DEA SCIENCES DE L'INFORMATION ET DE LA COMMUNICATION

OPTION 3

Description de documents textuels Indices pour une typologie prenant en compte le contexte et la finalité de la communication

BEN ABDALLAH Nabil

Sous la direction de

Sylvie LAINE-CRUZEL

ÉCOLE NATIONALE SUPÉRIEURE DES SCIENCES DE L'INFORMATION ET DES BIBLIOTHEQUES

-1993-

DEA SCIENCES DE L'INFORMATION ET DE LA COMMUNICATION

OPTION 3

UNIVERSITÉ LYON 2 UNIVERSITÉ LYON 3 ENSSIB

Description de documents textuels Indices pour une typologie prenant en compte le contexte et la finalité de la communication

BEN ABDALLAH Nabil

Sous la direction de

Sylvie LAINE-CRUZEL

ÉCOLE NATIONALE SUPÉRIEURE DES SCIENCES DE L'INFORMATION ET DES BIBLIOTHEQUES

-1993-

Table des matières

PARTIE III

INTRODUCTION

Les systèmes documentaires informatisés tels que les catalogues en lignes, les banques de données bibliographiques ou / et textuelles, les banques de données financières et juridiques, etc. , ont nettement influencé le processus de recherche d'information et par conséquent le comportement de l'usager. Un éventail des recherches a été mené afin d'adapter les différents produits des systèmes documentaires aux besoins exprimés des utilisateurs. Mais, le problème de l'inadéquation habituellement constaté entre la demande exprimée par le chercheur et les ressources décrites par le système persiste encore. A nos jours presque tous les chercheurs s'entendent sur le fait que le problème de l'inadéquation constatée découle essentiellement du processus d'indexation et de la manière avec laquelle les chercheurs formulent leurs questions.

Le dit problème, alors qu'il constitue une entrave au développement des systèmes documentaires informatisés, est considéré comme mineur quand il s'agit des systèmes documentaires conventionnels (bibliothèques, centres de documentation). Dans ces derniers, l'existence du dialogue entre humains (utilisateur et bibliothécaire ...) apportait les précisions nécessaires à tous termes mal compris, ou ambigu. Cette constatation a été à l'origine de plusieurs travaux qui s'intéressaient au rôle joué par le bibliothécaire dans le processus de recherche d'information. Ce dernier possède une formation en documentation permettant de bien connaître les outils documentaires, et une formation en discipline de l'usager lui assurant la compréhension des besoins de l'utilisateur. Ayant ce profil, le bibliothécaire pourrait apporter "le remède" nécessaire au problème de l'inadéquation, et il contribue par conséquent au succès des systèmes documentaires.

La maîtrise des outils documentaires facilite essentiellement la localisation des documents recherchés. Cette localisation s'effectue en utilisant soit des attributs externes au contenu de document tels que: le nom de l'auteur, le nom de l'éditeur...,soit des attributs internes (mots clés, ...) relatifs au contenu de document. Les premiers sont fortement liés au document en tant que produit tangible, et de ce fait ils n'engendrent pas des problèmes d'ambiguïté et d'imprécision s'ils sont utilisés pour localiser le document, les seconds (attributs internes), résultent d'un processus de compréhension du texte. Ce processus pourrait engendrer des interprétations différentes et par conséquent des problèmes d'ambiguïté et d'imprécision dans la description du contenu de document. En effet, il s'agit bien d'un problème relatif à l'indexation qui permet la représentation du contenu de document par un ensemble de termes " signifiants ", alors que le sens du message véhiculé par le texte est loin

d'être représenté par de simples mots objets (mots clés) qui d'ailleurs diffèrent d'un indexeur à un autre.

Le bibliothécaire utilise beaucoup sa propre connaissance du contenu de fonds afin de satisfaire les besoins de l'usager car il sait bien que les mots clés n'assurent pas une bonne représentation du contenu du document.

La compréhension des comportements des usagers face aux différents problèmes de recherche d'information présuppose une connaissance de certains concepts reliés au transfert de l'information en milieu documentaire. Ainsi la notion de besoin d'information largement utilisée par les responsables des systèmes documentaires (informatisés ou non) nécessite d'être bien étudiée. Il s'agit en effet d'une notion beaucoup plus complexe qu'on le croit généralement, en raison de son caractère tout à fait immatériel.

Les recherches concernant le besoin d'information ont démontré qu'il s'agissait d'un besoin hiérarchisé. En fait, il existerait plusieurs niveaux de besoins d'information, il y avait d'abord le besoin "exprimé" qui correspond à un besoin conscient et aboutit généralement à une demande; le besoin "non exprimé" qui correspond à un besoin ressenti mais qui ne s'exprime pas convenablement; ainsi que le besoin "non activé" qui correspond à un besoin latent chez l'individu. Le bibliothécaire grâce à sa pratique quotidienne (dialogue avec les usagers), et sa propre expérience, il arrive généralement à bien cerner les besoins de l'usager et donc de les satisfaires.

Cette brève présentation du processus de transfert d'information dans le système documentaire (bibliothèque), montre qu'avec la discussion entre humains des problèmes relatifs au fonctionnement du système pourraient être évités. Ainsi, l'usager peut demander des précisions concernant tout terme mal compris, ou ambigu, utilisé pour présenter le contenu de document. Ces caractéristiques disparaissent dans la plupart des systèmes documentaires informatisés, et l'usager se trouve généralement seul face à la complexité du matériel et des logiciels d'interrogation. Ainsi, en plus des problèmes relatifs à l'indexation (des mots clés non signifiants...), s'ajoute les difficultés que l'usager rencontrait pour formuler sa requête. Pour s'en sortir les utilisateurs utilisent une variété des mots clés qu'ils estiment représentatives du sujet de recherche. Généralement ces termes de recherchesont mal choisis, d'où le degré supérieur d'incertitude dans la recherche réalisée.

A partir de ces constatations qui reflètent les problèmes auxquels sont confrontés tous les systèmes documentaires. On voit tout l'intérêt qu'il y aurait à travailler sur la modélisation de ce contexte de communication, au niveau de la description des documents prenant en compte le profil du futur utilisateur, et au niveau de la formulation d'une requête en tant que complément de la question formulée par l'utilisateur .

Dans la première partie, nous essaierons d'explorer à partir d'une bibliographie qui traite les problèmes de la communication sous des angles différent, les typologies déjà réalisées concernant les acteurs de la circulation de documents textuels, les support (périodiques, monographies) et les circuits ou réseaux par lesquels transite l'information. Dans la deuxième partie, nous présenterons les modèles existant d'indexation et de recherche d'information. Notre but est de mettre en évidence l'éventail des "solutions" disponibles en matière de recherche d'information. A la troisième partie nous justifierons la prise en compte de l'utilisateur dans la description de documents.

PARTIE I

1- Typologie des documents

1.1 Définition de document

Selon le petit ROBERT : " le document est un écrit qui sert de preuve ou de renseignement ". Cette définition renferme trois concepts: l'écrit, la preuve et le renseignement. Le premier concept traduit l'existence du texte en tant que moyen de communication, le second la valeur probante du document qui découle de son caractère tangible et le dernier l'information que renferme le document .

La définition du petit Robert n'a pas souligné l'existence d'un support d'une manière explicite, mais il serait aisé de déduire implicitement à partir de la signification des trois concepts (écrit, preuve, renseignement) l'existence de ce support. Le document serait donc à la fois le support et l'information qu'il renferme. Selon cette définition, l'article scientifique, le livre, le disque magnétique, le disque optique numérique ..., pourraient être des documents .

Une typologie de premier niveau consiste à séparer les documents textuels des documents non textuels, puis à partir de chaque type, on peut obtenir une typologie de second niveau et ainsi de suite.

Dans le présent travail nous nous intéresserons aux documents véhiculant une information textuelle. Le document se présente sous une structure bien définie qui touche à la fois la forme physique ou / et intellectuelle. Cette structure varie d'un document à l'autre, mais il existe des traits communs qui serviront de base pour une typologie .

1.2 Monographies

GUINCHAT Clair (90) décrit la monographie comme suit: " une monographie comprend en général une couverture, une page de titre, un texte divisé en plusieurs parties, une table des matières et des notes en bas de page ou enfin de chapitre, qui peuvent compléter les indications données dans le texte (référence, remarque, notes, etc.), une préface généralement écrite par une autre personne que l'auteur, une introduction, un avant propos ou un avertissement. On trouve aussi une ou des

bibliographies placées en fin de volume ou des chapitres, un glossaire ou lexique, des index et des annexes contenant des données complémentaires."

Chaque élément de cette structure véhicule une information bien déterminée et par conséquent le traitement que va subir le document sera tributaire de l'élément pris en considération. La page de titre, la couverture serviront de base pour la description matérielle du document, tandis que l'introduction, l'avant propos et la préface peuvent servir à décrire le contenu intellectuel: le sujet les intentions ...

La monographie est considérée comme une forme achevée du discours, sa structure actuelle traduit l'effort investi en matière d'organisation des connaissances; Cette structure présente la monographie comme une véritable oeuvre obéissant à des lois. D'abord celle de la normalisation qu'implique la fabrication en série de tout objet produit industriellement. Et aussi celles du marché, qui incite l'éditeur à adapter son produit à un public virtuel quitte à en sacrifier pour des raisons de prix, l'esthétique du livre (exp.: couverture illustrée par des oeuvres d'arts ...) .

1.3 Périodiques

Le nom de périodiques que l'on donne à certaines publications en séries découle du fait qu'ils paraissent à des intervalles de temps définis et réguliers. Les périodiques se présentent sous forme de numéros à contenus différents. Cette spécificité les différencie des monographies, qui elles ne sont produites qu'une fois, et traite généralement d'un seul sujet .

Pour GUINCHAT C. (90) : " la publication périodique se compose d'une couverture, toujours la même pour chaque livraison et d'un texte comportant les éléments suivants: un sommaire donnant la liste des articles et rubriques, plusieurs articles accompagnés ou non d'un résumé et d'une bibliographie. Une partie informative concernant soit la vie de l'organisme éditeur, soit l'actualité dans la discipline ou le domaine considéré. A ces éléments peut s'ajouter un éditorial, sorte d'avis de l'éditeur, signé par la direction du périodique, qui présente directement son opinion sur un sujet d'actualité ou sur les articles".

Comme la monographie, le périodique a lui aussi une structure bien déterminée et il obéit a des lois régissant son existence et sa continuité. D'abord celles de la normalisation de tout objet produit industriellement, et aussi celles du marché avec

toutes ses contraintes. Ces points communs ne cachent pas certaines spécificités des périodiques, que nous pouvons résumer comme suit :

- *rapidité de publication ;*
- *moyen de communication pour la communauté des chercheurs ;*
- *propre à un domaine ou une discipline .*

Selon B. LINE Maurice (92) :"le système de publication scientifique et technique est concentré sur les périodiques, bien que les rapports de conférences et les livres jouent un rôle important ... les chercheurs accordent une priorité à la rapidité.. . il est improbable que leurs travaux soient synthétisés ultérieurement dans des livres d'où l'importance des périodiques ."

C'est en raison de la précarité de l'information dans le secteur scientifique que le facteur temps étant particulièrement important pour les chercheurs. Ils vont surtout faire appel aux articles de périodiques et plus rarement aux monographies, sans doute parce qu'ils représentent une source généralement plus à jour. Ce rôle joué par les périodiques scientifiques en matière de diffusion des connaissances explique l'évolution que connaît ce type de document malgré l'existence de quelques problèmes conjoncturels .

B. LINE a essayé de recenser dans une typologie les différents supports des articles scientifiques . Il a cité neuf types en l'occurrence :

- texte intégral du journal;
- journal résumé;
- mini journal;
- New paper style journal;
- journal en ligne;
- disque numérique;
- digital tape;
- forme mixte (publication sur demande) .

Tous ces supports devraient répondre aux exigences de ceux qui interviennent dans le système de production des articles scientifiques (auteurs, éditeurs, bibliothèques...).

Avec l'évolution des banques de données et la vulgarisation de l'informatique, le support papier qui dominait jusqu'à présent la quasi-totalité des documents scientifiques, cédera la place a des nouveaux supports qui seront plus appropriés au traitement automatique de l'information. Les supports optiques avec leur capacité de stockage importante pourraient jouer ce rôle

1.4 Les documents non publiés

Les rapports de conférences, les brevets, les thèses..., sont des documents non publiés, appelés aussi la "littérature souterraine" ou "littérature grise". Ils sont caractérisés par des structures variables du fait qu'ils n'obéissent pas à des normes préalablement établies. Cette littérature est très sollicitée, elle échappe aux contraintes de publication et par conséquent elle est utilisée dés l'achèvement des travaux. Les documents non publiés circulent par des canaux formels tels que les bibliothèques spécialisées et les centres d'informations. Le contrôle de ces documents s'effectue par élaboration d'une bibliographie spécialisée. La littérature grise peut comprendre des documents circulant par des canaux informels de prise d'information, notamment par les contacts interpersonnels. Plusieurs enquêtes ont souligné la préférence des chercheurs pour ces canaux .

DEMAILLY 1978 - MENZEL 1968 ont cherché l'explication de ce phénomène dans les qualités intrinsèques de ces canaux: transmission rapide des informations (sans attendre leur publication), établissement de relations sélectives entre les chercheurs, circulation d'une information déjà "digérée" par autrui, définition de lignes d'action, communication du non publié (petites astuces de savoir faire, détails de montage, procédés de calcul, etc.), rétroaction instantanée.

Pour les travaux scientifiques, cette manière de collecte d'information, pose un problème de référence du fait que chaque chercheur devra apporter la preuve que ce qu'il affirme existe. La notion de référence nécessite l'accessibilité au document concerné sinon cette référence n'a pas la valeur de preuve (la définition même du document).

Selon B. LINE " les conférences peuvent donner une opportunité à la réalisation d'une nouvelle recherche ou à la discussion d'une recherche récente, mais les travaux, tant qu'ils ne sont pas publiés ne peuvent pas acquérir le statut de référence ".

Dans d'autres travaux tels que ceux des journalistes, la citation des sources n'est pas obligatoire, et la collecte d'information par les canaux informels devient une spécificité du journalisme. Selon Guinchat :"un journaliste, par exemple, peut refuser d'indiquer la source où il a puisé sa documentation; celle-ci ne demeure pas moins valable car cette façon de procéder est admise par la déontologie du métier ".

2 Typologie des producteurs d'information textuelle

2.1 auteurs

Bien qu'il soit un utilisateur averti de l'information textuelle, cette typologie présente l'auteur comme le producteur exclusif du contenu intellectuel du document. Pour déterminer une typologie des auteurs, on admettrait l'existence d'un premier niveau de distinction qui consiste à traiter chaque auteur par rapport à un secteur disciplinaire bien défini.

Actuellement il est "convenu" que les connaissances humaines sont réparties en trois secteurs disciplinaires:

- *le secteur "sciences" renferme*: la chimie, le génie, l'informatique, etc. Ce secteur est sans doute celui qui est le plus confronté a l'explosion documentaire. De ce fait, il est la cible, de presque tous les travaux en matière de recherche d'information .
-*Le secteur "sciences sociales":* il est beaucoup plus difficile à cerner que le milieu scientifique. Selon SAVARD Réjean (88) :" une étude de la fédération internationale de documentation mentionne un document de l'Unesco qui y inclut: anthropologie, géographie, gestion, linguistique, psychologie, science politique, sociologie."
-*Le secteur "science humaine":* il regroupe les disciplines tel que les arts, l'histoire, langues, la littérature, sciences religieuses, etc.

Dans ce qui suit, nous présenterons les besoins de chaque type d'auteurs (auteurs scientifiques, auteurs en sciences sociales, auteurs en sciences humaines), puis nous essaierons de voir un deuxième niveau de distinction présentant une typologie des auteurs scientifiques.

Selon SAVARD R. (88), le facteur temps étant particulièrement important pour les chercheurs scientifiques, ceux-ci vont surtout faire appel aux articles de périodiques et plus rarement aux monographies. Une étude de Meyriat J. (1984) Use of information in science and research, citée par SAVARD. Montre que dans des travaux scientifiques les articles sont utilisés quatre fois plus que les monographies. Dans le secteur sciences sociales où le facteur temps perd de sa valeur, les chercheurs utilisaient autant la littérature monographique que périodique. Le type d'information recherché en sciences sociales est souvent d'ordre statistique ou numérique, ou encore d'ordre méthodologique ou conceptuel. Dans le secteur sciences humaines, le facteur temps à beaucoup moins d'importance et les chercheurs réclament des

documents qui ne sont pas nécessairement de publication récente. C'est dans cette catégorie que l'on retrouve les plus grands utilisateurs de monographies.

L'étude présentée par S. R. illustre bien le lien étroit existant entre les besoins en information et les secteurs disciplinaires, ce qui donne à la typologie fondée principalement sur la distinction entre les champs disciplinaires des assises solides.

Aujourd'hui, il est admis que les travaux scientifiques, diffèrent des autres travaux par leur structure. Dans presque la quasi-totalité des recherches scientifiques, on trouve les items suivants: le thème, la méthode, le résultat, la discussion... A partir de cette structure presque standard, on pourrait déterminer une typologie présentant les auteurs scientifiques selon l'item le plus élaboré. Par exemple, les auteurs qui s'intéressent plus à la description des méthodes seront classés ensemble.

