés virtuelles.
1995. Addisson-Wesly, France. ISBN: 2879080894. 311 p.

ROESTER M., HAWKINS D.T.
Intelligent Agent.
in Online, Juillet 1994, Vol 18, N° 4. pp 19-25.

ROLE, François
La norme SGML, pour décrire la structure logique des documents

Spirit Consultation
in manuel utilisateur. T-GID Société. 1994

TOMASIC A., GARCIA-MOLINA H.
Query Processing and Inverted Indices in Shared-Nothing Document Information Retrieval Systems.
in VLDB Journal, July 1993, vol 2, N° 3. pp 243-275.

VICTORRI B., THOMAZO L., BOYREAU G., MADELAINE J., COULON J.F., HANRIOT D., LE CROSNIER H., GIROLLET D.
The front-end server : an intelligent interface to documentation.
in INFORMATIQUE 92. International Conférence Interface to Real and Virtual Worlds, 23-27 march 1992, Montpellier, France.

VIII)Glossaire

Bouclage de pertinence

Technique du monde de la RI [DACH90], ne pouvant néanmoins comme le remarque [BELK87a] cité par [COND94], être qualifiée de technique de RI puisque visant à affiner le modèle de requête, en prenant en compte les jugements de pertinence émis par l'utilisteur au sujet des réponses précédentes pour reformuler automatiquement la requête suivante.

Son efficacité n'étant plus à prouver, elle a été adoptée par les SIRI [CROF87a] cité par [COND94], et semble sur le point de gagner les systèmes hypertextes [LECR91] cité par [COND94], où elle permettrait d'élaguer quelque peu les arborescences de lien. De plus, sa simplicité et son adhésion à la philosophie des systèmes coopératifs permettent de créer des interfaces intuitives ([THIN91], [HALI89], [PRIT93] cités par [COND94]).

Relevance feedback

Butineur, Feuilleteur

Désigne un outil qui permet de naviguer "librement" au sein de données en suivant par association d'idées des liens hypertextes préexistants ([CONK87] cité par [COND94]).

BROWSER

Diffusion sélective de l'information, DSI

Désigne le filtrage d'un magma d'information au travers d'un tamis, le profil utilisteur, qui reflète les centres d'intérêt de ce dernier, afin de ne conserver que les données pertinentes et de les présenter de manière hiérarchisée. Quoique passifs au regard des SRI, puisqu'alimentés autoritairement par des flots d'information en provenance d'agence de presse ([JACO90] cité par [COND94]), ou de messageries électroniques ([FISC91a], cité par [COND94], ces systèmes ([MCCL94] cité par [COND94]) partagent une grande partie de la problématique de la RI.

Selectiveinformationdelivery,

SID

Hypertexte

Les systèmes de cette famille ([SAVO90], [LECCR91], [CONK87], [ABOU93] cités par [COND94]), reposent sur l'idée d'associer, grâce à divers liens, des informations stockées dans une base de données à des représentations graphiques, l'utilisateur pouvant se déplacer le long de ce maillage à l'aide de butineurs.

hypertext

Indexation

Selon ([HALI89] p.13, cité par [COND94]), "*C'est une étape fondamentale qui donne au document un statut conceptuel dans la base gérée par le* SRI". D'un point de vue «intuitif, les index associés à un

document sont censés représenter au mieux son contenu à l'aide d'un formalisme standardisé, afin d'en faciliter l'accès par la suite. Toutefois, les termes «au mieux» et «standardisé» masquent une complexité telle, que les méthodes à priori plus frustres d'indexation automatiques bénéficient dorénavant d'un bon rapport qualité/prix par rapport à l'indexation manuelle.

Indexing

Indexation sémantique

Ce terme hérite de ([NEBE90b] cité par [COND94]), qu'il convient de ne pas confondre avec le mécanisme d'indexation utilisé en RI pour décrire le contenu d'un document, est une transposition/adaptation dans le monde des langages teminologiques d'une technique d'optimisation classique du monde BD : à savoir la pose d'index autorisant un accés quasi immédiat aux individus d'un concept donné, qui couplée au mécanisme de classification de requête permet d'accélérer de façon drastique l'exécution de celle-ci, en minimisant le nombre d'individus effectivement confrontés à la requête.