DEMAILLY A. (86) distingue quatre catégories de chercheurs:

- *expérimentateurs, rétifs à la théorie et se limitant essentiellement ou exclusivement au recueil de données empiriques.*
- *compilateurs qui ne lisent que la littérature et ne disposent ni de temps ni d'intérêt pour l'expérimentation*
- *gens de "relation", partout présent, discutant et échangeant des avis, connaissant les un et les autres, mais superficiellement liés au véritable travail scientifique (théories et/ou pratiques)*
- *théoriciens et philosophes des sciences qui généralisent, coordonnent, systématisent, et dégagent des lois générales... sans investir dans le travail empirique et expérimental.*

2.2 Les éditeurs

Bien étant un objet intellectuel le document textuel (livres, périodiques, ...) est aussi un phénomène économique qui reste dépendant de la politique et de l'économie générale du pays. La production du document textuel devait se conformer aux normes du marché notamment la fameuse relation qui lie l'offre à la demande. Cette notion de marché aurait besoin naturellement d'être approfondie, mais nous nous contentons de présenter un aperçu historique qui résume la dynamique du marché des documents textuels.

Selon MARTIN H. J. (90) :" jusqu'au début du vingtième siècle , le marché de la chose imprimée apparaissant comme porteur, les capitalistes n'hésitèrent pas à y investir. Il n'en reste pas moins que l'apparition d'une machinerie complexe et coûteuse renversera tous les rapports entre les différents types d'impression. Les presses multiplient désormais avant tout des journaux... privilégier les gros tirages. Il fallut dans ces conditions redoubler d'effort pour élargir le marché du livre, et ce médium à vocation largement élitaire tendit désormais à perdre de son importance sur le plan économique".

Dans ce contexte du marché, ESCARPIT Robert (82) soulève la problématique suivante: "Est ce l'auteur qui déclenche le mouvement de l'édition ou le lecteur qui, dans le circuit de la consommation, oriente l'éditeur et par suite l'auteur... c'est l'antécédent économique de l'achat par les lecteurs qui déclencherait avec quelque temps de retard la création collective des auteurs".

Cette problématique résume bien les liens étroits existants entre l'auteur, l'éditeur et le lecteur. Elle nécessite par conséquent à elle seule toute une étude. En fait, des études du marché ont montré qu'en période de hausse de la demande, les éditeurs préféraient augmenter les tirages plus que le nombre des titres nouveaux, et inversement dans la baisse préféraient réduire le nombre des exemplaires en investissant d'avantage dans les titres nouveaux. Ce phénomène est dû au fait que les gens qui achètent diminuent, d'où la baisse des exemplaires. Mais, pour maintenir l'intérêt il faut conserver le plus possible de titres nouveaux.

Cette description rapide du système d'édition montre que l'industrie des documents textuels est régie par les lois du marché. De ce fait la recherche d'une typologie des éditeurs doit tenir compte de cette réalité.

Selon B. LINE M. (92) il existe deux types d'éditeurs:

- *éditeur non commercial*, il peut être un organisme national ou international, il contribue à la diffusion de certains travaux des institutions de recherche. La règle de la gestion financière se base sur le principe de recouvrement du coût de la publication. Il s'agit bien de la définition d'une institution à but non lucratif.
- *éditeur commercial* est un véritable entrepreneur d'une entreprise à but lucratif. De ce fait, son premier souci reste le bénéfice. A nos jours ce type d'éditeurs détient la grande partie du marché de "savoir" .

Les deux types d'éditeurs manifestent parfois les mêmes besoins:
- habileté à attirer les bons auteurs;
- contrôler l'usage de leurs productions.

Ce dernier besoin touche les notions de la propriété patrimoniale et morale, qui découlent du droit d'auteur, et des lois régissant ce secteur d'activité humaine. L'approfondissement de l'étude de ce sujet pourrait aboutir a une typologie plus affinée des éditeurs.

Enfin, nous pouvons conclure que l'élaboration d'une typologie d'éditeurs sur la base de champs disciplinaires ne permet pas de tenir compte du contexte économique dans lequel s'effectue cette activité humaine. De ce fait, nous avons du mal à imaginer un éditeur (entrepreneur) qui va se contenter de la publication des travaux scientifiques alors qu'un autre produit représenterait la demande réelle du marché.

2.3 Les Producteurs des banques de données

Selon GUINCHAT C. (90): "les producteurs collectent l'information originale émise par les auteurs et éditée sous toutes formes (manuscrites, imprimées, sonores...). Ils sélectionnent, identifient l'information et l'analysent, c'est le traitement documentaire. Ce dernier se double d'un traitement informatique qui conduit les producteurs à préparer des disques magnétiques ou optiques pour le serveur. Ils créent des banques de données".

En effet, les producteurs des banques de données exploitent l'information primaire émise par les deux autres types de producteurs (auteurs, éditeurs) pour produire une information secondaire (compte-rendu, description, extrait, synthèse). Les éléments de l'ensemble de cette information s'ils sont organisés sous une structure bien définie (classement par champs disciplinaire ...), et édités sur des supports informatiques, constituent une banque d'information ou information tertiaire.

CHAUMIER J.(86) présente dans son livre "système d'information" le travail de R. BECA qui a essayé de déterminer une typologie englobant les différents producteurs de banques de données.

Dans ce qui suit nous nous contentons de présenter cette typologie qui distingue:

- *les producteurs indépendants:* ceux-ci sont producteurs d'une banque de données unique (mais souvent de grande dimension) et qui est distribuée par plusieurs

serveurs afin d'assurer la couverture du marché la plus grande possible. Exemple le Chimical Abstract Services.

- *Les producteurs en exclusivité:* il s'agit là encore, le plus souvent de producteurs d'une banque de données, mais réservant la distribution de celle-ci à un serveur exclusif. Dans ce cas , il n'est pas rare de voir le serveur contribuer au développement de la banque de données.

Exemple de ce type : les fichiers brevets de l'INPI sur télé systèmes.

- *Les producteurs intégrés:* dans ce cas le producteur est également diffuseur de sa banque de données. Le plus souvent, il s'agit de producteurs très spécialisés.

Exemple la banque Gaphyor de l'university Paris sud.

- *Les coopératives d'informations:* deux types peuvent être distingués dans cette catégorie. D'une part les banques de données orientées vers le transfert de technologie telles que Technotec de Control Data et d'autre part les banques de données réalisées au sein d'organismes professionnels sont réservées aux adhérants de ceux-ci comme les banques de l'Union des Industries Métallurgiques et Minières.

3- les systèmes documentaires

3.1 définition

Un système documentaire est un système qui stocke, gère ,et manipule un ensemble de documents de façon à permettre aux utilisateurs de trouver ceux qui correspondent à leurs besoins. On tend souvent à considérer les systèmes documentaires comme un sous ensemble des systèmes de gestion des données. Or, il existe des différences profondes entre ces deux types de systèmes. Dans les systèmes de gestion des données, les questions d'accès physique aux données sont déterminantes: temps d'accès, et la capacité à poser des questions complexes mais précises. Dans les systèmes documentaires, c'est l'accès intellectuel aux informations qui est en jeu. L'utilisateur n'a pas généralement une question précise, et le système doit lui apporter l'aide nécessaire (reformulation de la requête...).

Cater Steven C. cité par (Le Grosnier 90) a défini un système documentaire comme un octuplet ordonné:

$$SD = (\ DOC, QU, ED, EQ, EP, Fi, Fq, Fp\ )$$

- DOC: un ensemble fini de documents, dans le cas de système informatisé , les documents électroniques sont utilisés comme objets de la recherche

- QU: l'ensemble des requêtes possibles
- ED: l'espace documentaire, c'est à dire la représentation abstraite des documents qui sert lors de la comparaison entre la question et le document
- EQ: l'espace des questions représente les questions acceptables par le moteur de recherche, qui effectue la comparaison entre les questions et les documents
- EP: l'espace de pertinence, constitué par un ensemble de valeurs, en général rapporté à l'intervalle fermé [0 1], permettant d'indiquer le degré de pertinence formelle d'un document par rapport à une question
- Fi: est la fonction d'indexation qui permet de décrire les documents dans l'espace documentaire
- Fq: est la fonction de traduction qui permet de rapporter les questions exprimées en formulation libre sous une forme qui puisse être utilisée par le moteur de recherche.
- Fp: est la fonction de pertinence formelle, qui permet de déterminer le coefficient de satisfaction des documents en regard de la question posée.

3.2 Les bibliothèques:

La bibliothéconomie traditionnelle a classé les bibliothèques en trois catégories: les bibliothèques publiques, les bibliothèques universitaires et les bibliothèques scolaires. Cette typologie est basée sur des études effectuées au près des clientèles des bibliothèques. Elle révèle que pour chaque type de bibliothèque, il y a un public bien déterminé.

SAVARD R. (88) a signalé les études qui se sont penchées sur l'utilisation traditionnelle de *la bibliothèque publique*, notamment a des fins de loisirs: visite à la bibliothèque, lecture sur place, emprunt de livres, etc. Il a cité les travaux de ZWEIZIG ET DERVIN qui mentionnaient que l'utilisation de la bibliothèque pour des fins d'informations est beaucoup plus limitée: moins de cinq pour cent des personnes mentionnent la bibliothèque publique comme source dans leur quête d'information de toutes sortes! L'étude de Zweizig et Dervin a été effectuée en 1977, depuis ce temps il y a eu une grande évolution dans les services présentées par les bibliothèques publiques ce qui nécessite une mise à jour de cette recherche .

Quant aux facteurs qui influencent le taux de fréquentation de la B. P., les recherches ont recensé les facteurs suivants: le niveau d'éducationnel, niveau socio-économique, une grande diversité des supports documentaires disponibles... .

Les bibliothèques universitaires: elles se caractérisent par un public spécifique, plus homogène que celui des B. publiques. Il y a trois grands types de clientèles: les professeurs, les étudiants de deuxième et de troisième cycle, et les étudiants de premier cycle. Les professeurs utilisent d'avantage la bibliothèque pour leurs recherches et leurs publications, les étudiants de deuxième et troisième cycle pour chercher de la documentation dans le cadre d'un cours, et les étudiants de premier cycle pour y travailler, souvent avec leurs propres documents. Les types de documents utilisés par ces trois catégories ne semblent pas être les même : les professeurs et les étudiants de troisième cycle utiliseraient surtout les périodiques, tandis que les autres des manuels de base .

Les bibliothèques scolaires constituent un autre milieu qui a fait l'objet d'étude sur les modèles d'utilisations, mais dans une moindre mesure que pour les deux premiers milieux mentionnés. Une étude de Bernhard P., citée par REJEAN S. montre que la bibliothèque scolaire est généralement sous utilisée, tant par les étudiants que par les professeurs. De plus l'utilisation qu'on en fait semble se limiter pour le moment aux monographies.

3.3 les centres de documentation:

Les types observés sont davantage liés aux secteurs disciplinaires(secteur sciences, secteur sciences sociales, secteur sciences humaines). Les besoins de chaque secteur ont été discutés pour présenter une typologie d'auteurs. Cette discussion reste valable pour les centres de documentation, ce qui conduit à admettre l'existence de trois types de centre de documentation: centre de documentation scientifique et technique, centre de documentation des sciences sociales, et centre de documentation des sciences humaines.

Nous pouvons élargir cette typologie si en tenant compte du secteur économique (l'entreprise). Il s'agit des centres de documentation des entreprises. A nos jours les auteurs s'entendent pour affirmer qu'il y a deux types d'informations dans une entreprise: l'information interne, servant essentiellement à des fins de gestion, et l'information externe servant à alimenter la prise de décision à la fois du côté de la gestion et à la fois du côté de la production,. Selon WHITE D. A. (86): "les différents besoins d'information en milieu de travail formeraient un continuum allait du très général au très spécifique... certains départements de l'entreprise nécessiteraient d'avantage d'information générale , et d'autres des informations plus spécifiques (information opérationnelle), ... ".

3.4 Les Serveurs:

Le serveur est un organisme exploitant un système informatique permettant a des utilisateurs de consulter à distance un ensemble de banque de données. Ils génèrent les fichiers produits par les producteurs afin de les mettre à la disposition des utilisateurs par l'intermédiaire d'un logiciel d'interrogation. Le premier programme d'ampleur a été dévéloppé pour que tous les participants au projet d'APOLLO puissent accéder à la banque de données de la NASA. A partir des recherches sur ce mode de diffusion de l'information, le serveur DIALOG a connu le jour. Parallèlement, le Nationnal Library of Medicine a mis en ligne la banque de donnée médicale Medline en 1971. Le premier serveur a été confié à télé système Questel à la fin des années 70.

CHAUMIER J.(82) a cité une typologie des serveurs présentée par R. BECA. Nous nous contentons dans ce qui suit de citer cette typologie.

- *les serveurs polyvalents:* les serveurs de ce type, dont le prototype est le serveur américain Dialog, il compte prés de 600 banques offertes aux utilisateurs. Ces serveurs cherchent à couvrir l'ensemble du marché par le grand choix possible de fichiers et de domaines abordés. Ils développent une stratégie commerciale autour de quelques grands fichiers attractifs comme Chimical Abstract.

- *les diffuseurs spécialisés:* il s'agit ici de serveurs pratiquant une politique de créneaux avec une couverture en profondeur de ces créneaux.
- *les serveurs intégrés:* ou producteur et serveur se confondent pour diffuser leurs propres produits documentaires.

- *les fournisseurs de services spécialisés autour d'une banque de données.* Dans ce cas, ces diffuseurs sont vendeurs d'un ensemble de prestation dans le domaine de l'information et les banques de données ne représentent qu'un élément par fois minoritaire dans leur activité commerciale.

- *les réseaux de temps partagé:* il s'agit de serveur venu à la diffusion de banque de données par le biais de leur activité de service informatique. La finalité principale de leur activité reste la vente d'énergie informatique.

4 Les Utilisateurs

La complexité de l'industrie d'information (notamment les systèmes d'information en ligne) a fait naître une distinction entre l'utilisateur final et l'utilisateur intermédiaire. Ce dernier connaît bien le marché de l'information, les outils mis à la disposition des utilisateurs, et maîtrise les techniques de la recherche documentaire (rédaction de la question de recherche,...). Cette distinction est importante du fait qu'elle révèle la complexité de la recherche d'information.

Pour déterminer les améliorations qu'il faut apporter aux systèmes de recherche d'information, nous devrons repartir de l'utilisateur final, et s'attacher à bien définir les raisons pour lesquelles il a besoin d'informations et ses voies habituelles pour les obtenir. Ces voies ont été présentées dans le paragraphe 2 réservé aux producteurs d'information textuelle.

Dans ce qui suit nous essaierons de présenter une typologie des utilisateurs finals. Elle sera déterminée à partir des besoins "réels" de l'utilisateur et de son comportement face à un problème de recherche d'information.

KIRCZ J.(91) présente une typologie des lecteurs d'articles scientifiques (utilisateurs finals) comprenant quatre types: le lecteur averti "the informed reader", le non lecteur "the non reader", le lecteur partiellement averti "the partially informed reader", le lecteur non averti "the uninformed reader".
Les non-lecteurs: il s'agit d'un groupe important, constitué principalement par des administrateurs qui sont concernés par le fait qu'une personne ou un groupe de recherche a publié un article ou travail sur un thème donné. L'information bibliographique usuelle (nom d'auteur, nom de journal, volume, ...) est largement suffisante pour ce type d'usager ... l'article scientifique est utilisé essentiellement pour constituer une propre documentation et dans le cas de besoins, ils ne lisaient pas l'article entier.

Le lecteur averti: il s'agit du scientifique appartenant au même champ disciplinaire que celui de l'auteur, ou une personne ayant une connaissance de ce champ. Ce type de lecteurs connaît bien ce qu'il cherche et peut trouver la littérature rapidement. Généralement le nom de l'auteur est connu et les mots clés qui ne sont pas précis seront complétés par leur propre connaissance. ... Les questions des recherches sont claires et les techniques standards de recherche d'information en ligne sont généralement suffisants.

Le lecteur partiellement averti: ce type de lecteur comprend les personnes débutantes dans un domaine, ou celles qui s'intéressent à un champ adjacent à le leur. Le nom de l'auteur est connu à partir des références bibliographiques, mais son importance n'est pas totalement appréciée...pour ce type de lecteur les sélections par mots clés n'assurent pas la localisation exacte des articles recherchés.

Le lecteur non averti: dans cette catégorie nous trouvons le groupe le plus intéressant, il désire lire les nouvelles choses. Les lecteurs de cette catégorie sont généralement informés par leurs collègues ou par la lecture des nouvelles histoires de la presse... L'imprécision est la spécificité de leurs questions...ils ne lisaient habituellement de l'article que l'introduction, la conclusion et la liste de références.

KIRCZ J. G. signale que le scientifique dans certains cas manifeste des comportements identiques à ceux des autres catégories de lecteurs. Au même temps un article particulier pourrait servir à toutes les catégories de lecteurs bien qu'il n'est pas nécessairement à égale distance.

Cette typologie est intéressante dès lors qu'elle traite le comportement de l'utilisateur d'un type de document bien déterminé (article de périodique) dans un champ disciplinaire bien défini (le champ scientifique). Toutefois, la diversité des utilisateurs, les facettes évolutives d'un même individu (les besoins en phase initiale du travail ne sont pas identiques à ceux en phase finale) et la variété des buts de recherche rendent cette typologie non exhaustive. En effet, entre les quatre types on trouve toutes les nuances. Nous y reviendrons en detail sur cette notion de typologie d'utilisateur dans la troisième partie.

PARTIE II

1- Les différentes approches d'indexation

1.1 Indexation manuelle

L'indexation manuelle est l'opération qui consiste à déterminer un certain ensemble de concepts sur lesquels le document apporte des informations. L'analyste dans un premier temps extrait les idées fondamentales véhiculées par le document puis il procède à leur traduction, en respectant le vocabulaire et la syntaxe d'un langage documentaire.

On peut distinguer deux types d'indexation manuelle:
-Une indexation libre, où l'indexeur choisit lui même les termes d'indexation (mots-clés unitermes, descripteurs composés...)
-Une indexation contrôlée, dans laquelle les termes d'indexation sont sélectionnés à partir d'une liste préétablie. On distingue alors les " listes d'autorités ", qui ne gèrent que des relations de renvoi entre les termes, et les " thesaurus " qui définissent des relations sémantiques entre les termes.

UMBERTO Eco (1989) admet que:" le message est encore la forme vide à laquelle le destinataire pourra attribuer différents sens selon le code qu'il y applique". Si le texte est le message émis par l'auteur, c'est le récepteur (lecteur, analyste,..), qui crée le sens de ce message. Pour ce faire, il mobilise à la fois le texte , ses propres connaissances (connaissances linguistiques, connaissances du domaine, connaissances socioculturelles) et ses intentions de lecture. Donc, la lecture d'un texte et son analyse relèvent d'un processus d'interprétation du sens qui ne peut jamais s'épuiser dans une approche unique.

Ce bref examen du fonctionnement de la communication confirme la délicatesse de l'opération d'indexation. En effet, si le sens même du texte n'est pas bien défini, comment décider que tel ou tel descripteur représente bien le contenu d'un document?

Or, il est courant de lire que l'opération d'extraction est simple, qu'il suffit de lire introduction et conclusion, de parcourir les intertitres et de survoler les textes pour déceler les mots qui semblent les plus riche de signification.

Il reste à signaler que l'indexation et la condensation (la production d'un résumé) sont similaire quant à l'extraction du contenu des documents, et ce n'est qu'au moment de la traduction de ce contenu qu'elles se distinguent. Leur spécificité, quant au produit final, réside dans le degré de profondeur de l'analyse.

1.2 Indexation automatique

Avec le développement des banques de données, une véritable industrie d'indexeurs a pu émergée. Ainsi, l'indexation présente pour le système documentaire un enjeu économique très important. Les coûts d'indexation manuelle sont élevés, ils varient en fonction des compétences intellectuelle des indexeurs. Ce phénomène purement économique a braqué l'attention sur les possibilités et les limites de l'indexation automatique.

A l'état actuel des choses substituer l'indexation automatique par l'indexation manuelle, c'est admettre que la première était performante au même titre que la seconde. Or, la brève présentation de l'indexation manuelle dans le paragraphe précédent montre, qu'elle s'agit bien d'une activité intellectuelle, et de ce fait confier cette tâche à la machine, nécessite que le processus de compréhension du texte soit parfaitement illustré par un programme informatique, chose actuellement impossible malgré, les résultats relativement prometteurs des recherches en intelligence artificielle.

Pour pallier à ce problème du sens de texte, les chercheurs en indexation automatique, estiment quant en mêlant plusieurs types de modèles: le modèle statistique (travaillant sur l'état de surface de texte), le modèle fondé sur une approche " IA " (fonctionne à partir d'un réseau sémantique travaillant sur le sens du texte), et le modèle linguistique, on arrive à concevoir un système d'indexation automatique capable d'extraire du texte les mots clés représentatifs ou mieux encore produire un résumé du contenu.

A nos jours, nous ne pouvons en aucun cas considéré les produits logiciels existant sur le marché comme des simulateurs de l'activité humaine en matière d'indexation des documents. Il s'agit des outils d'aide à l'indexation permettant d'améliorer la productivité des indexeurs en produisant plus rapidement une représentation plus efficace d'un document au moment de son insertion dans le système documentaire. C'est une jonction entre la puissance de calcul de l'ordinateur et la capacité à juger de façon pragmatique de l'indexeur.

1.2.1 Les méthodes statistiques

Les méthodes statistiques d'indexation découlent essentiellement des travaux de LUHN J. B (1958). De ce fait, elles sont les méthodes automatiques les plus anciennes.

LUHN définissait ainsi cette méthode: " au lieu de tirer des échantillons au hasard, comme le fait normalement le lecteur, la nouvelle méthode automatique choisit les phrases qui représentent le mieux l'information pertinente ... La fréquence d'un mot dans un article fournit une mesure utile de la signifiance de ce mot... la co-occurrence relative dans une phrase de mots auxquels ont été affectés des poids de signifiance est une mesure utile de la signifiance de la phrase.. Plus certains mots sont trouvés souvent en compagnie les uns des autres dans une phrase, plus on peut dire que ces mots sont lourds de sens... "

Le système repère d'abord chaque mot du texte - un mot est la chaîne de caractères comprise entre deux espaces - ensuite, il calcule la fréquence d'apparition (occurrence) des termes significatifs, c'est à dire de tous les termes ne figurant pas dans un anti dictionnaire (liste de terme ayant peu de valeur documentaire, notamment les mots grammaticaux). La fréquence de ces termes est ensuite comparée à des tables de fréquences moyenne. Les termes ayant un taux de fréquence situé en dessous d'un seuil supérieur et au-dessus d'un seuil inférieur sont retenus comme termes d'indexation.

Le texte à analyser peut être comparé à un texte de référence, et sont retenus comme termes d'indexation les termes ayant une fréquence d'apparition supérieure à la constante C déterminée par l'application de la loi de Zipf .

Dans son ouvrage le traitement linguistique de l'information documentaire Chaumier J. (77) présente la loi de Zipf ainsi: "Si l'on dresse une table de l'ensemble des mots d'un texte quelconque, classés par ordre de fréquence décroissante, on constate que la fréquence d'un mot est inversement proportionnelle à son rang dans la liste, et se présente par la formule: f x r = c dans laquelle f représente la fréquence d'un mot et r le rang du mot dans le classement des occurrences".

L'évolution du vocabulaire des textes par apport au corpus de base pris comme référence pour déterminer les termes significatifs, et le comptage des occurrences

sur les formes flexionnelles et les synonymes sont à l'origine de faiblesse de cette méthode.

Salton G. (1981) a présenté une méthode statistique d'indexation basée sur deux principes fondamentaux:

- il existe une relation entre la fréquence d'un terme à l'intérieur d'un document et son importance pour la représentation de document
- il existe une relation inversement proportionnelle entre l'importance d'un terme pour l'indexation d'un document et le nombre total de documents contenant ce terme dans le système documentaire.

En adoptant ces deux principes Salton a pu substituer l'utilisation du corpus de référence par un calcul de l'indice de pertinence, qui, comme on va le constater, dépend du nombre de documents du fonds documentaire, d'où l'émergence du problème de la variation de l'indice de pertinence au cours du temps (insertion des nouveaux documents dans le système).

Dans ce qui suit on se contente de présenter les étapes de cette méthode:

1- Identifier les mots dans le document (analyse morphologique)

2- éliminer les mots vides en utilisant un antidictionnaire (liste de mots ayant peu de valeur documentaire)

3- établir une normalisation des termes destinés à représenter les documents. Cette normalisation s'effectue en utilisant soit un dictionnaire rassemblant toutes les formes possibles des entités lexicales, soit un lexique renfermant les formes canoniques des termes auxquelle on associe des règles de flexion.

4- calculer le poids de chaque terme d'indexation T_j dans le document Di

tfij = fréquence du terme Tj dans le document Di
idfj = 1/ Nombre de documents possédant le terme Tj

A partir de ces deux facteurs on peut définir un poids

$$Wij = tfij \times idfj \qquad (1)$$

5- calcul de la valeur de discrimination de chaque terme d'indexation (VDTj):
dans un système documentaire un document est représenté par 3 catégories de termes:
- un terme à usage très fréquent, c'est à dire utilisé pour représenter le contenu de plusieurs documents du système. Ce terme engendre un espace documentaire (voir la partie 1, 3.1) comprimé avec une forte densité de présence de document au tour du terme. On parle alors d'une indexation peu discriminante par des termes généraux, elle est utilisée généralement pour améliorer le taux de couverture d'une recherche.
- un terme à usage très rare qui, à l'opposé du premier, est utilisé pour représenter peu de documents du système. La densité au sein de l'espace documentaire résultant de l'adoption de cette catégorie est faible. Il s'agit d'une indexation discriminante utilisée généralement pour augmenter la précision d'une recherche.
- Un terme à usage moyen est utilisé pour représenter une partie des documents du système qui traite des sujets relativement semblables.

La densité Q de l'espace documentaire est définie comme étant la moyenne de la similitude calculée entre tous les documents pris deux à deux:

$$Q = \frac{2}{m\,(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} S(Di, Dj) \qquad (2)$$

m = le nombre de document du système
S (Di, Dj) = la similitude entre les documents (Di et Dj)
Di = < T1 Wi1; T2Wi2; .....TnWin>: La représentation d'un documents Di par ses n termes d'indexation

De l'équation (2) il résulte que la densité de l'espace documentaire varie avec l'ajout ou le retrait d'un terme d'indexation. On peut calculer ainsi la nouvelle valeur de Q après retrait du terme Tj

$$Qj = \frac{2}{m\,(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} S(D_i^j, D_j^j) \qquad (3)$$

C'est la différence Qj- Qi qui va déterminer la valeur de discrimination de Tj . On a ainsi: $VDTj = Qj - Q$ (4)

6) Grouper les termes de faible fréquence qui sont caractérisés par un VDTj >0. Il s'agit des termes trop précis tendant à éloigner les documents les un des autres dans l'espace documentaire.

7) constituer des descripteurs à partir des termes trop généraux ayant une VDT < 0 .
8) Indexer chaque document avec les termes de groupage, les descripteurs composés et les termes originaux. Chaque descripteur est affecté d'un poids qui sera déterminé par la formulation faisant intervenir la valeur de discrimination du Tj

$$Wij = (tfij \times idfj) \times VDTj \qquad (5)$$

En introduisant la VDT , qui est en fait une pondération des informations des informations contenues dans les descripteurs, cette méthode d'indexation a pu adapter les représentations aux besoins documentaires du système. L'indexation n'est plus centrée uniquement sur le document sans prendre en considération le fonds documentaire dans lequel il doit prendre place.

1.2.2 Les méthodes fondées sur une approche " I A".

Dans les méthodes statistiques d'indexation, des connaissances d'ordre morphologique et lexical sont mobilisées ainsi que des techniques de jugement de l'importance d'un mot à représenter le contenu du document. Dans la présente méthode , c'est la propriété de l'appartenance du mot a un réseau sémantique complexe qui est recherché. Les descripteurs ne sont pas issus de surface du texte mais, recomposés à partir des connaissances du système concernant les liens sémantiques entre plusieurs termes. C'est une tentative de simulation du processus d'indexation d'un indexeur humain.

Cette notion de réseau sémantique résulte d'un besoin de classification des concepts et d'établissement d'un modèle du " monde" (une situation, un acte, un discours).
Les réseaux peuvent être considérés comme une extension de la notion de thesaurus dont la représentation du domaine de connaissance est assez rudimentaire. Les noeuds du réseau représentent des concepts, et les arcs illustrent les relations entre ces concepts. Ces relations donnent la sémantique du réseau.

Le "sens" d'un noeud du réseau est défini par l'ensemble des concepts qui lui sont rattachés par l'intermédiaire des arcs. On définit souvent les relations de type:
"est un (e)"qui relie un terme à son ou ses génériques

" sorte de" qui est plus spécifique
" partie-de" est spécifique des classes d'objet complexe.

Dans cette approche d'indexation, en plus de l'analyse morphologique visant la reconnaissance du mot et l'analyse lexicale permettant la réduction d'un mot a sa forme canonique, une analyse syntaxique s'est avérée nécessaire pour ne conserver que les phénomènes syntaxiques théoriquement porteuse de sens. Ainsi les homographes seront levées par la reconnaissance des termes n'appartenant pas à la même catégorie grammaticale telle que ferme= substantif, ferme = adjectif, ferme = verbe; les formes vides seront éliminées sur la base de la reconnaissance de la catégorie grammaticale comme pour or = conjonction et or = substantif.

Les expressions retenues du texte après ces analyses font l'objet d'une comparaison avec les termes descripteurs (noeuds) du réseau sémantique. Cette approche malgré qu'elle se présente comme une phase avancée de l'indexation automatique du fait qu'elle applique des méthodes sémantiques utilisant des outils documentaires évolués (réseau sémantique,...), elle reste loin de la resolution des problèmes d'ambiguïtés sémantiques du texte.

Cette faiblesse résulte de la conception même du réseau de représentation des connaissances. Ce réseau ne tient pas compte de l'évolution du domaine couvert.

1.2.3 L'approche linguistique

Dans son article l'indexation automatique assistée par ordinateur CHAUMIER J. (92) admet que, le modèle linguistique engendre cinq niveaux d'analyse, permettant de lever les ambiguïtés du langage. Ces cinq niveaux sont:
- le niveau morphologique
- le niveau lexical
- le niveau syntaxique
- le niveau sémantique
- le niveau pragmatique.

Le dernier niveau intègre la représentation sémantique dans le contexte, et éventuellement il la modifie. La représentation pragmatique est construite à partir des règles pragmatiques et des réseaux pragmatiques. Nous y reviendrons dans le paragraphe suivant.

Ces cinq niveaux d'analyse on les rencontre aussi dans l'architecture en série d'un système de traitement automatique des langues naturelles présenté par Sabah G. (90) dans son livre l'intelligence artificielle et langage. Sans entrer dans les détails de ce modèle, d'ailleurs ce n'est pas le but du présent travail, nous pouvons dire qu'il s'agit d'un modèle cohérent étant donné qu'il recense pratiquement tous les niveaux d'analyse permettant la compréhension du sens de texte mais, ceci ne nous empêche pas de signaler les limites des règles sémantiques et pragmatiques, utilisées respective dans l'analyse sémantique et l'analyse pragmatique.

Les méthodes d'indexation présentées plus haut, n'arrivent généralement pas à décrire convenablement le contenu de document. Les problèmes qui sont à l'origine de cette indexation "incertaine" sont nombreux. On cite à titre de rappel, le problème relatif au processus de lecture. En fait, ce qui est pris comme une description du contenu de document, dans le meilleur des cas, , n'est qu'une simple transcription de l'interprétation de l'indexeur. Dans une opération d'indexation "l'outil" de cette transcription n'est autre que le mot du lexique de la langue. Or, l'utilisation d'un tel outil s'est avérée aussi problématique .

M. Le Guern(91) admet que: " les mots de la langue, en tant qu'ils sont mots de la langue ne signifient que de propriétés, jamais des entités: ils signifient des attributs et non des substances, tant qu'ils ne sont pas mis en oeuvre dans le discours. Le descripteur, quant à lui, signifie une entité, une substance, au sens de la philosophie d'Aristote. Le descripteur ne peut donc être considéré, à l'instar des mots de la langue, comme un symbole sans référence....Cette notion de valeur référentielle, qui est la nature même du discours a nécessité l'utilisation d'une unité du discours au lieu d'une unité du lexique de la langue".

L'approche du groupe SYDO* consiste à extraire du texte des syntagmes nominaux considérés comme étant les plus petites unités du discours porteuses d'une valeur référentielle.

Bouché R. (89) résume en trois points les objectifs du modèle linguistique adopté par l'équipe SYDO:
- permettre l'identification des syntagmes nominaux
- déterminer la structure de ces syntagmes en mettant en évidence les relations entre ses constituants. Ceci permet le stockage d'une représentation du syntagme nominal facilitant la recherche d'informations.

---

* - L'équipe SYDO regroupe actuellement : (R.Bouché, M. Le Guern, J.-P. Metzger, S. Lainé, M.Hassoun).

- Bien montrer le mécanisme de passage partant des mots (prédicats fonctionnant dans une logique intentionnelle) et arrivant à une unité valeur référentielle (le syntagme nominal, dans le cadre d'une logique extensionnelle).

Ce qui diffère ce modèle des autres modèles linguistiques c'est l'absence d'une étape d'analyse sémantique, permettant l'interprétation du sens du texte. Le non recours à cette étape à été justifié par le fait que la structure syntaxique en elle même véhicule un sens.

Le Guern M. (91) argumente cette thèse ainsi: " que le syntagme nominal soit une structure syntaxique est une évidence mais il faut y voir aussi une structure logico-sémantique. Si la syntaxe se limite à l'étude du fait de position, de rection et d'accord, elle n'a que peu d'intérêt, et je considère qu'il ne s'agit là que d'une partie de la syntaxe, de sa partie superficielle". Nous y reviendrons sur la relation entre la syntaxe et la sémantique dans la troisième partie.

1. 3 Interprétation pragmatique du texte

Selon Sabah G.(92), une réelle communication demande de dépasser le sens littéral et d'en calculer les interprétations possibles. Toutes nos connaissances sont alors utiles: la réalité est ses contraintes, les aspects culturels, les informations sur la situation, permettent de découvrir les sous-entendus cachés derrière l'énoncé (on parle alors de pragmatique). Dans ce contexte, Récanati F.(92) admet aussi que : " la principale leçon de la pragmatique est que le message communiqué est toujours très largement sous déterminé par le matériel linguistique utilisé et ne se confond pas avec la signification grammaticale de la phrase". De ce fait, pour comprendre le message véhiculé par un énoncé il ne suffit pas d'être linguistiquement compétent et de saisir la phrase au sens grammaticale du terme.

L'interprétation pragmatique est généralement la dernière étape d'un processus de traitement automatique du texte, elle cherche à modifier les résultats de l'analyse sémantique par la prise en considération des éléments extérieurs au texte.

Un exemple, tel que : " pouvez-vous me dire l'heure qu'il est", est une question qui, au sens propre appelle une réponse par oui ou par non. Les conventions sociales, qui déconseille les formes trop impératives du type " dites l'heure qu'il est ", conduisent tous ce qui entendent cette question à en modifier les sens et à l'interpréter comme une manière polie de demander l'heure.

L'idée, d'affirmer qu'il ne suffit pas de comprendre le sens de surface d'un texte, et qu'il faut pouvoir en comprendre les implications profondes, principalement les buts les intentions et les stratégie du locuteur, résultent du fait que la plupart des informations transmises par le discours en langage naturel sont implicites.