Semantic indexing

Interface conviviale

A notre sens, c'est une interface dont la convivialité provient plus de l'enrobage graphique, que d'une démarche ergonomique profonde. Ainsi,de nombreuses interfaces commerciales de RI en ligne QUESTEL-ACCESS ([QUES91] cité par[COND94]), facilitent la tâche de l'utilisateur grâce à des menus, des scrollers et des éditeurs pour visualiser les résultats ou saisir les requêtes, mais la philosophie du mode de dialogue sous-jacent n'est guère différente de celle héritée des systèmes documentaires fondés sur un langage de commande.

Friendly Interface

Interface en langage naturel

En adoptant une syntaxe de formulation du besoin, proche de la langue parlée ou écrite, les interfaces se réclamant de cette approche ([PRIT93] cité par[COND94]), visent à recentrer le mode d'interaction sur l'utilisateur, par opposition aux autres techniques dont les fondations mathématico-logiques absconses, plus ou moins affleurantes priviligient dûment la machine au détriment de l'être humain.

Naturallangage interface

Modélisation de l'utilisateur

Se réfère à l'introduction au sein d'un système, de connaissances relatives à l'utilisateur, à des fins d'adaptation du comportement interne et externe de l'application ([BRAJ87] p.306 cité par [COND94]). Ce problème est particulièrement sensible dans le cas des systèmes coopératifs ([MAST90] cité par [COND94]), où l'utilisation de critiques pour être constructive, passe par une qualité du dialogue homme/machine directement subordonnée au niveau du premier.

Usermodelling

Problème du musée d'art

Cette métaphore empruntée à Carolyn Foss et citée par [LECR91] p.288, illustre l'un des problèmes rencontrés par l'utilisateur qui chemine au sein d'un lustre l'un des problèmes rencontrés par l'utilisteur qui chemine au sein d'un hypertexte : à force de flâner sans but précis, le visiteur est incapable d'assimiler l'information, c'est à dire de conceptualiser à partir des items entr'aperçus.

Art museum problem

Passerelle decommunication

Ces systèmes [PAOL92], peuvent être vus comme des antéserveurs centralisés qui permettent d'accéder via un guichet unique à moult serveurs et bases. A la différence des antéserveursproprement dits, le stade du produit commercial est atteint depuis longtemps EASYNET ([SCHO93], DGIS [KUHN88] cités par [COND94]), INFORMATION-INTELLIGENTE [PAOL92], et la taille critique amplement dépassée, puisque le premier cité fédère 13 serveurs et 1200 bases.

L'interface dans le cas de ces produits, est plus proche de l'outil évolué (LCC) que de l'assistant intelligent.

Gateway

Surcharge intellectuelle

Se réfère aux difficultés rencontrées par l'utilisateur lorsqu'il doit mener de front plusieurs tâches mentales. Un facteur particulièrement aggravant est lorsque les différentes tâches concurrentes n'ont pas de rapport direct avec la problématique propre de l'utilisateur .

L'apparition de ce problème est un bon indicateur pour l'emploi d'une interface intelligente ([SHAR91] cité par [COND94])

Cognitive overhead

Système coopératif

Ces systèmes ([MAST90] cité par [COND94]), visent à réintégrer l'utilisateur au sein du cycle de résolution de problème d'où il avait été exclu indûment par l'approche système expert. ce faisant, l'utilisateur étant redevenu un agent actif au même titre que le système, chacun peut collaborer à la résolution du problème, en se concentrant sur les aspects relevant de ses aptitudes, qui pour l'être humain concernant l'analyse des informations planifiées, ainsi que de la collecte des indices nécessaires aux prises de décision.

Cooperative system

Système de recherche d'information conceptuel

Sous ce vocable, on peut regrouper tous les systèmes qui tentent de dépasser le pallier atteint par les méthodes classiques de RI, en intégrant un zeste d'IA, sans pour autant basculer dans le monde du TLN. Pour ce faire, on utilise généralement des taxinomies, pour

indexer en *compréhension* les informations et faciliter leur recherche, Cf.TAIGA [MADI92] cité par [COND94]. Toutefois, la phase de construction s'est révélée rédhibitoire dans les premiers prototypes COREL [CROS86], GRANT [COHE87] cités par [COND94, du fait de son caractère manuel.

Pour remédier à ce problème, RUBRIC-TOPIC [TONG88], I3R [CROF86], proposent de rechercher des évidences textuelles exprimées à l'aide d'un formalisme traditionnel, par exemple le langage booléen. Les conceptes étant repérés de manière ad hoc, puis les indicesrelevés étant combinés à l'aide d'un mécanisme de raisonnement plausible [TONG89] cité par [COND94].

Conceptuel nformation retrieval system

Système de recherche d'information en ligne

Cette branche des SRI se distingue par une problématique héritée de celle plus générale de la RI, mais enrichie des écueils constitués par l'utilistion de banque de données en ligne [HAWK88] cité par [COND94]. De fait, alors que dans le monde de la RI documentaire,les textes sont disponibles en local et donc susceptibles d'être remaniés à loisir sans surcoût, en ligne on ne peut en revanche qu'accepter le mode de représentation de l'information choisi par le fournisseur du service, ainsi que le mode de récupération proposé, en général le langage booléen. De plus, le mode de facturation actuel des serveurs [LECR90] p.326, conduit rapidement à des coûts de fonctionnement élevés. En conséquence, les SRI en ligne ne peuvent se permettre le luxe de finasser, mais doivent plutôt circonscrire rapidement un espace de travail compatible avec un télédéchargement vue d'un post-traitement éventuel par des outils plus élaborés.

Online information retrieval system

Traitement du langage naturel, TLN

Un malentendu persiste quant au positionnement de l'IA (Intelleigence Artificielle) vis à vis du LN (LangageNaturel). S'il est évident que ces deux domaines possèdent une intersection commune, il est par contre irrecevable que l'IA subsume le TLN. En conséquence, effectuer des raitements linguistiques ([DEBI88] cité par [COND94]), n'implique nullement une approche conceptuelle et encore moins une ébauche de compréhension.

Naturel langageprocessing, NPL

IX) Annexes

Cette annexe fait l'objet d'une évocation de quelques systèmes extraits en majorité de la thèse de B.CONDUCTIER et de mon travail dans le cadre de ma note de synthèse.

• CARTE-EXPERT

Ce système développé par la société de même nom, filiale de QUESTEL, et subventionné dans le cadre du projet européen IMPACT est un excellent exemple de ce que pourrait être un antéserveur; mais illustre également l'apparente antinomie à faire développer ce type de produit par un fournisseur d'information de premier plan (à confronter avec EASYNET développé par TELEBASE qui a une position beaucoup plus saine dans ce domaine).

De fait, si [PAOL92] page 94, avance des caractéristiques prévisionnelles plus que flatteuses (accès à 15 serveurs / 40 bases, point d'accès unique, LCC, interface conviviale, présélection des bases), [CHAT92] page 32, revoit à la baisse la nombre de sources (10 bases, 3 ou 4 serveurs) et, fait plus important, note que le rôle temporaire (automatisation début 1993) de l'antéserveur est de centraliser les demandes des utilisateurs qui seront par la suite traitées manuellement.

Ce pallier manuel n'est pas sur le point d'être franchi [1] [MARI93], et si l'on se fie au message véhiculé par ce document, il est fort à craindre que l'intelligence de l'interface soit définitivement remisée au profit de celle des courtiers d'un service de RI à haute valeur ajoutée [2]. Mais à ce propos, quel intérêt pour l'utilisateur à s'escrimer sur une interface informatique alors qu'il peut obtenir des services tout du moins similaires par simple appel téléphonique (services SVP) ?!

• CODER (COmposite Document Expert Retrieval)

Ce grand prototype de recherche (de la trempe d'I3R) s'est assigne comme objectif [FOX87] de tester les techniques à connotation IA en les confrontant à la problématique de la RI. Un intérêt tout particulier a été porté à l'architecture à tableau noir [WEAV88].De même, la représentation de connaissances à l'aide d'un langage de schéma est un élément clef de ce système [WEAV89]. En effet, le nombre d'informations à prendre en compte (dictionnaires, utilisateurs, documents, ..) nécessite de disposer d'un véritable SBC. Les concepteurs se réclament d'ailleurs de l'école des langages terminologiques. Bien qu'utilisant le TLN, ce système ne néglige pas pour autant les acquis des techniques classiques et tâte par exemple de la recherche probabiliste [BELK87a]. Des tests ont été effectués sur des messages électroniques (Cf. INFORMATION-LENS, INFOSCOPE). Suivant une étude prospective [PORT88b] page13, #*This is the sort of system that could implement information retrieval functions experimentally in 1995* # .

• DARWIN

Ce logiciel d'indexation plein texte (analogue à SPIRIT) commercialisé par CORA, est utilisé par la Banque de France pour explorer son règlement interne [BLAS93], ainsi que par l'EDF-DER pour explorer des documentations de fiches d'application [LEY93].