Berrendonner A. (91) admet que: " Ces informations implicites ne sont pas littérales, c'est à dire qu'il n'existe pas de couplage immédiat et nécessaire associant chacune d'elles a une occurrence de signifiant linguistique."

Pour déterminer ces informations implicites véhiculées par un texte, plusieurs chercheurs se sont efforcés à construire des règles pragmatiques, généralement emprunté de la logique mathématique. L'approche le plus simple consiste à admettre l'existence d'un système de "script". Les scripts sont des séquences d'événement correspondant au déroulement typique d'une action parfaitement banale. Déroulement que l'auteur d'un texte n'a aucune raison d'expliciter. Par exemple, l'auteur suppose que chacun de ses lecteurs sait que " se rendre à son bureau " se décompose fréquemment en: sorti de son logement rejoindre à pieds un moyen de transport, l'utiliser... En effet, tout ce passe comme si l'auteur n'exprime pas une situation " se rendre à son bureau" qui nécessite les étapes cités plus haut mais, il exprime la différence entre la situation qui veut décrire et celle qu'il sait que le lecteur connaît déjà.

Cette approche utilisée, dans la compréhension de certaines situations engendrant des événements simples caractérisés par des actions parfaitement typiques, n'est plus valable, quand il s'agit des situations complexes aux quelles, on devait appliquer des processus de compréhension plus développés. En effet, il n'est pas possible d'énumérer l'ensemble des scripts auquel est associé chaque énoncé. Cela est dû principalement au fait que les déductions interprétatives opèrent toujours sur un ensemble indéterminé de scripts

Si on essaie de traiter l'énoncé suivant: " vous constaterez que l'art. 8 prévoit une composition de la nouvelle commission très différente de celle de l'ancienne. Je voudrais donc, par la présente, vous remercier vivement de l'action que vous avez eue (...) et vous dire le plaisir que j'ai eu à collaborer avec vous. ".

Dans cet énoncé, on peut connaître aisément deux informations littérales:
- la composition de la nouvelle commission est différente de celle de l'ancienne
- un remerciement...

Ces deux informations ne puisent pas le sens profond de l'énoncé car en fait le message "réel" est implicite, et il se situe entre les deux informations littérales. Le locuteur et pour cause d'euphémismes n'a pas utilisé l'expression: " vous ne faites plus partie de la commission".

La situation décrite par cet énoncé est loin d'être appréhendée en appliquant les systèmes des scriptes.

Selon Berrendonner A.: " un discours sera perçu comme cohérent s'il admet une interprétation fortement connexe (...) cette connexité repose crucialement sur deux implicites: dans le "graphe" que constitue une interprétation cohérente, la plupart des éléments, sommet (objets, états des choses) et arcs (relations), sont des informations qui n'ont pas été verbalisées littéralement".

Le travail interprétatif consiste donc à chercher l'élément de l'énoncé établissant les relations existantes entre les informations littérales et / ou implicites.

B. A. propose trois classe de phénomène illustrant ce fait: les connecteurs des anaphores associatives et les métonymies.

En utilisant les connecteurs, il commente l'énoncé cité plus haut comme suit: le connecteur "donc" n'enchaîne pas sur le contenu littéral qui vient d'être asserté, mais sur un sous-entendu de ce dernier implicité pour cause d'euphémisme: vous n'êtes plus membre de la commission.

Prenons l'exemple, qui a été traité par Kayser D. (87): "le professeur a envoya l'élève chez le censeur parce qu'il voulait lancer des boulettes".

La compréhension du sens de cet énoncé nécessite le recours à un mécanisme de calcul inférentiel. Selon Kayser (90): " étudier le langage sans prendre en compte les opérations inférentielles que l'usage du langage présuppose, c'est ignorer délibérément un élément fondamental du problème".

Pour mieux saisir cette notion d'inférence, on cite le point de vue de Berrendonner A.: " l'inférence c'est l'opération par laquelle un sujet, à partir de CRIS (connaissances, représentations, informations, significations) des statuts variés (contenus linguistiques explicites, évidences perceptives, savoirs implicites) construit

une nouvelle CRIS implicite. L'inférence en ce sens n'est autre que l'unité élémentaire de raisonnement".

De ce qui précède, on peut admettre que le calcul inferentiel permet de déterminer les sous-entendus de l'énoncé à partir de ce qui a été dit et des connaissances dont nous disposons.

Dans le cas de l'exemple cité plus haut:
ce qui a été dit:
- le professeur envoya l'élève chez.....
- il voulait lancer des boulettes

Les connaissances dont nous disposons:
- il est usuel que les élèves s'ennuient
- une personne qui s'ennuie peut chercher à se distraire
- un élève qui se distrait en classe risque une punition

Donc on peut inférer qu'il s'agit bien de l'élève qui voulait lancer des boulettes.

La question qui se pose maintenant est de savoir si le calcul inferentiel est conforme au modèle de la construction logique.

Lors de traitement d'un énoncé, un calcul d'implicite comprend souvent plusieurs pas inférentiels successifs, la conclusion de chaque pas peut être réutilisé comme thèse lors d'un pas ultérieur. De même un processus logique analogue par exemple, à celui de la démonstration, s'opère par application répétée de règles d'inférences . Ainsi on peut assimiler la compréhension à un modèle de démonstration logique

Par exemple, si on prend l'opération d'inférence déduit on peut schématiser ce type d'inférence sur le modèle du modus Ponens.

P vrai .....thèse
P $\Rightarrow$ Q vrai.....loi implicative
.....
Q (vrai) ..... conclusion

Kayser (85) conteste l'assimilation de la compréhension à un processus logique, pour lui: alors que le raisonnement logique ordinaire est "monotone" (lorsque des

informations nouvelles arrivent, il est possible d'en déduire des nouvelles conséquences, mais pas de remettre l'acquis en cause) les raisonnements usuels tolèrent souvent des conclusions par "défaut"; par exemple on sait une connaissance comme : "si une personne P veut avoir les nouvelles du jour, il peut s'acheter un journal

$$P \Rightarrow Q$$

nouvelles du jour $\Rightarrow$ journal

Cette relation quasi implicative peut mener à des conclusions fausses si les ouvriers du journal sont en grève ou si la personne est aveugle; mais on se donne le droit de l'utiliser sans prendre le soin de vérifier qu'aucune exception ne s'applique.

L'assimilation de la compréhension à un processus logique demeure un problème très complexe mais, nous pouvons admettre que le calcul inférentiel obéissent à des régularités, puisqu'on parvient à communiquer par ce moyen, et que la logique intervient pour exprimer cette régularité.

2- Les modèles de recherche d'information

Le but ultime d'une fonction de recherche d'information documentaire est de déduire ou de rechercher les réponses aux questions de l'utilisateur. Ce but est très loin d'être atteint dans sa généralité. Toute la difficulté consiste à apparier l'intention des utilisateurs aux besoins multiples, exprimés sous une certaine forme dans leur propre langage, avec le même sujet exprimé sous une forme différente par l'auteur du texte.

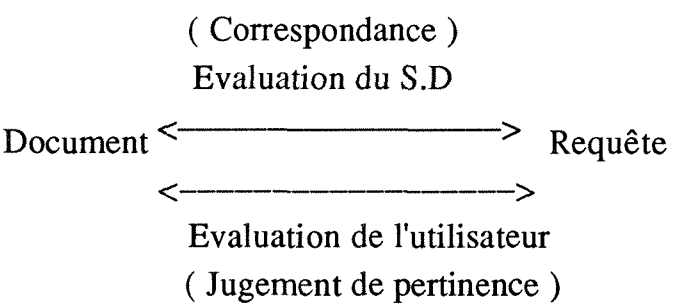
Dans un contexte de communication où l'utilisateur n'a accès qu'aux informations secondaires (information traitée), le problème de l'appariement entre les besoins exprimés par l'utilisateur et les informations dont dispose le système, est un problème résultant essentiellement de la représentation adoptée par le système documentaire. Généralement cette représentation ne prend en compte que la formulation choisie par l'auteur, et donc si cette représentation véhicule un sens, ce sens sera celui qui a été mis par l'auteur et jamais celui qui est construit par le lecteur.

Cette approche de la représentation du sens de texte est ancrée dans la littérature des sciences de l'information, elle n'est autre qu'une tentative de saisie de la stabilité

sémantique d'un texte indépendamment de son usage (contexte, compétence pragmatique du lecteur, etc.). Malgré son caractère réducteur (le texte véhicule un seul sens), cette approche n'arrive même pas à extraire le sens voulu par l'auteur étant donné la diversité avec laquelle les mêmes choses peuvent être exprimées par la langue.

Si le but de la fonction recherche demeure la détermination des réponses aux questions de l'utilisateur, ce bref rappel du problème de la description de documents textuels nous conduit à admettre que les réponses ne peuvent être satisfaisantes que si la représentation du sens du document est la plus proche possible de celle de l'utilisateur. La fonction recherche d'information d'un système documentaire doit assurer l'appariement entre le besoin formalisé de l'utilisateur (la requête) et l'information disponible (documents du système). Cet appariement est une évaluation du système qui peut être fondée soit sur une comparaison stricte des documents et de la requête soit sur une comparaison plus souple, prenant en compte la sémantique de la requête et celle des documents.

Cette évaluation au niveau du système et généralement accompagnée par une évaluation au niveau de l'utilisateur. En fait, lorsqu'un document est fourni comme réponse à une requête, l'utilisateur évalue lui aussi la correspondance du document à son besoin. Son évaluation n'est pas nécessairement celle du système. La relation entre documents, requête et évaluation peut être représentée comme suit



2.1 Le modèle booléen

La plupart des systèmes documentaires utilisent le modèle booléen et, on peut même admettre qu'il sert de base de référence à presque tous les modèles de recherche. Ainsi, par exemple, le modèle de Salton est un modèle booléen étendu.

Dans le modèle booléen, un document (D) est représenté par un ensemble de termes (mots ou expressions composées selon les choix suivis au niveau de l'indexation). Une requête est une expression logique composée de termes connectés par les opérateurs logiques ET, OU et SAUF. Les documents sont retrouvés ou non suivant la présence ou l'absence des termes utilisés dans l'équation de recherche. Le modèle est donc caractérisé par une évaluation de pertinence binaire ("vrai" ou "faux")

Si on définit une fonction de recherche P(D,Q) pour évaluer la pertinence d'un document D pour une requête Q. L'évaluation booléenne peut être plus précisément définie ainsi:

- $P(D,ti) = 1$ si $ti \in D$ (ou ti est un terme)
- $P(D,t1 \wedge t2) = 1$ si $P(D,t1) = 1$ et $P(D,t2) = 1$
- $P(D,t1 \vee t2) = 1$ si $P(D,t1) = 1$ ou $P(D,t2) = 1$
- $P(D,\neg t) = 1$ si $P(D,t) = 0$

- *Les opérateurs booléens:*

le connecteur ET exige la présence simultané des termes dans la représentation d'un document.

exemple: communication et (réseau téléphonique)

le connecteur SAUF permet d'éliminer les documents représenté par un terme donnée.

exemple (Indexation automatique) SAUF (Méthodes statistiques)

le connecteur OU exige la présence de l'un ou moins des termes dans la représentation des documents localisés

- *Les limites du modèle booléen*

bien que la structure des requêtes booléennes permette à l'utilisateur d'exprimer les éléments principaux de sa recherche, le taux de satisfaction de ses besoins en matière d'information pertinente demeure faible et dépend essentiellement du degré de maîtriser de la technique de formulation de la requête. Des critiques nombreuses ont été portées sur le modèle booléen, notamment sur sa rigidité:

- Une requête: W1 OU W2 OU W3 extraira les documents qui contiennent au moins l'un des trois termes. Si un document contient les trois termes ou deux des trois termes, il ne sera pas favorisé et présenté en premier lieu à l'utilisateur. Ainsi, les documents (résultats de la recherche) ne sont pas ordonnés en fonction de leurs degrés de pertinence pour la question de l'utilisateur.

- Une requête: W1 ET W2 ET W3 extraira les documents qui sont décrits par les trois termes indépendamment de l'importance relative pour l'utilisateur des trois termes. Un document contenant deux des trois termes sera rejeté de la même manière qu'un document n'en contenant aucun. En d'autres termes, le modèle booléen ne peut pas fournir un résultat de type: un document traite principalement de tel sujet, mais évoque tel autre et survole un troisième.

- La logique booléenne est difficile à assimiler par l'utilisateur si elle est utilisée pour traiter des expressions linguistiques. Par exemple, l'expression " commerce et industrie " doit s'exprimer "commerce OU industrie". Le sens booléen des connecteurs ET, OU et SAUF est diffèrent du sens qu'ils ont généralement dans la langue quotidienne.

2.2 Le modèle probabiliste

Dans le paragraphe précédent nous avons évoqué que parmi les limites assignées au modèle booléen, on recense son incapacité à fournir à l'utilisateur des résultats ordonnés par leur degré de pertinence pour l'usager. Pour dépasser ce problème, les documents extraits par une requête doivent être classés en fonction de leur probabilité décroissante de répondre à cette requête, cette probabilité étant évaluée à partir des données dont le système dispose.

Etant donné une requête Q, elle découpe la banque de donnée en deux ensembles: les documents pertinents et ceux qui ne le sont pas. La probabilité que la réponse soit pertinente étant donné D comme réponse est exprimée sous une forme conditionnelle:
P (per/ D) = P1
D'après le théorème de Bayes:

$$P1 = \frac{Pp(D) \times P(per)}{P(per) \times Pp(D) + P(n\ per) \times Pn.p(D)}$$

Pp (D) et Pn.p (D) représentent respectivement la probabilité pour D d'être un document pertinent ou non pertinent.

P(per) et P(n per) sont respectivement la probabilité a priori de trouver des documents pertinents et non pertinents.

Pour un corpus donné, P (per) et P (n per) sont supposés fixés. Ainsi l'estimation de P1 est donc ramenée à celle de Pp (D) et Pn.p (D). Un moyen de les estimer consiste à établir les probabilités de pertinence et de non-pertinence d'un document en fonction des probabilités des termes contenus. Ces probabilités seront notées:
Pp (ti), Pn.p (ti). Elles sont évaluées de la façon suivante:

$$Pp\ (D) = \prod_{ti \in D} Pp\ (ti)$$

$$Pn.p\ (D) = \prod_{ti \in D} Pn.p(ti)$$

Le choix initial des documents pertinents et non pertinents rend les probabilités établies non crédibles. L'hypothèse que les termes d'indexation sont indépendants les uns des autres est trop simplificatrice. En fait, pour une raison de sens, l'évidence sémantique nécessite que les termes ne soient pas indépendants (Voir 1.2.3).

2.3 Le modèle vectoriel

Dans ce modèle les termes d'indexations forment une base de l'espace vectoriel dans lequel chaque document est un vecteur représenté par une combinaison linéaire de ces termes:

$$D = ai1T1 + ai2T2 + \text{------} ainTn$$

Les coefficients a sont les poids du terme T pour le document D . Ce dernier n'est en réalité représente que par un nombre limité de terme d'indexation de ce fait on peut considéré que la majeur partie des coefficients sont nuls.

Une requête Q est aussi formalisée par un vecteur de termes de recherche pondérés:

$$Qi = bi1T1 + bi2T2 + \text{----------} binTn$$

b étant le poids du terme T dans la requête.

Durant le processus d'évaluation de la pertinence d'un document en regard de la question Q, le système sélectionne les documents D dont la similarité avec la requête est la plus grande.

Il existe diverses méthodes pour calculer cette similarité dont la plus importante est celle de Salton. Elle consiste à mesurer le cosinus de l'angle des deux vecteurs Q ET D. Les documents sont alors retrouvés en fonction de leur proximité avec la question

.

$$\cos(Q, D_i) = \frac{\sum_{j=1}^{n} bj\ aij}{\sqrt{\sum_{j=1} bj2 \sum_{j=1} aij2}}$$

Les documents pertinents ont une valeur de cosinus qui tend vers 1. Ainsi, la réponse à la requête est constituée par un ensemble de documents classés en fonction de cette mesure. Cette possibilité de jugement de pertinence est utilisée pour reformuler les questions. Le vecteur de la requête se déplace pour se rapprocher des vecteurs représentant les documents non pertinents et s'éloigne des vecteurs représentant les documents non pertinents. La nouvelle équation de recherche s'écrit en fonction des documents pertinents:

$$Q = \sum_{j=1} \left(bj + \sum_{Di.per} \alpha\ aij - \sum_{Di\ n.per} \beta\ aij \right) Tj$$

Les valeurs de α et β permettent de moduler la prise en compte des types de documents (pertinents et non pertinents)

Ainsi conçu, le modèle vectoriel parait cohérent mais comme pour le modèle probabiliste l'affirmation de l'indépendance des termes d'indexation rend ce modèle trop simplificateur. En effet, malgré la prise en compte du poids d'un terme pour un document précis, cela ne permet en aucun cas de substituer les liens sémantiques existant entre les termes et qui sont à l'origine de la force du langage naturel.

2.4 Les modèles sémantico-linguistiques

Ce modèle vise d'une part à rendre l'accès à l'information plus aisé en permettant à l'utilisateur de formuler sa requête en langage naturel, d'autre part, à fonder l'opération de recherche sur la sémantique (utilisation d'un réseau sémantique). Cette possibilité de formuler la requête en langage naturel facilite énormément la

tâche aux utilisateurs occasionnels n'ayant pas une bonne connaissance des termes autorisés (terme d'indexation) et ne sont pas familiarisé avec la logique booléenne.