Dans les deux cas, il convient de noter l'utilisation de l'outil pour créer automatiquement un réseaux hypertexte à partir d'une terminologie. La maquette développée par l'EDF se singularisent par l'utilisation de la méthodologie de construction de BdC KADS pour gérer la terminologie d'indexation. Ainsi à la différence d'un système propriétaire comme SOFER, bénéficie-t-on de la réutilisation de composants logiciels industriels. [TRIQ94] annonce le développement du système multilingue CAROLUS dans le cadre d'EUREKA, une extension de DARWIN permettant la traduction de concepts (Cf. LEXIC et MITI).

- **DIALECT**

Ce prototype de recherche [MEKA91], est qualifié par ses concepteurs comme étant un système multi-experts pour la recherche documentaire. Et de fait ce système dispose-t-il autour d'une architecture à tableau noir (Cf. I3R, CODER), de différents experts, dont les plus complexes sont ceux consacrés aux connaissances linguistiques. Celles-ci occupent une place privilégiée, et permettent réellement de parler de TLN (Cf. SPIRIT, LEXIC), contrairement aux systèmes à formulation libre et bouclage de pertinence (Cf. WAIS, FULL-TEXT).

Outre l'immédiate application du LN à l'indexation et à l'expression du besoin, c'est au niveau de la reformulation de requête que l'apport est le plus manifeste [DEBI 88],

- **EASYNET**

Cette passerelle de communication développée par TELEBASE en 1982, permet l'accès à 13 serveurs et 1200 banques de données, à travers une interface unique fondée sur un système de menus, ainsi qu'un LCC [WILC88]. Elle peut être considérée comme la première grande passerelle en activité (Cf. DGIS et INFORMATION-INTELLIGENTE), et son succès, quoique limité aux USA, ne s'est pas démenti depuis sa création [SCHO93]. Ce dernier, page 55, attribue une partie de ce succès à la position neutre de TELEBASE, qui n'est pas une société sous contrôle d'un quelconque fournisseur d'information en ligne (à la différence par exemple de CARTE-EXPERT).

Si les fonctionnalités proposées permettent effectivement d'effectuer des recherches multi-bases et multi-serveurs, le LCC destiné aux utilisateurs experts n'est en revanche opérationnel que sur 5 serveurs, et n'a donc de commun que le nom.

- **EPRINET**

Cette passerelle de communication développée par l'EPRI [EPRI91], témoigne d'une volonté rare dans le monde de la recherche documentaire de mettre le système réellement dans les mains des utilisateurs (Cf. Les LCC de DGIS et EASYNET tournés principalement vers les documentalistes). Pour ce faire, elle a opté pour une interface conviviale, utilisable presque sans formation, à base de menus et de formulation libre. Un point important à signaler est la forte intégration de fonctionnalités que l'on retrouve actuellement dans les collecticiels: messageries électroniques, forums, infocentre partagé. Toutes fonctions vitales lors de la coordination de projets de plus en plus éclatés, et qui parviennent à motiver un public néophyte jusque là ignoré par les logiciels de recherche documentaire (Cf. HOOVER), dégageant par là même le fameux marché manquant au lancement des antéserveurs [HORW88]

- **EURISKO**

Ce prototype de RI en ligne est un bon exemple de l'approche système-expert appliquée à la recherche documentaire (Cf. IOTA). Le principal effort de ce projet [BART87], [BART88], a été concentré sur la planification de la stratégie d'interrogation. En revanche, les autres fonctions d'un antéserveur n'ont pas reçu d'attention spécifique. Des tests ont été effectués sur des bases de TELESYSTEMES et du CEDOCAR.

- **FREEWAY**

Ce logiciel d'interrogation déporté sur un poste de travail de type ordinateur personnel [CASA93], ne peut cependant être gratifié de l'appellation d'antéserveur (tout comme CYPRESS, DIALOG-LINK, ou QUESTEL-ACCESS), du fait d'un accès limité au seul serveur FT-PROFILE, ainsi que d'une convivialité entièrement héritée d'une interface graphique, plutôt que de fonctionnalités d'assistance réellement *#intelligentes#*. En effet, la réussite de ce produit provient d'une utilisation raisonnée de menus, de formulaires de saisie, ainsi que d'une formalisation d'un câblage de la connaissance sur les bases via des menus hiérarchiques, permettant à un utilisateur occasionnel de s'orienter vers la bonne source d'information, tout en s'abstrayant au maximum d'une quelconque syntaxe.

A ce sujet, il convient de signaler que le langage d'interrogation booléen est tout autant abscons sous une forme graphique que textuelle. En effet, l'éditeur de requête visuel proposé, s'il permet de donner un semblant de modernité, ne solutionne aucunement les véritables problèmes, à savoir la structuration de la requête et le choix du vocabulaire.

- **FULL-TEXT**

Ce produit commercialisé par FULCRUM (Cf. SFQL pour sa participation) permet dans sa sixème version [GUEZ93], d'effectuer des recherches d'information en plein texte «*suivant une nouvelle méthode de recherche intuitive et interactive*». En fait de nouvelle méthode, on est en présence de la technique bien éprouvée de bouclage de pertinence, qui semble retrouver une nouvelle jeunesse, si l'on considère son adoption par WIN ou WAIS. « Comme pour ce dernier, la philosophie d'interface obtenue est facilement assimilable et doit permettre à un utilisateur non spécialiste de se «*débrouiller*» seul.

- **GOPHER**

Selon [LANG93] page 14, «*ce protocole est le plus simple qui permette d'opérer des opérations de recherche documentaire en mode réparti et d'obtenir un accès aux documents sélectionnés: GOPHER est principalement un système général de diffusion de documents*». Devant la multiplication des sources, la méta-base VERONICA décrivant les différentes bases accessibles a été implémentée [NOTE93]. Cette approche, similaire à celle adoptée par WAIS, n'est pas le seul lien rapprochant ces deux systèmes. En effet, suivant une attitude moins frileuse que la traditionnelle politique protectionniste de replis sur soi, des ponts ont été établis avec ses principaux rivaux, dont WAIS qui peut être sollicité de manière transparente au travers du protocole GOPHER.

- **INFORMATION-INTELLIGENTE**

Cette passerelle de communication [PAOL92], conçue par INFOTAP et GEONET pour répondre à un appel d'offre de la CEE en 1988 (Cf. IMPACT) est accessible dans le cadre du service GEOMAIL, et autorise l'accès à 11 serveurs et 120 banques de données à travers l'utilisation d'un système de menu hiérarchiques.

La démarche adoptée est très proche de celle d'EASYNET, et en conséquence les fonctionnalités disponibles sont très éloignées des ambitions affichées par DGIS ou CARTE-EXPERT (bien que ces dernières soient également revues à la baisse).

• IOTA

Ce SRI plein texte de la famille des systèmes experts appliqués à la RI, possède un certain air de famille avec EURISKO, bien que ce dernier soit un système de recherche en ligne. En revanche, il se distingue par l'utilisation de la stratégie d'interrogation pour déterminer le modèle d'utilisateur à employer [CHIA87]. Le principe étant d'affecter l'utilisateur dans une typologie ou une autre, en se basant sur la dégradation de la requête primaire vis-à-vis de la requête finale, le système ayant procédé aux adaptations nécessaires. De plus, sa dernière version [BRUA89] dispose de fonctionnalités avancées de constitution automatisée de la BdC, en particulier de la terminologie d'indexation (Cf. Egalement SPIRIT et DARWIN).

• IR-NLI (Information Retrieval Natural Language Interface)

Ce SRI est un prototype d'interface en LN, construit au dessus d'un système expert, comparable è EURISKO ou IOTA, qui intègre entre autre les tactiques d'interrogation telles qu'elles sont énoncées dans les vade-mecum sur la RI en ligne. Un réseau sémantique sert à modéliser la connaissance détenue sur les banques de données ainsi que sur l'utilisateur. Une attention toute particulière a d'ailleurs été accordée à cette dernière fonctionnalité [BRAJ87], qui repose pour une grand part sur une hiérarchie de stéréotypes d'utilisateurs, auxquels l'utilisateur est dynamiquement rattaché en fonction de l'activation de certaines conditions provoquées par l'analyse du dialogue courant, ainsi que l'étude de l'historique des sessions antérieures.

• LEXIC

Ce système parrainé par le MRT et diffusé par la société IDEEM du groupe SYMBIOSE [IDEE93], permet d'effectuer une recherche plein texte (Cf. SPIRIT, FULL-TEXT, DARWIN,IOTA), à l'aide d'un réseau neuronal (Cf. L'activation de proximité déjà au menu de GRANT et D'I3R), dans un corpus de texte préalablement indexé par un mécanisme (secret ...) capable grâce à des traitements linguistiques de faire ressortir la signification des mots. L'utilisation d'inférences approximatives autorise des réponses nuancées. Ce nouveau-né babille 4 langues (français, anglais, espagnol, italien) et a pour parents adoptifs outre le MRT, l'arsenal de Toulon, le centre de tir de Kourou, RHONE-POULENC, et d'EDF-SQR.