L'interprétation de la requête est un processus qui fait appel à une analyse linguistique (une analyse morpho-syntaxique est une reconnaissance d'expression idiomatique, une désambiguisation grammaticales) et une analyse sémantique utilisant généralement une base de connaissance comprenant des relations sémantiques entre les termes. Ces relations permettent la représentation de différents domaines de connaissances. Mais, à cause de la complexité de la langue naturelle, le processus d'interprétation de la requête comme celui d'indexation ne peut jamais représenter exactement la requête. Une certaine perte de précision est toujours introduite. De ce fait, on ne peut pas admettre que la mesure de correspondance, entre la représentation interne de la requête (interprétation du système) et celle du document (des mots clés), fournie par le système est identique au jugement de pertinence de l'utilisateur.

A ce stade d'étude, nous retenons que pour la plupart des modèles sémantico-linguistique l'évaluation de la requête consiste à mettre en correspondance les termes fournis par l'usager et ceux du système (terme d'indexation). Cette opération est réalisée via un appariement syntaxique et sémantique. Une fois cette transformation réalisée (le passage du langage naturel au langage de représentation du système), on aboutit à une expression booléenne de termes d'indexation permettant la sélection des documents. Ces derniers sont ordonnés selon la valeur de la mesure de correspondance avec la requête.

2.5 Le modèle hypertexte

Un document n'est jamais un produit intellectuel isolé. Au contraire, il existe par son environnement. Il tend à apporter des réponses, des contradictions ou à enrichir un débat en posant des nouvelles questions. Un document est un élément au sein d'un ensemble de documents, produit dans le " micro-monde " intéressé par des thèmes similaires. La réalité d'un document n'est pas uniquement incarnée dans ce document lui même , mais dans les "forces de liaison" au sein du micro-monde documentaire dans lequel il s'insère. (Le Grosnier H. 90)

Cette définition "exhaustive" du document n'est pas en contradiction avec l'idée de Vanne Var Bush, qui, en 1945 dans son article intitulé " as we may think ", préconise que la plupart des systèmes d'indexation et d'organisation des informations en usage

dans la communauté scientifique sont artificiels. Chaque item n'y est classé que sous une seule rubrique et le rangement y est purement hiérarchique (classe, sous classe, etc.). Or, dit Vanne Var Bush, l'esprit humain ne fonctionne pas ainsi mais par association.

Pour mieux illustrer ce principe, considérons l'exemple suivant. Quand nous rédigeons un article, nous affirmons quelque chose en faisant référence à une autre chose. Ainsi par le biais d'un numéro de référence, nous renvoyons le lecteur à une autre partie de l'article qui, souvent elle même revoie à un autre ouvrage ou un autre article. Ces liens établis entre, d'une part, les différentes parties du texte, et d'autre part, entre le texte lui même et les autres textes traitant des items similaires, sont à l'origine du pouvoir convainquant du document

Cette conception de la production documentaire à inspiré à Ted Nelson (dans les années 60) la réalisation d'un projet qui groupe toute la littérature sur un domaine. Ce projet permet de circuler entre les textes par des liens non hiérarchiques. Chaque document s'inscrit dans le contexte de tous les autres et entretient des rapports explicites (la citation) ou implicite (analogie) avec un certain nombre d'entre eux. Pour designer ce projet Ted Nelson invente le terme hypertexte

Techniquement, un hypertexte est un ensemble de noeuds connectés par des liens. Les noeuds peuvent être des mots, des pages, des images, des graphiques ou parties de graphiques, des documents complexes....Les items d'information ne sont pas reliés linéairement mais chacun d'eux étend ses liens en étoile, sur un mode réticulaire. Ces liens permettent de naviguer d'un noeud à l'autre très rapidement.

Fonctionnellement, un hypertexte est un environnement logique pour l'organisation des connaissances ou des données, l'acquisition d'information et la communication. Avec un logiciel de type hypertexte, il est possible de concevoir des systèmes d'information " orienté utilisateur" adaptés à la problématique de chacun d'eux bien que le système global soit le même pour toute une catégorie d'entre eux. La recherche d'information s'effectue au gré des problèmes rencontrés, selon une démarche intellectuelle propre à l'utilisateur.

Si nous admettons l'hypothèse que, l'être humain procède par association des idées et des informations pour acquérir ou améliorer des connaissances sur un domaine donné, on voit tout l'intérêt qu'il y aurait à concevoir des programmes permettant de circuler de façon non séquentielle (par association) au sein d'une masse

d'information. En revanche, ce mode d'accès à l'information (hypertexte) pose plusieurs problèmes aux utilisateurs , que Carolyn Foss cité par Le Grosnier nomme:

- le problème des digressions en chassées (the embedded disgression problem)
- le problème du musée d'art (the art museum problem)

Dans le premier cas , l'utilisateur suit des divers chemins de recherche et finit par se perdre dans la masse d'information. Le second cas représente la situation d'une personne qui a vu de nombreux éléments d'information, mais qu'il finit par ne plus savoir les distinguer les uns des autres et par ne plus savoir généraliser à partir de ces éléments d'information pour en faire un savoir cohérent.

L'hypertexte doit être structuré à fin de faciliter le butinage, mais doit aussi permettre de filtrer l'information en fonction des buts d'un utilisateur.

PARTIE III

1- Modélisation de l'utilisateur

Les besoins des utilisateurs sont d'une variété quasi illimitée. De ce fait la modélisation de l'utilisateur, bien que difficile, est une nécessité car elle servira de référence lorsqu'il faudra procéder à l'évaluation de la pertinence des documents indexés par rapport à la requête interprétée. Donc, il est préférable d'aboutir à un modèle même incomplet et imprécis que de ne pas avoir de modèle du tout.

Dans le paragraphe D de la première partie, nous avons essayé d'établir une classification des utilisateurs en lecteurs avertis, les non-lecteurs, les lecteurs partiellement avertis et les lecteurs non avertis. Cette classification reste générique est nécessite la détermination des paramètres qui entre en jeux dans la définition des catégories des lecteurs. Ainsi, le passage du générique (profil type d'utilisateur) au spécifique (profil d'un utilisateur donné) est obtenu par une simple pondération de ces paramètres.

Le profil type de l'utilisateur que nous essaierons de déterminer s'appuie sur trois niveaux: le premier représente les connaissances dont dispose l'utilisateur (niveau cognitif), le second décrit les aspects techniques de la réalisation des tâches informatisées (niveau technique), et le dernier est relatif au but de la recherche d'information.

1.1 Niveau cognitif

Etant donnée la complexité de la dimension cognitive, nous ne pouvons en aucun cas prétendre que les éléments invoqués pour décrire ce niveau reflètent d'une manière exhaustive et précise le problème et le processus cognitif de l'utilisateur. Toutefois, nous estimons que si ces éléments sont bien explicités, le comportement de l'usager en quête d'information pourrait être déterminée et par conséquent l'établissement d'un profil type d'utilisateur devient possible.

1.1.1 Niveau éducationnel

Nous avons procédé à une classification académique. Ainsi, on trouve les niveaux suivants: niveau primaire, niveau secondaire, premier cycle universitaire, deuxième cycle universitaire, troisième cycle universitaire (DEA) et recherche.

1.1.2 Champs disciplinaires

Nous admettons que le savoir humain est réparti en trois champs disciplinaires: secteur " sciences et techniques ", secteur " sciences sociales " et secteur "sciences humaines ". Cette classification son exhaustivité, elle renferme des limites intrinsèques. En fait, cette répartition des savoirs en trois classes n'ayant aucune relation entre elles, pose un problème d'appartenance quant-il s'agit de déterminer la classe d'une science interdisciplinaire. En plus la mise à jours de cette classification (insertion des sciences nouvelles) n'est pas toujours évidente. Ainsi, une personne qui n'est pas familiarisé avec cette classification académique trouve du mal à situer son domaine d'activité.

1.1.3 Niveau de familiarité avec le domaine de la recherche

Nous considérons que le niveau de familiarité de l'usager avec le domaine de recherche pourrait être déterminé avec une précision acceptable si nous tenons compte des données suivantes:
- le niveau éducationnel de l'utilisateur;
- les constituants de la requête;
- les canaux habituels de prise d'information;
- les parties consultées du document;
- le champ disciplinaire;
- le type de document.

L'utilisation seulement des constituants de la requêtes pour déterminer ce niveau de familiarité, d'ailleurs l'approche adoptée par presque tous les systèmes documentaires, ne permet, dans le meilleur des cas, que l'évaluation du niveau de familiarité de l'usager avec les termes d'indexation (exp. le thesaurus du système documentaire), on y reviendra dans le paragraphe suivant.

Toutefois, si on entend par constituants de la requête les attributs externes (titre, auteur, ISBN, etc.), ou internes (descripteurs,...) du document, on pourrait admettre que l'utilisation des premiers, dans des circonstances bien déterminées (l'usager demande un ou plusieurs documents précis; un ou plusieurs auteurs précis), révèle la maîtrise du domaine de la recherche.

1.1.4 Niveau de précision de la requête

Nous estimons que ce niveau dépend essentiellement: des constituants de la requête, du but de la recherche d'information, du niveau éducationnel de l'usager, et du nombre de publication demandé.

Dans une opération de recherche d'information, l'utilisateur formule sa requête en utilisant soit des attributs externes aux documents soit des attributs internes. La requête est généralement une expression logique composée de termes connectés par les opérateurs logiques: et, ou et sauf. La maîtrise de la logique booléen et l'utilisation des termes précis (termes spécifiques) sont deux éléments essentiels à l'évaluation du degré de précision de la requête. Cette précision demeure relative dès lors qu'elle dépend du modèle de représentation adopté par le système et du besoin de l'utilisateur.

En effet, un haut degré de précision relatif à une évaluation du système dont la représentation du contenu est définie par un thesaurus ne peut être obtenue qu'avec une requête constituée par des termes extraits de ce thesaurus. De même pour un système dont l'indexation repose sur la Classification Décimale de Dewey. Les constituants de la requête doivent être rapporté aux indices de Dewey.

Quand il s'agit de déterminer le niveau de précision de la requête par rapport aux besoins de l'utilisateur, les constituants de la requête à eux seuls ne suffisent pas. En fait, l'utilisateur peut utiliser des termes très précis (les termes du thesaurus de système..) mais malgré ça, il n'arrive pas à bien interpréter ses besoins. Donc, pour mieux évaluer le niveau de précision de la requête, il faut inclure le profil de l'utilisateur ainsi que le but de sa recherche. Ces derniers seront des informations complémentaires aux constituants de la requête

1.2 Niveau technique

1.2.1 Types de documents

Dans le présent travail, nous ne nous intéressons qu'aux périodiques et aux monographies en tant que documents véhiculant une information textuelle. Les périodiques se présentent sous forme de numéros à contenu différent. Les monographies quant à elles ne sont produites qu'une fois, et traite généralement d'un seul sujet. Les périodiques se distinguent des monographies par:

- la rapidité de publication;
- sont un moyen de communication pour la communauté des chercheurs;
- sont propre à un domaine ou une discipline.

1.2.2 Les parties consultées du document

Le raisonnement que renferme le texte est défini par l'auteur au moment de la rédaction . Ce texte traduit en fait les opérations d'esprit qui ont conduit l'auteur de l'observation de certains faits empiriques à l'énoncé de propositions diversement nommées: thèses, hypothèses, interprétations, conclusions, explications, etc.

La diversité de ces termes suggère qu'il existe certaines différences dans la façon de concevoir la nature de l'exercice, son but, ses modalités. Mais, ceci n'empêche pas l'existence d'une propriété commune, à savoir le principe même de la connaissance scientifique, qu'il faut bien tenir pour caractéristique du savoir des spécialistes et des experts. Les textes scientifiques doivent par conséquent offrir tous, par leur construction, un reflet de cette propriété commune, quelque soit l'hétérogénéité des objets ou des phénomènes dont ils traitent, à la différence des textes que d'autres auteurs étrangers à la communauté scientifique s'avisent parfois d'écrire sur les même sujets (Gardin J.C. 86).

Cette réflexion sur les textes scientifiques, nous conduit à les considérer comme des objets construits, à des fins et par des voies particulières. L'objectif des constructions scientifiques dans des disciplines telles que la physique, la biologie, etc. est d'élaborer des représentations symboliques de faits d'observations qui paraissent manifester entre eux certaines relations, et d'agencer ces représentations sous forme de théories, systèmes ou modèle dont on puisse vérifier la valeur par des observations nouvelles. Ainsi, la structure même du texte scientifique ressemble à la logique mathématique très utilisée dans la recherche, le texte standards est presque complètement codifié est formulé dans une structure avec la même rigueur d'argumentation faite en logique formelle qui est la base même des analyses de résultats scientifiques (Kircz J. 92). Les éléments du contenu des documents sont soumis à des critères de standardisation et de régulation . Ces standards servent à la fois aux lecteurs et aux auteurs. Ils consistent en objet, méthode, résultats et discussion.

Ce bref exposé de la construction des textes scientifiques, nous conduit à admettre que ces derniers possèdent une macrostructure refermant les éléments suivants:

- introduction: énonce l'objet de la recherche en délimitant ses champs: champ d'observation et champ d'application.
- méthode: représente l'argument nécessaire au passage de l'observation de certains faits empiriques à l'énoncé des interprétations de ces faits sous formes de théories, systèmes...On distingue notamment les interprétations empirico-inductives où l'on part des données d'observation pour aboutir à la conclusion, et l'interprétation hypotico-déductives au l'on procède inversement d'une hypothèse dont on établit la valeur par des observations empiriques.
- résultats: énoncent les explications, les interprétation...
- discussion: c'est le mécanisme de validation qui permettra de confirmer ou d'infirmer les résultats.
- résumé: reflètent les éléments du contenu de document, sa structure est basée sur l'ordre le plus prévisible de ses constituants (le but, la méthode, les résultats, la conclusion).

Or, il est courant de lire qu'un document est hiérarchiquement divisé en un ensemble d'unité textuelle à différents niveaux structurels: chapitre, sous chapitre, paragraphe et sous paragraphe. Cette microstructure est utilisée pour améliorer la description du contenu de documents (Kerkouba 86, cité par Nie J. 90). A une unité d (chapitre, paragraphe, etc.) et un terme isolé t, deux mesures statistiques sont associées: la représentativité de l'unité pour le terme Rep.(d, t), qui mesure le développement du thème t dans l'unité d relativement au reste du corpus et la représentativité du terme pour l'unité Rep.(t,d), qui mesure l'importance du thème t dans l'unité d. Nous reviendrons sur cette approche d'indexation plus loin.

Certes, l'utilisation de la macrostructure apporte une certaine efficacité à l'indexation mais, les limites de cette approche sont évidentes. Elles résultent essentiellement du fait que la structure (chapitre, sous chapitre..), ne traduit pas les propriétés inhérentes de chaque type de documents (scientifiques, littéraires, etc.).

1.2.3 Les canaux de prise d'information

Dans sa quête d'information, l'utilisateur suit les canaux formels ou/et les canaux informels de diffusion. Le but de la recherche, le niveau éducationnel..., déterminent le choix du canal de prise d'information. On recense parmi les canaux formels:
- les bibliothèques scolaires;
- les bibliothèques publiques;
- les bibliothèques universitaires;

- les centres de documentation ou bibliothèques spécialisées;
- serveurs documentaires.

Les quatre premiers ont pour objectif d'offrir aux utilisateurs un accès aux documents primaires à partir de la connaissance des références de ces documents. Les serveurs documentaires quant à eux, offrent l'accès à un large éventail de banques de données renfermant généralement des informations secondaires (références bibliographiques...).

Parmi les canaux informels de prise d'information nous citons les contacts interpersonnels. Cette manière de collecte d'information est sollicitée par la communauté des chercheurs. Les canaux informels assurent: une transmission rapide des informations, la circulation d'une information déjà "digérée" par autrui, et la communication du non publié (petites astuces de savoir faire, détails de montage, procédés de calcul, etc.).

1.2.4 Le nombre de publication demandée

Le nombre de documents que l'utilisateur souhaite retrouver est un élément, parmi d'autres, que nous pouvons utiliser pour juger la précision de la requête. En effet un usager qui veut explorer des nouveaux concepts sur des sujets non connus, n'accorde généralement pas un intérêt au nombre de documents, il favorise la couverture au détriment de la précision . Dans ce cas, l'utilisateur ne fixe pas le nombre de documents, l'attribut "illimité" sera assigné au nombre de documents demandés. Dans le cas contraire l'attribut "limité" sera accordé à cet élément du niveau technique.

1.2.5 Les constituants de la requête.

Les systèmes documentaires fonctionnant sur un mode booléen traditionnel permettent la formulation des requêtes sous une forme d'expression logique composée de termes connectés par les opérateurs logiques et,ou et sauf. Ce type de formulation, s'il est acceptable par un documentaliste, est très pénible et ardu pour un utilisateur non averti. Au contraire, les systèmes en " formulation libre " permettent à l'utilisateur de poser sa requête dans les termes qui lui viennent spontanément à l'esprit. Dans les deux modes de formulation, les constituants de la requête peuvent être des termes sujets, des mots de titre, des noms d'auteurs etc.

Les termes sujets en "formulation libre" peuvent être une succession des mots (traitement automatique information), éventuellement séparés par des signes de ponctuation (analyse, données) ou reliés par des articles (représentation des connaissances).

1.3 Le but de la recherche

Nous estimons que chacun des éléments suivants: travail de recherche, activité professionnelle, études, formation..., pourrait être le but d'une recherche d'information. Ces buts se distinguent par des besoins spécifiques.

PERDERSEN cité par (Dachelet R. 90) recense trois types de besoins:

- besoins de vérifications: " l'utilisateur veut vérifier ou retrouver de l'information sur des éléments d'information aux caractéristiques connues". C'est le cas de la recherche d'information pour une activité professionnelle, l'utilisateur a besoin d'une donnée supplémentaire pour conduire sa prise de décision par exemple, retrouver des données numériques sur des valeurs d'actions boursières, la précision des recherches est alors déterminante. Ce type d'information relève plus spécifiquement des systèmes de gestion des données.

- besoins conscients concernant un sujet: " L'utilisateur veut clarifier, passer en revue ou approfondir certains aspects d'un sujet bien connu ".Ce besoin correspond bien a une activité de recherche". L'utilisateur (chercheur) sollicite une information lui permettant de construire une réflexion sur un domaine précis. par exemple, le cas du scientifique qui établit une bibliographie...

- besoins flous concernant un sujet: " l'utilisateur veut explorer de nouveaux concepts sur des sujets non connus ". C'est le cas des recherches à but d'étude ou de formation. par exemple, le dossier bibliographique constitué par un étudiant en début de recherche. Dans cette phase initiale, l'étudiant à besoin de l'aide du système documentaire. A l'heure actuelle l'intermédiaire en information (documentaliste, etc.) assure cette aide.

1.4 Les attributs des éléments du modèle utilisateur

1.4.1 Les attributs des éléments du niveau cognitif

Elémeents	Attributs
a) Niveau éducationnel	- primaire - secondaire - premier cycle universitaire - deuxième cycle unvirsitaire - troisième cycle universitaire - recherches
b) Champs disciplinaires	- siences et techniques - sciences sociales - sciences humaines
c) Niveau de familiarité avec le domaine de la recherche	- Bon, Moyen, Mauvais
d) niveau de précision de la requête	- Bon, Moyen, Mauvais

1.4.2 Les attributs du but de recherche

Eléments	Attributs
But de la recherche	- activité de recherche - activité professionnelle - études - formation - autres

1.4.3 Les attributs des éléments du niveau technique

Eléments	Attributs
a) Types de documents	- monographies - périodiques
b) Canaux de prise d'information	- bibliothèques universitaires - bibliothèques scolaires - bibliothèques publiques - centres de documentation ou bibliothèques spécialisées - serveurs - contacts interpersonnels
c) Le nombre des publications demandées	- limité - illimité
d) Les constituants de la requête	- auteur - titre - sujet: descipteurs, ou langue naturelle
e) Les parties consultées	- résumé - introduction - résultats - discution - références bibliographiques - le documents entier

1.5 Détermination du profil d'un utilisateur

La diversité des utilisateurs, les facettes évolutives d'un même individu (les besoins en phase initiale du travail ne sont pas identiques à ceux en phase finale) et la variété des buts de recherche se combinent en une complexité de données contradictoires. Il n'est pas donc surprenant qu'a partir des éléments invoqués pour décrire notre "modèle utilisateur ", on n'arrive pas à décrire le profil d'un nombre d'utilisateurs.

Notre approche de détermination du profil d'un utilisateur consiste à assigner à chaque élément du "modèle" un ou des attributs bien déterminés. Ces attributs peuvent être des données que l'utilisateur fournit au système ou une évaluation à partir des attributs fournis (exp. les attributs bon, moyen et mauvais assignés après évaluation au niveau de familiarité avec le domaine).

Nous estimons que la prise en compte de la catégorie socio-professionnelle (les attributs du but de la recherche, les attributs du niveau éducationnel)et du type de connaissances (niveau de familiarité avec le domaine de la recherche), nous a permis de dépasser la classification générique des utilisateurs (les lecteurs avertis, les non lecteurs, les lecteurs partiellement avertis et les lecteurs avertis). Toutefois, il est à signaler que le modèle utilisateur élaboré dans le cadre de ce travail ne peut être considéré comme un modèle conceptuel traduisant la représentation mentale que l'utilisateur a du processus de recherche d'information. Cette représentation détermine le comportement de l'usager. Ce comportement n'est que l'effort investi par l'usager pour s'adapter aux conditions de la tâche (recherche d'information) qu'il s'est fixé.

1.5.1 Les données fournies par l'utilisateur

Dans une opération de recherche, l'utilisateur doit fournir au système les données suivantes:
- le but de la recherche d'information
- le niveau éducationnel
- le champ disciplinaire
- les canaux habituels de prise d'information
- types de documents
- les parties consultée du document
- les constituants de la requête

1.5.2 Evaluation du niveau de familiarité avec le domaine

Nous avons vu au paragraphe 1.1.3 que le niveau de familiarité avec le domaine de la recherche dépend essentiellement des éléments suivants:
- le niveau éducationnel
- le champ disciplinaire
- les canaux habituels de prise d'information
- types de documents

- les parties consultées du document
- les constituants de la requête.

Pour illustrer cette évaluation du niveau de familiarité avec le domaine, nous utilisons deux exemples différents.

Exemple (1):
- niveau éducationnel: recherche
- champ disciplinaire: science
- types de documents: périodiques
- les parties consultées: le document entier
- les canaux de prise d'information: bib. universitaire; serveurs; contacts personnels
- les constituants de la requête: nom d'auteur

Exemple (2):
- niveau éducationnel: recherche
- champ disciplinaire: sciences
- types de documents: monographies
- les parties consultées: le document entier
- les canaux de prise d'information: bib. universitaire, serveurs
- les constituants de la requête: sujet

a) interprétation de l'évaluation:

On constate que les attributs de trois éléments du niveau technique sont différents:

(1) périodique # monographie
(2) auteur # sujet
(3) bib. univ.; serveurs; contacts per. # bib. uni.; serveurs

(1) En raison de la précarité de l'information dans le secteur scientifique, le facteur temps étant particulièrement important, les chercheurs vont surtout faire appel aux articles de périodiques et plus rarement aux monographies. Sans doute parce qu'ils représentent une source généralement plus à jours. Une étude de (Meyriat J.1984) montre que dans des travaux scientifiques, les articles sont utilisés quatre fois plus que les monographies (Partie I. 2.1). Cette relation entre le champ disciplinaire et le type de document, nous permet d'accorder, dans le cas du "secteur scientifique", à l'attribut périodique un "poids" supérieur à celui de l'attribut monographie.

(2) Un chercheur connaît généralement la communauté scientifique de son domaine. De ce fait, il ne trouve pas de difficulté à localiser un article, un ouvrage...Les noms

des auteurs sont généralement connus (Partie I 4). Toutefois, rien n'empêche ce chercheur a utilisé la recherche par mots clés. Dans ce cas, deux interprétations sont possibles: la première le chercheur veut rendre sa recherche plus exhaustive(risque des noms inconnus), la seconde, le chercheur est en phase initiale de recherche et les noms des auteurs sont encore inconnus pour lui.

De ce qui précède , nous estimons que la recherche par auteur demeure une recherche précise (voir introduction). Cette précision devient presque certaine si l'utilisateur a le niveau éducationnel "recherche".

(3) Les enquêtes de DEMAILLY 1978 - MENZEL 1968 (PartieI 2.4) ont montré que les chercheurs préfèrent les contacts interpersonnels dans leurs quêtes d'information. Ce phénomène se traduit par le fait que ces canaux de prise d'information permettent: une transmission rapide des informations (sans attendre leur publication) et l'établissement de relations sélectives entre les chercheurs. Cette dernière caractéristique prouve que la recherche par contacts interpersonnels demeure réservée aux chercheurs bien intégrés dans la communauté scientifique. De ce fait, nous admettons que la mention par l'utilisateur de l'attribut "contacts personnels" pourrait être retenu comme un indice de bon niveau de familiarité avec le domaine de la recherche.

b) Résultat de l'évaluation:
A partir de (1), (2) et (3), nous pouvons conclure que "l'utilisateur 1" a un bon niveau de familiarité avec le domaine, et que "l'utilisateur 2" a un niveau moyen de familiarité avec le domaine.

c) Discussion
L'utilisation de la typologie présentée au paragraphe (3) de la première partie, ne permet pas de révéler la nuance existant entre les deux profils (utilisateur 1 et utilisateur 2). En effet, cette typologie est basée sur des éléments tels que: le niveau éducationnel, le champ disciplinaire et le type de document, pour déterminer les différentes catégories des lecteurs. Or, il s'est avéré que le problème est encore plus complexe d'ailleurs, en invoquant deux autres éléments (les canaux habituels de prise d'information et les constituants de la requête) nous avons pu aboutir à une typologie plus affinée.

Une étude encore plus poussée pourra invoquer d'autres éléments, et présentera par conséquent une typologie encore plus affinée que la notre. C'est pour cette raison

nous avons affirmé au paragraphe 1.5 qu'il n'est pas surprenant qu'à partir des éléments invoqués on n'arrivera pas à déduire le profil d'un nombre d'utilisateurs.

1.5.3 Evaluation du niveau de précision de la requête

Au paragraphe 1.2.5 nous avons essayé de déterminer les éléments intervenant dans l'évaluation du niveau de précision de la requête. On a recensé:
- le but de la recherche d'information
- le niveau éducationnel
- le nombre de publication demandée
- les constituant de la requête

Pour illustrer cette évaluation du niveau de précision de la requête, nous maintenons les exemples 1et 2 traités au paragraphe précédent. Ainsi on a les attributs suivants:

	Exemple (1)	Exemple(2)
- but de la recherche	travail de recherche	travail de recherche
- niveau éducationnel	recherche	recherche
- le nombre de publication	limité	illimité
- les constituants de la requête	auteur	sujet

a) Interprétation

On constate que deux éléments du niveau technique diffèrent par leurs attributs:

(1) auteur # sujet

(2) limité # illimité

(1) voir explication paragraphe précédent.

(2) L'indication par l'utilisateur du nombre de documents qu'il souhaite retrouver, pourrait refléter la précision de sa requête. Cette précision devient presque certaine quand l'utilisateur effectue une recherche par auteur. De ce fait, on peut considérer que le niveau de précision de la requête est bon pour l'utilisateur(1). L'utilisateur (2) quant à lui, il a accumulé deux imprécisions, l'une relative à la recherche par sujet (voir introduction et partie 2) l'autre à l'attribut "illimité" (voir 1.2.4). Mais, le fait qu'il a un niveau éducationnel recherche, cela nous permet d'admettre qu'il pourrait investir ses connaissances pour pallier à ces imprécisions, et par conséquent on accorde l'attribut moyen au niveau de précision de sa requête.

1.5.4 Conclusion

Il est courant de lire que pour réussir une activité de recherche d'information, il est nécessaire de repartir du besoin de l'utilisateur. Mais concrètement, en quoi consiste ce besoin?

PERDERSEN (voir 1.3) a essayé d'établir une typologie des besoins en besoins de vérifications, besoins conscients concernant un sujet et besoins flous concernant un sujet. Cette typologie à caractère trop générique ne permet pas de cerner tous les besoins possibles des utilisateurs. Entre les trois types des besoins, on trouve toute les nuances. Il est donc intéressant de se demander qu'elle est l'opportunité d'une telle typologie pour une notion aussi complexe que les besoins. Cette complexité résulte essentiellement du caractère tout à fait immatériel des besoins.

Ce caractère immatériel du besoin a conduit les chercheurs à focaliser leurs études sur le comportement de l'utilisateur. Ce comportement n'est autre que les manifestations observables de l'interaction avec l'environnement. De ce fait, l'analyse du comportement est une détermination des éléments ou des paramètres décrivant ces manifestations.

Notre approche dans la modélisation de l'utilisateur était de recenser le maximum des éléments qui pourraient refléter les besoins de l'usager. Le modèle élaboré, nous ne pouvons en aucuns cas le considérer comme un "modèle cognitif". Ce dernier établit des liens complexes entre les différents éléments décrivants le comportement de l'utilisateur. En évaluant le niveau de familiarité avec le domaine de la recherche et le niveau de précision de la requête, nous avons essayé d'établir des liens entre les différents éléments (exp. entre le niveau éducationnel et les constituants de la requête ..) mais , ces liens restent rudimentaires comparés à ceux du modèle cognitif. Toutefois, nous estimons que la résolution de l'éternel problème de satisfaction de l'usager auquel sont confrontés les différents systèmes documentaires (bibliothèques, serveurs,...) nécessite une évaluation bien fondée du niveau de familiarité de l'usager avec le domaine de recherche autrement dit, il faut arriver à constituer une idée sur la représentation mentale que l'utilisateur a sur son sujet de recherche. Ainsi, on arrive à bien évaluer le degré de précision de la requête non par apport au modèle de représentation du système (thesaurus) mais par rapport aux besoins réels de l'utilisateur.

2) Description de documents textuels

En matière de description de documents textuels deux approches d'indexation se présentent. La première correspond au règne absolu du texte (tentative d'une saisie de la stabilité sémantique). Elle part du postulat que chaque texte contient les informations utiles à sa compréhension. Cette approche est profondément enracinée dans la tradition documentaire, malgré ses imperfections, elle continue à dominer presque tous les systèmes documentaires, et un changement radical n'est guère envisageable actuellement. La seconde approche pose l'investissement du lecteur (compétence pragmatique, contexte ...), elle part de l'hypothèse que la lecture est un processus actif où le lecteur crée lui même le sens du texte, il mobilise pour cela à la fois le texte, ses propres connaissances et les intentions de lecture. Les recherches en pragmatique ont approuvé cette approche, elles considèrent que les ambiguïtés rencontrées dans un texte ne peuvent être levées part l'utilisation de la grammaire et du sens des mots.

Pour les spécialistes d'information (bibliothécaire , ...) dont la tâche consiste précisément à représenter de façon cohérente le contenu des documents, le choix entre les deux approches appraitra problématique. En effet, la première approche est opérationnelle mais ses imperfections sont nombreuses. Le problème de l'inadéquation habituellement constaté entre la demande exprimée par le chercheur et les ressources décrites par le système reflète ces imperfections. La seconde approche quant à elle , elle est au stade de la recherche théorique. A notre connaissance aucun système documentaire n'est actuellement opérationnel avec cette approche. Certes la seconde approche se base sur une théorie séduisante mais, l'espace qui la sépare de la pratique est grande. Il s'agit d'une activité profondément humaine (processus de lecture...) que la théorie doit formaliser.

Au paragraphe précédent nous avons essayé d'établir un modèle utilisateur mais il s'est avéré que le travail en ce domaine est très complexe, pour cela nous étions obligé de nous contenter de citer quelques éléments pouvant être utilisés dans l'explication du comportement de l'usager. Néanmoins nous avons conclu que la réussite de l'activité documentaire dépend largement des résultats des recherches en matière de modélisation de l'utilisateur . Cette conclusion est en corrélation parfaite avec le principe de communication qui présuppose un choix adéquat de la représentation (description) en fonction d'un utilisateur (récepteur), et en fonction du canal de communication (système documentaire) , nous y reviendrons plus loin. Donc, nous estimons que le choix de la seconde approche (description en fonction de

l'utilisateur) est légitime mais étant donné la délicatesse de son application, la solution pourrait être envisagée dans le choix d'une position intermédiaire entre les deux approches

Au paragraphe (2.2), nous essaierons d'expliciter les fondements de la seconde approche de description, et de déterminer les éléments qui pouvaient nous servir à justifier un éventuel choix d'une position intermédiaire.

2.1 Les limites des modèles classiques de description

Dans la deuxième partie nous avons distingué entre l'indexation automatique et l'indexation manuelle . Cette distinction était nécessaire pour déterminer l'apport de chaque approche en matière de description de document. Les résultats auxquels nous avons abouti ne montrent en rien qu'il y avait une telle distinction. En effet, les approches automatiques d'indexation ne sont qu'une tentative d'imitation de l'approche usuelle. Le principe de règne de texte (tentative de saisir la stabilité sémantique du texte) sur lequel se base l'approche d'indexation manuelle, est parfaitement respecté et même accentué par les différentes méthodes d'indexation automatiques. Donc, on peut conclure que c'est uniquement au niveau de la technique de réalisation de la tâche d'indexation qu'il y avait une différence. D'ailleurs, le passage même de l'indexation manuelle à l'indexation automatique s'est imposé suite à des contraintes purement économiques (les coûts d'indexation manuelle trop élevé) et non dans un but d'améliorer la représentation du contenu.

Dans ce qui suit nous resterons dans le cadre de l'approche traditionnelle de description de documents textuels (tentative de saisir la stabilité sémantique du texte) et nous essaierons de présenter en bref les limites de chaque méthode d'indexation.

- Indexation manuelle

Malgré la pratique d'une indexation contrôlée dans laquelle les termes d'indexation sont sélectionnés à partir d'une liste préétablie (liste d'autorité, thesaurus), on trouve le même texte représenté par des termes différents. La question de comment décider que tel ou tel descripteur représente bien le contenu d'un document demeure donc à jour. La multiplicité des termes de représentation, n'est pas l'unique critique apportée à l'indexation manuelle. En effet , le choix même du mot (catégorie "nom" et "adjectif") comme unité principale de représentation du contenu de document s'est avéré problématique . Le mot seul en tant qu'élément du lexique ne peut pas , considéré isolément, faire référence à un objet de la réalité extra- linguistique de

l'auteur du document. Il ne peut pas exprimer " ce dont parle le document ". Il ne peut donc pas être descripteur (Bouché R. 89).

- méthodes statistiques

Nous essaierons de présenter les critiques apportées à la méthode de Salton. Cette dernière est considérée comme étant la méthode statistique la plus développée. Le choix des termes d'indexation s'effectue non seulement sur la base de la fréquence du terme dans le document mais aussi en tenant compte du nombre de documents du système représenté par le même terme. Un terme rare, décrivant peu de documents sera privilégié par rapport à un terme général se trouvant dans un grand nombre de documents. L'indice de pertinence du terme est donc fonction du nombre de document du système. L'évolution du fonds documentaire conduit alors a un changement de l'indice de pertinence ce qui nécessite, une répétition de l'ensemble des calculs à chaque introduction de nouveaux documents. Ces calculs deviennent lourds à gérer quant il s'agit d'un système documentaire à plusieurs centaines de milliers de documents.