• RUBRIC-TOPIC

Ce système de recherche d'information conceptuel mis au point par ADVANCED-DECISION-SYSTEMS et connu sous le nom de prototype de RUBRIC [TONG85a], [TONG85b], [TONG87], [TONG88], [TONG89], est devenu le produit TOPIC de VERITY, distribué en FRANCE par SYSECA, jusqu'à la création de VERITY-FRANCE.

La philosophie de base est pragmatique et consiste à préférer au TLN trop coûteux une approche basée sur la recherche en plein texte, dans des documents de tout format, éventuellement distribués au sein d'un réseau, d'évidences textuelles permettant d'inférer plus ou moins fiablement la présence d'un sujet associé (Cf. La récupération de cette technique par I3R). Ces sujets sont eux-mêmes organisés en une hiérarchie mise à profit par un mécanisme de raisonnement plausible pour ordonner les réponses. Des modules autorisent un couplage avec des SGBDRs, ainsi qu'une utilisation en temps réel, le système analysant en continu des messages ou dépêches électroniques (Cf. SCISOR).

Quoique fortement implanté aux USA, (400 références gouvernementales et grands comptes, en particulier la CIA «*was a key customer for VERITY's TOPIC*»
[DYSO91] page 18), son assise en France est plus modeste, mais compte toutefois une référence de poids en la cellule de veille technologique de THOMSON.

De même, cette technologie est-elle intégrée au logiciel de travail en groupe NOTES 3.0 de LOTUS, pour filtrer les messages [KESE94], [GUEZ93] p. 19, (Cf. HOOVER avec lequel il a été couplé [KESE94], INFORMATION-LENS, INFOSCOPE, WIJIT, BEYOND-MAIL), et sur le point d'être au système CAROUSEL d'échange de document d'ADOBE. Cette société a d'ailleurs pris une participation dans le capital de VERITY.

- **SFQL (Structures Full-text Query Language)**

De par son patronyme, on peut se douter que ce protocole [DYSO91], pour les architectures client-serveur, de recherche d'informations textuelles stockées sur CD-ROM, fait la part belle à la structure du document et de fait est il plus rigide que CD-RDX ou WAIS. Son appui sur un schéma fédérateur pour stocker le document, et l'utilisation d'une variante de SQL pour interroger, le rapproche de la problématique des SGBDOOs utilisées pour gérer et retrouver des textes au format SGML ou ODA.
Initialement créé en 1987 par des compagnies aériennes (BRITISH AIRWAYS, BOEING, KNOWLEDGESET, MAXWELL DATA, TMS).

- **SPIRIT (Syntactic and Probalistic Indexing and Retrieval of Information in Texts)**

Ce logiciel de SYSTEX est couramment utilisé par l4EDF-GDF-SERVICES lors de la consultations de banques documentaires en LN [BEZA93]. Tout comme DARWIN, ce SRI plein texte utilise intensivement le TLN [FLUH92], lors de la phase d'indexation, ainsi que lors de la phase de la reformulation de la requête [DEBI88]. De même, utilise-t-il des traitements statistiques pour établir la proximité des textes de la requête. Comme RUBRIC-TOPIC, il autorise un accès à l'information à l'aide de graphes de concepts, toutefois à la différence de ce dernier, il dispose d'une automatisation de la phase de constitution de cette taxinomie [FLUH92] p. 122.

- **WWW (World-Wide-Web)**

Ce logiciel développé par la CERN, n'a pas usurpé son patronyme, puisqu'il tente littéralement de tisser une toile d'araignée reliant par des liens hypertextes des sources d'informations réparties sur le réseau INTERNET. Des passerelles vers ses confrères (GOPHER, WAIS) sont d'ores et déjà disponibles. Mais l'hypertexte ne pouvant à lui seul relever le defi, ([LANG93] page 18, #l'hypertexte, par la création de liens actifs entre différents éléments d'information, permet , dans certaines limites, de #reproduire# l'organisation des idées d'un utilisateur, alors que l'acquisition d'information permet d'établir des associations d'idée en fonction du contenu de l'information#), des requêtes portant sur un mot simple permettent de créer des liens dynamiques (Cf. FARE).