- Approche linguistique

Cette approche d'indexation recense pratiquement tous les niveaux d'analyses (voir partie 2) permettant la compréhension du sens du texte. L'analyse sémantique utilise un réseau dont les noeuds représentent des concepts et les arcs illustrent les relations sémantiques entre les concepts. Les limites de ce niveau d'analyse résultent de la conception du réseau sémantique qui ne tient pas compte de l'évolution du domaine couvert. L'analyse pragmatique fait intervenir des scripts, des règles d'inférence (voir partie 2), elle recherche à modifier le résultat de l'analyse sémantique par la prise en considération d'éléments extérieurs au texte. Les recherches en pragmatique ne sont pas encore achevées, et la discipline n'est qu'à ses débuts il serait donc prématuré de juger ses résultats. Toutefois, nous estimons que l'avenir du document de la documentation dépend largement des résultats de recherche en pragmatique, nous y reviendrons plus loin.

On a voulu par ce bref rappel des limites de méthodes d'indexation montrer à quel point le domaine de la description du contenu de texte est complexe. Chaque méthode a ses principes par exemples, dans les méthodes statistiques les résultats ne peuvent être considérés comme donnant une bonne approximation des phénomènes analysés que si les fréquences observées sont suffisamment importantes. Cette diversité dans les principes conduit à des descriptions de qualité " différentes ".

Toutefois, nous estimons que les jugements apportés à la représentation doivent tenir compte des limites de l'approche de description (tentative de saisie de la stabilité sémantique). En fait, un indexeur qui décrit le contenu en utilisant uniquement le " sens" véhiculé par les mots du texte n'arrive généralement pas à traduire convenablement le message texte. Cette description insuffisante du contenu n'est pas dû aux outils de représentation (mots clés ou ensemble de termes structurés) mais, à l'approche de la lecture adoptée par l'indexeur.

2.2 Description en fonction du besoin de l'usager

Le but ultime de la description du contenu de document est d'arriver à une représentation adéquate à la recherche d'information. Cette représentation consiste à associer au document soit un ensemble amorphe de "mots clés" ou descripteurs" soit un ensemble structuré de "terme" (syntagmes nominaux. ...) issus ou non d'un lexique établi à priori et plus ou moins organisé.

(GARDIN J. C. 72) définit l'analyse documentaire (description de documents textuels) comme étant l'extraction du sens des documents. Le sens implique une référence à quelque chose que l'on sort" du texte et que l'on désigne par des symboles (descripteurs...) qui ne se trouve pas nécessairement dans celui ci .

Gardin dans sa définition admet que le sens du texte ne peut être extrait qu'à partir des mots de celui ci. Cette approche de description de contenu à dominer presque tous les systèmes documentaires existant (voir paragraphe précédent), et ce n'est que récemment, avec les recherches en pragmatique, qu'une nouvelle approche de description de documents a pu émerger. Avant de décrire cette approche, il serait nécessaire de présenter en bref l'évolution en matière de description qui a précédé son émergence.

Deux axes de recherches ont dominé la description de document textuel telle qu'elle a été définie par Gardin: Le premier axe, est une tentative de saisir la sémantique du texte par des méthodes quantitatives (méthodes statistiques), elle consiste à détecter des mots ou des phrases supposées plus " significatifs" que d'autre dans un corpus donné au moyen de divers indices numériques. De nombreux critiques ont été apporté à la méthode staztistique. A part quelques travaux intéressant tels que ce de Salton, cette méthode n'a pas connu une évolution considérable malgré ses débuts prometteurs.

La seconde axe de recherche est une description du contenu de document (extraction du sens) à partir de sa construction grammaticale, elle consiste à déterminer la nature des relations syntaxique entre les termes. Ces relations se rapportent à des rapports de causalité, de classification, de but, de localisation,.... , elles traduisent donc une certaine logique nécessaire à la construction du sens de certains textes. Mais , le fait qu'aucune grammaire ne soit encore en mesure de proposer un inventaire raisonnablement complet des différentes façons d'exprimer dans une langue quelconque par exemple un lien de causal entre un antécédent X et un conséquent Y, à conduit les analystes à ignorer cette partie de la théorie grammaticale qui pouvait fournir une certaine rationalisation du processus (Gardin J. C. 74).

Ce point de vu de GARDIN sur les limites de la grammaire, se confirme par les représentations des faits empiriques dans les constructions scientifiques (voir 1.2.2). En effet, faute de cette limite, le langage de représentation dans les constructions scientifique est différent du langage naturel: les représentation de la nature telles que nous les connaissons aujourd'hui en physique, en chimie, en biologie, n'ont plus grande chose de commun avec les descriptions que nous pourrions en donner avec les seuls moyens de nos langue maternelle, aussi "avancée" soient elles.

A ce stade d'étude nous pouvons admettre que la description de documents basée sur la seul analyse syntaxique (approche linguistique)ne peut aboutir à une représentation adéquate à la recherche d'information. Mais, les défendeurs de l'approche linguistique en matière de description de document voient les choses autrement: comment se peut il qu'un exposé des règles qui permettent à une certaine personne de comprendre une langue donnée, n'ait aucun rapport avec la faculté qu'a cette personne de dégager le sens des textes écrits dans la même langue. Cette réflexion est claire dans ses objectifs, il s'agit d'accorder à l'analyse syntaxique, considérée comme une étape nécessaire de toute description de document, une priorité absolue en d'autres termes, l'idée est d'opérer essentiellement du texte vers le sens ou encore du symbole S vers l'interprétation profondes défini sur le plan syntaxique (Chomsky cité par Gardin 74).

Nous avons voulu, en citant en bref les deux axes de recherches en matière de description de document (méthodes statistiques et méthodes linguistiques), montrer à quel point la notion du "sens" du texte est liée aux mots de celui ci. La quasi-totalité des travaux effectués en domaine de description de documents n'a pas envisagé des améliorations de la représentation du contenu en dehors de cette conception du sens (stabilité sémantique du texte) . De ce fait, notre travail a une certaine originalité dès

lors qu'il traite la description de document non seulement à partir des mots du texte mais en tenant compte aussi du lecteur en tant qu'utilisateur final du document. Toutefois, nous ne prétendons pas qu'à ce stade de recherche (le présent travail), nous pouvions présenter une nouvelle approche de description de document qui substituera une approche parfaitement enracinée dans les traditions documentaires. Le fait que nous avons soulevé d'une part les problèmes relatifs aux limites de l'approche de description existante et d'autre part les problèmes relatifs à l'absence d'un modèle utilisateur permettant de cerner les besoins de celui ci, fait que nous estimons qu'un pas a été franchi mais le travail est loin d'être achevé.

Dans ce qui suit , nous nous attacherons à justifier cette prise en compte de l'utilisateur en matière de description de document . Pour cela, il serait nécessaire de décrire quelques concepts clés de la communication : les connaissances , la compréhension, la pertinence.

2.2.1 La notion de communication.

Comprendre comment s'effectue la communication ou encore mieux comprendre le modèle de la communication serait intéressant à notre sujet si, nous considérons que la recherche d'information est une forme de cette communication.

Dans ce paragraphe nous essaierons d'apporter une réponse à la question "comment les humains communiquent-ils entre eux". Certes, la réponse ne sera guère exhaustive mais, l'essentiel à notre niveau de recherche est de connaître les éléments de cette communication. Ces éléments vont nous permettre de déduire le rôle de l'utilisateur dans la communication écrite (recherche d'information).

Pour apporter une repose à notre question, il serait intéressant de commencer par un bref exposé des différentes tentatives d'élaboration d'une théorie générale de la communication.

Selon DAN SPERBER (89):" tout le monde semble considérer qu'une théorie générale de la communication est possible et nécessaire. D'Aristote aux sémioticiens modernes, toutes les théories de la communication ont été fondées sur un seul et même modèle , que nous appellerons le modèle du code. Selon ce modèle, communiquer, c'est coder des messages. Récemment, plusieurs philosophes , dont Paul Grice et David Lewis, ont proposé un modèle tout à fait différent, que nous appellerons le modèle inférentiel, communiquer, c'est produire et interpréter des

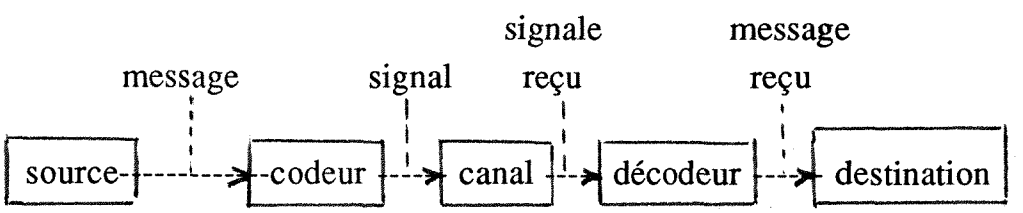
indices. Ces deux modèles ne sont pas incompatibles entre eux; on peut les combiner de différentes manières. Au cours des vingt dernières années , les travaux des pragmaticiens, des philosophes du langue et des psycho- linguistes ont montré que la communication verbale met en jeu à la fois un processus de codage et des processus inférentiels".

Cette classification des différentes théories de la communication en trois catégories (modèle de code, modèle inferentiel, combinaison des deux modèles) est intéressante par rapport à notre propos . En effet nous allons essayer de décrire la description de documents par rapport aux différents modèles de la communication. En d'autres termes, nous déterminerons le modèle qui explicite mieux le processus de la communication.

- Le modèle du code

Un code est un système qui associe des signaux et qui permet à deux dispositifs de traitement de l'information (des machines ou des organismes) de communiquer. Un message est une représentation interne à l'un des dispositifs. Un signal est une modification de l'environnement qui peut être produite par un des deux dispositifs et détectée par l'autre.

Le diagramme de Shannon et de Weaver (1948) montre comment l'utilisation d'un code rend la communication possible



Dans la communication écrite la source et la destination sont les mécanismes cognitifs centraux de l'auteur et du lecteur; le codeur et le décodeur sont constitués par leurs capacités linguistiques; le message est une pensée de l'auteur et le canal est le document.

Selon ce modèle la communication écrite n'est autre qu'un mécanisme de codage et de décodage. Le lecteur (récepteur) investit ses compétences linguistiques (décodeur) pour comprendre le message voulu par l'auteur. En d'autres termes, il suffit de lier les deux compétences linguistiques par un canal (document) pour que le

message voulu par l'auteur arrive au lecteur. Cette conception "technique" de la communication réduit la compréhension du texte au simple décodage d'un signale linguistique . En termes grossièrement simplifiés, il suffit d'une grammaire qui associe des symboles (des mots) à des représentations sémantiques pour que le message passe entre l'auteur et le lecteur.

Avec cette théorie de la communication fondée sur le modèle du code, on rejoint parfaitement l'approche de description qui tente de saisir la stabilité sémantique du texte. Le lecteur et l'auteur ont besoins d'un code commun (compétence linguistique commune)pour communiquer. Cette notion de code commun rend difficile la coexistence d'interprétations différentes.

- Le modèle inférentiel

L'hypothèse de base de ce modèle de communication est que la représentation sémantique construite à partir d'une structure syntaxique est loin de coïncider avec les pensées qui peuvent être communiquées en énonçant cette phrase. On passe de la représentation sémantique à la pensée communiquée non par un surcroît de codage (structure syntaxique) mais au moyen d'inférences.

Selon Dan Sperber (89) la représentation sémantique qu'une grammaire assigne à une phrase ne tient pas compte des propriétés extra-linguistiques des énoncés de cette phrase, par exemple, le moment et le lieu de l'énonciation, l'identité du locuteur, ou ses intentions . La représentation sémantique d'une phrase correspond en quelque sorte au noyau de sens qui est commun à tous les énoncés de la phrase en question. Mais, diffèrents énoncés d'une même phrase peuvent avoir - et, en général, ont - différentes interprétations.

Dan Sperber à bien distingué entre la représentation qui relève de la grammaire et l'étude de l'interprétation des énoncés relevant de ce qui est appelé aujourd'hui la pragmatique (voir partie 2)

Lorsqu'on travaille sur la description de documents textuels, la distinction entre la représentation sémantique et l'interprétation à l'avantage d'apporter une explication aux limites de l'approche traditionnelle de description qui parte du texte pour définir le sens de celui ci. En effet, cette approche que nous pouvons assimiler au modèle de code opère sur un niveau explicite (les phrases du texte) . Or, il s'est avéré d'après le modèle inférentiel que la plus part des informations transmises par le discours (~ texte) en langage naturel sont implicites.

Le modèle inférentiel rend caduc l'idée même du sens dès lors que celui ci est saisi non pas lorsque le lecteur reconnaisse le sens linguistique de l'énoncé mais lorsqu'il en infère le "vouloir dire" de l'auteur. De ce fait , nous ne pouvons pas parler d'une description adéquate, quel que soit la représentation adoptée, si celle ci n'a pas pris en compte la capacité inférentielle du lecteur.

Cette conclusion à la quelle nous avons abouti justifie bien notre hypothèse de travail qui consiste à admettre que la description de document ne peut réussir sans tenir compte des caractéristiques de l'usager. En d'autres termes, les capacités inferentielles du lecteur doivent être bien cerné. En quoi consiste cette capacité inférentielle? C'est la question à la quelle nous nous attacherons à répondre dans ce qui suit.

2.2.1.1 La compréhension

Le fonctionnement du système cognitif repose sur la mémoire, qui est un processus résultant du fonctionnement physiologique du cerveau et qui conserve les unités d'information qui sont élaborées à partir des stimuli d'entrée reçue par les organes des sens. Les élaborations sont faites dès l'activation des organes sensoriels puis essentiellement par l'activité de la mémoire elle même en tant que support de processus de traitement (BISSERET A. 92)

Les connaissances élémentaires (information traitée) sont organisées en des ensembles complexes. Deux structures d'organisations sont particulièrement importantes à mentionner: les réseaux sémantiques et les schémas.

- Les réseaux sémantiques

Nombres de nos connaissances s'organisent entre elle comme des graphes comportant des noeuds reliés par des liens. Les noeuds représentent les objets et concepts d'un domaine de connaissances donné et les liens représentent les relations qui existent entre les noeuds

- Les schémas

De nombreux résultats expérimentaux montrent qu'à partir d'expérience dans des situations similaires, nous formons en mémoire, des ensembles de connaissances coordonnées et typiques pour chaque situation familière. Ces ensembles de connaissances typiques sont formalisés par des schémas qui sont des structures plus vastes que les propositions sémantiques mais construites de la même façon. Les

schémas peuvent comprendre des scriptes (voir partie 2) qui sont des structures définies comme des schémas mais qui organisent les connaissances relatives aux actions

A partir de ces éléments du cadre théoriques du fonctionnement cognitif nous pouvons établir une idée sur le processus de compréhension du langage.

Dans ce cas les données de la situation se présentent sous forme symbolique (les mots de texte) L'activité de l'individu est entièrement orientée vers la construction d'une représentation de ce dont parle le texte et non pas du texte lui même . Dans la grande majorité des situations, la représentation du texte, du vocabulaire, de la syntaxe joue un rôle d'outil intermédiaire, certes indispensable, mais vite oublié, tandis que la représentation de la situation évoquée grâce au texte , elle, est installée en mémoire. Ainsi la représentation finale construite par un individu à partir , une grande part provient de ses connaissances préalables qui certes ont été activé par la lecture du texte mais qui n'étaient pas dans le texte (BISSERET A. 92).

Par ce beve exposée de processus de la compréhension du langage, on rejoint parfaitement la théorie de la communication qui combine les deux modèles (le modèle du code et le modèle inférentiel). Les connaissances linguistiques seront utilisées dans un but de codage ou décodage du texte lui même, les connaissances pragmatiques, quant à elles, serviront à la construction d'une représentation de ce dont parle le texte (les informations implicites).

2.2.1.2 Les connaissances pragmatiques

Nous pourrions penser que les connaissances pragmatiques ne sont nécessaires que parce que nous nous exprimons mal, que nous laissons trop ambiguïté dans le texte. Mais, en réalité l'ambiguïté est une conséquence négative d'une qualité fondamentale des langues naturelle: la concision. Cette dernière n'est possible qu'entre Deux individus qui ont énormément de choses en commun. Par exemple, les connaissances professionnelles partagées permettent aux individus d'uniformiser le vocabulaire, la syntaxe et la structure de leur langage. Cette constatation a conduit les chercheurs à admettre l'existence d'un sous système linguistique de communication. Ce système a été baptisé langages de spécialité.

Certes, les connaissances professionnelles partagées facilitent la communication intra-communautaire mais, elles ne sont suffisantes pour enlever l'ensemble des ambiguïtés

véhiculées par le langage naturel. En effet, d'autres connaissances doivent être mises en oeuvre à fin d'écarter le reste des ambiguïtés .

Selon (Pitrat J. 85): "si nous voulons traiter des textes de n'importe qu'elle culture, nous serons obligés d'indiquer les limites d'utilisation de certaines connaissances: à telle époque, à tel endroit les gens pensaient telle chose... Par exemple , quand nous lisons un roman étranger , le traducteur doit rajouter des notes pour nous fournir les connaissances que tout concitoyen de l'auteur possède et qui sont indispensables pour comprendre l'histoire ".

De ce qui précède nous admettons que le processus de communication nécessite des connaissances partagées. De ce fait, il serait interéssant de savoir à quel point les humains partagent- ils des connaissances? Tous les humaines vivent dans le même monde physique. Durant leur vie, ils s'efforcent de tirer de l'information de cette environnement commun et d'en construire une représentation mentale (connaissances organisées) aussi bonne que possible. Ils ne construisent pas tous la même représentation, d'une part, à cause de différences dans leurs environnements physiques locaux et, d'autre part, à cause de différences dans leurs capacités cognitives. L'efficacité des capacités perceptives varie d'un individu à l'autre. Les capacités inférentielles varient aussi. Les humaines apprennent des langes différentes, ils maîtrisent des concepts différents; en conséquence, les représentations qu'ils peuvent construire et les inférences qu'ils peuvent effectuer diffèrent. Ils ont aussi des souvenirs différentes, et des théories différentes qu'ils appliquent à leurs expériences de manière variées. Donc, même si les humains partageaient tous le même environnement physique, ce que nous proposons d'appeler leur environnement cognitif serait néanmoins différent (Dan Sperber 89)

Ce point de vue Sperber sur les capacités inférentielles des humaines est très intéressant à notre propos, il illustre à quel point la description des documents en fonction des besoins d'utilisateur est difficile. En effet, si notre objectif est d'arriver à un représentation de la connaissance véhiculée par un texte. Cette représentation ne peut être identique à celle de l'utilisateur que dans le cas ou en tient compte de ce que Sperbere l'appeler " environnement cognitif ". Certes, c'est l'approche de description idéale, elle nous permettra d'éviter le problème de l'inadéquation habituellement constaté entre la de mande exprimée par le chercheur et les ressources décrites par le système. Mais, son application pose d'énorme problèmes. En d'autres termes en doit concevoir un modèle cognitif de l'utilisateur.

2.2.1.3 Relation pertinence connaissances

Un être humain est un dispositif efficace de traitement d'information. Par ce traitement, il vise à améliorer ses connaissances du monde autant que le lui permettent les ressources dont il dispose. Améliorer la connaissance, cela veut dire acquérir d'avantage d'informations.

L'être humain s'attache généralement à traiter trois types d'informations:
- Les informations anciennes: elles sont déjà contenues dans la représentation du monde de l'individu et de ce faite elles ne nécessitent pas un nouveau traitement.
- les informations nouvelles qui n'ont aucun rapport avec quoi que ce soit dans la représentation que l'individu a du monde. Ces informations ne peuvent s'ajouter à cette représentation que de manière isolée, ce qui entraîne en général un coût de traitement trop grand pour un bénéfice trop faible.
- les informations nouvelles qui sont en rapport avec des informations plus anciennes. Ces informations en rapport entre elles sont utilisées conjointement en tant que prémisses dans un processus d'inférence, elles engendrent d'autres informations nouvelles: des informations qui n'aurait pu être inférées sans cette combinaison de prémisses anciennes et nouvelles. Quand le traitement d'informations nouvelles donne lieu à un tel effet de multiplication, nous disons que ces informations sont pertinentes. plus l'effet de multiplication est grand, plus grande est la pertinence.

De ce qui précède nous admettons que l'individu possède non seulement le savoir présenté dans son esprit mais aussi le savoir qu'il est capable de le déduire du savoir dont il a la représentation. Cette affirmation de la nécessité d'un savoir accumulé pour inférer de nouveaux savoirs ne doit être pris dan l'absolu car, le problème et encore plus complexe. En fait, une hypothèse peut être manifeste sans être mentalement représentée. Pour DAN SPERBER cela découle du fait que les hypothèses sont différentes du savoir en ce qu'elles ne sont pas nécessairement vraies. Par exemple, l'hypothèse selon la quelle la communication verbale est un processus de codage et de décodage, ne reflète en aucun cas la réalité de la communication autrement dit elle n'est pas tout à fait vrais . pourtant , tellement elle est enracinée dans la pensée humaine que les gens ont oublié qu'il s'agit avant tout d'une hypothèse et non d'un savoir vrai.

le phénomène cognitif est encore plus compliqué et nous n'avons présenté qu'un noeud d'un réseau complexe dont il est nécessaire de découvrir ses différents noeuds

et les liens qui entreprennent entre eux. Toutefois, nous estimons que les quelques données recueillies, vont nous permettre de conclure cette partie de notre travail en présentant les grandes lignes de ce que nous appelons la description de documents en fonction des besoins de l'utilisateur.

2.3 Solution envisagée

De ce qui précède, nous pouvons admettre que dans une opération de recherche trois cas se présentent:
1) Si les informations apportées par une description de document sont en rapport avec les connaissances de l'utilisateur; ce dernier peut juger la pertinence ou non du document pour sa recherche.
2) Si les informations apportées par une description de document sont en rapport partiel avec les connaissances de l'utilisateur: il serait difficile à ce dernier de se décider sur la pertinence ou non du document pour sa recherche.
3) Si les informations apportées par une description n'ont aucun rapport avec les connaissances de l'utilisateur: ce dernier ne pourra jamais se prononcer sur la pertinence du document . Tout jugement de sa part sera considéré comme aléatoire.

C'est trois cas illustrent bien le problème de l'inadéquation habituellement constatée entre la demande exprimé par le chercheur et les ressources décrites par le système. En effet, dans le premier cas le problème ne se pose plus étant donné que l'indexeur (humain ou automate) a inféré le sens à partir des connaissances supposées identiques à ceux de l'utilisateur. Mais, en l'absence d'un modèle cognitif de l'utilisateur (voir les paragraphes précédents) cette approche de description n'est pas envisageable.

Dans le second cas le problème de l'inadéquation pourrait se manifester étant donné que seulement quelques connaissances de l'utilisateur ont été prises en considération au moment de la description. Donc, il y a risque que des documents pertinents pour l'utilisateur seront jugés non pertinent, faute d'une représentation ne véhiculant pas le sens voulu par l'auteur.

Dans le dernier cas, le jugement aléatoire de la pertinence rend possible voire évident le problème de l'inadéquation. Ainsi, des documents pertinents seront jugés non pertinents et vice versa . De ce fait, on se demande qu'elle serait l'utilité d'apporter une reformulation à la question de l'utilisateur alors que le jugement même de la pertinence est en cause.

Selon (Le Gresnier H. 90) il y a deux types de reformulation.
- la reformulation automatique: " elle consiste à utiliser les compétences linguistiques du système pour transformer automatiquement la requête de l'utilisateur en une requête plus complète ajoutant les synonymes ou les formes des termes de la requête ". Cette reformulation à réduit le problème de l'inadéquation à un problème purement linguistique (formes dérivées des termes, sémantique des mots) or, le problème est encore plus complexe. En fait, si la description adoptée par le système est identique à celle de la deuxième ou le troisième cas ci dessus mentionnés, le problème de l'inadéquation ne peut être réduit à un choix inadéquat de termes de recherches

- La reformulation supervisée: " elle correspond à la capacité du système à établir une nouvelle équation de recherche qui soit plus proche des besoins documentaires tels qu'ils ont été précisés par une première phase de jugement de pertinence sur un lot de documents extraits par une première requête". Mais, étant donné que le jugement de pertinence dépend de la description adoptée par le système, ce type de reformulation ne peut lui non plus être efficace.

Face à ce problème complexe de description de document, on se demande est ce qu'il pourrait y exister une ou des solutions envisageables.

Chaque document s'inscrit généralement dans le contexte d'autres documents de ce fait , il entretient des rapports explicites (la citation) ou implicite (analogie) avec eux.. Cette constatation nous laisse admettre que pour le même document on peut avoir des représentations différentes si en l'insère dans des différents corpus. En d'autres termes, on décrit le même document par rapport aux différents sujets traités.

Cette idée de lier le même document à des différents corpus n'est pas en contradiction avec le processus cognitif de l'utilisateur. En fait, la notion même du sens nécessite ce lien; donner du sens à un texte quelconque, revient à le relier, le connecté à d'autre textes donc l'insérer dans un corpus.

CONCLUSION

Dans le présent travail, nous avons essayé d'aborder le problème de description de documents textuels. Pour cela, nous avons tenté en premier lieu d'explorer les typologies déjà réalisées concernant les acteurs de la circulation de documents textuels, les supports (périodiques, monographies) et les circuits ou réseaux par les quels transite l'information. Notre but était de cerner le maximum d'éléments qui entrent en jeu dans le processus de la communication écrite. Les résultats aux quels nous avons abouti, ne sont pas très complets mais, ils sont suffisants pour refléter la complexité de la communication écrite. Par exemple, la typologie des lecteurs présentée n'est autre qu'une classification générique, elle ne reflète en rien la diversité des usagers et la variété des buts de la recherche d'information.

En seconds lieu, nous avons traité les modèles existants d'indexation et de recherche d'information. La détermination des limites des modèles d'indexation, nous a conduit à admettre qu'une nouvelle approche d'indexation doit être élaborée. L'étude des modèles de recherche d'information nous a été bénéfique. En effet, on a conclu que, quel que soit le niveau de convivialité d'un système de recherche (formulation libre...), les problèmes résultant d'une indexation incertaine (quelques soient les outils de représentation) ne peuvent être écartés. En fait, la question est: comment pourrait-on apporter un jugement au degré de précision de la requête alors que la représentation des documents est en cause.

Cette nouvelle approche de description que nous avons tenté d'expliciter, pose l'investissement du lecteur (compétence pragmatique, contexte...). Elle part de l'hypothèse que la lecture est un processus actif dans lequel celui-ci crée lui même le sens du texte. Il mobilise pour cela à la fois le texte, ses propres connaissances et ses intentions de lecture. Les recherches en pragmatiques ont approuvé cette approche, elles considèrent que les ambiguïtés rencontrées dans un texte, ne peuvent être levées par l'utilisation de la grammaire et du sens des mots. De ce fait, la modélisation de l'utilisateur s'impose comme un outil indispensable à cette approche d'indexation (voir partie 3).

A partir des typologies établies dans la première partie, nous avons essayé de travailler sur la modélisation de l'utilisateur. Nous ne pouvons en aucun cas considérer le modèle élaboré, comme un "modèle cognitif". Ce dernier établit des liens complexes entre les différents éléments décrivant le comportement de l'utilisateur. En évaluant le niveau de familiarité avec le domaine de la recherche et

le niveau de précision de la requête, nous avons tenté détablir des liens entre les différents éléments (exp. entre le niveau éducationnel et les constituants de la requête...) mais, ces liens restent rudimentaires comparés à ceux du modèle cognitif. Toutefois nous estimons que la résolution de l'éternel problème de satisfaction de l'usager auquel sont confrontés les différents systèmes documentaires (bibliothèques, serveurs...) nécessite une évaluation fondée sur le niveau de familiarité de l'usager avec le domaine de recherche. Autrement dit, il faut arriver à constituer une idée sur la représentation mentale que l'utilisateur a sur son sujet de recherche.

Il reste à signaler que le problème de la description de documents (extraction du sens d'un texte) continue à soulever un débat acharné entre des chercheurs des différentes disciplines (linguistique, psychologie, philosophie...).

Si le principal reproche adressé au modèle du code (la sémiologie...) concerne le peu de cas qu'il fait du contexte dans lequel s'inscrit la longue. La délicatesse de l'opération de formalisation du contexte demeure l'entrave essentielle au modèle inférentiel.

A ce stade de recherche, ce travail, sous bon nombre d'aspects, parait bien inachevé. En effet, les points suivants mérite d'être approfondis:

- la modélisation de l'utilisateur, bien que difficile, est une nécessité car elle servira de référence lorsqu'on procédera à la description de documents textuels. Les recherches effectuées en sciences cognitives seront des outils indispensables à ce genre de travail

- la description de document en fonction des besoins de l'utilisateur demeure le problème auquel on doit accorder une attention plus poussée car, nous estimons que la réussite de l'activité documentaire dépend essentiellement du degré de fiabilité de cette description.

BIBLIOGRAPHIE

ANDEREWS J.- Indexation consistency in information science abstracts. Journal of the American Society for Information Sciences.42 (1): 1-6; 1991

BELKIN N. J.- Users interfaces for information systems. Journal of Information Science.17 (3); 327-344: 1991

BERRENDONNER A. - Discours et raisonnement. Troisième école de l'été sur les langues et langage. Paris: Association pour la Recherche Cognitive: 468-508, 1-13 Juillet, 1991.

BISSERET A. - Concevoir une "compréhension" homme-machine, in : intrefaces intelligentes dans l'information scientifique et technique. Paris: INRIA, 1992.

B. LINE M. - The publication and availability of scientific and technical papers: an analysis of requirements and the suitability of different means of meeting them. Journal of Documentation. 48 (2); 1992.

BOUCHE R. - Le syntagme nominal, une nouvelle approche des bases de données textuelles, in: Méta, pp. 428-433, septembre 1989.

CASE D. - Collection and organization of written information by social scientists and humanists: a review and exploratory study. Journal of information Science. 12 (3); 97-104; 1986

CAVAZZA M.- Extracting implicit information from free text technical report. Information Processing and management. 28 (5): 609-617; 1992

CHAUMIER J.- L'indexation documentaire: de l'analyse conceptuelle humaine à l'analyse automatique morphosyntaxique. Documentaliste; 27(6): 275-279; 1990

CHAUMIER J.-Analyse et langages documentaire. Paris: Entreprise Moderne d'édition,1982

CHAUMIER J.- Système d'information marché et technologie. Etreprise Moderne d'Édition, 1986

CHAUMIER J.- L'indexation assistée par ordinateur: principe et méthode. Documentaliste. 29 (1): 3-6; 1992

DAN SPERBER; DEIRRE W. - La pertinence. Paris: les Éditions de Minuit, 1989.

DACHELET R.- Etats de l'art de la recherche en informatique documentaire: la représentation des documents et l'accès à l'information. in: Le Document Electronique, cours INRIA, 11-15 Juin 1990.2-*CHAUMIER J.*- Système d'information marché et technologie. Etreprise Moderne d'Edition, 1986

DUROSS L. E. -A Study of discourse anaphora in scientific abstract. Journal of the American Society for Information Science. 38 (4): 255-261; 1987

DRUSS L. E.- Structure of information in full text abstract. In conference on user-oriented content-based text and image handling. Grenoble: Centre de Haute Etudes Internatinales d' Information: 182-192; 21-24 Mars, 1988

DEMAILLY A.- Comportement de communication des chercheurs scientifiques. Documentaliste science de l'information.15: 10-18; 1978

ESCARPIT R.- systèmes partiels de communication. Paris: Ecole Pratique Des Hautes Etudes, 1972.

FARROW J. - A cognitive process model of document indexing. Journal of Documentation. 47 (2): 149-166; 1991

GARDIN J.-C. - Les analyses de discours. Coll. "Zethos", 1974.

GARDIN J.-C.; Guillaume O. - Systèmes experts et sciences humaines. Paris: édition Eyrolles, 1986.

GREIMAS A. J.- Sémantique structurale. Paris: Presses Universitaire de France, 1986

GUINCHAT C.; MENOU M.- Sciences et techniques de l'information et de la documentation. Paris : UNESCO, 1990

HARRIS Zellig - Structure mathématiques du langage. Paris: Dunod, 1971

HARVEY C. F.; SMITH P. - User models in intelligent information systeme. In conference on the structuring of information. London, University of York: The Association for Information Management; 105-120; 20-22 Mars, 1991

HSINCHUN C. - Knowledge based document retrieval: frame work and design. Journal of Information Science.18: 293-314; 1992

JODOIN L. - La condensation et l'indexation: l'apport des approches de type textuel. documentaliste. 38(2): 71-74: 1992

KAYSER D. - Des machines qui comprennent notre langue, in: la recherche en intelligence artificielle. Paris: Edition du Seuil, 1987

KIRCZ J.- The use of relational data bases for electronic and conventional scientifique publishing. Journal of Information science. 13(2) : 75-89; 1987

KIRCZ J.- retherocal structure of scientific articles: the case for argumentational analysis in information retrieval.Journal of documentation.48 (2); 1992

KITTREDGE R.- Sublanguages. American Journal of Computational Linguistics 8(2); 1982

KWOK K.- Experiments with document components for indexing and retrieval. Information Processing and management.24(4): 405-417; 1988

LE GROSNIER H. - Systèmes d'accès à des ressources documentaires vers des antéserveurs intelligents. Thèse. Université d'Aix-Marseille, 1990.

LE GUERN M.- Un analyseur morpho-syntaxique pour l'indexation automatique. Le Français Moderne, (1), 22-35, 1991.

LUCTKENS E.- Le point de vue de l'utilisateur de système documentaires. Cahier dela Documentation.45(1): 3-13; 1991

MARTIN H. J. - Histoire et pouvoir de l'ecrit. Paris: Société Nouvelle F. D., 1990

MAYES P.- A comparaison of the readability of synopses and original articles for engineering synopses.Journal of the American Society for Information science.29(6): 312-313; 1978

NAOMI S.- A methode of mesuring information in language applied to medical texts. Information Processing and management. 21(4): 269-289;1985

NAOMI S.- Computer analysis of sublanguage information structures. New York Academy of sciences:161-179; 1990

NIE J. - Un modèle logique générale pour les systèmes de recherche d'information: application au prototype RIME. Thèse. Grenoble I. 1990

RECANATI F.- La pragmatique linguistique. le Courrier du CNRS. N°79: 19;1992

RADECKI T. - Fuzzy set theoretical approch to document to document retrieval. Information Processing and Management. 15: 247-259; 1979

SABAH G. -traitement automatique des langues. Le courrier du CNRS. N°79: 21; 1992

SABAH G. - L'intellignece et le langage; vol. 2. Paris: Hermes, 1990.

SALTON G. - A blueprint for automatic indexing. ACM-SIGIR Forum. 16 (2): 22-38, 1981.

SARACEVIC T.- Information seeking and retrieving: background and methodology. Journal of the American society for Information Science. 39(3); 161-176; 1988

SARACEVIC T. - A study of information and retrieving: users questions, and effectiveness. Journal of the American Society for Inf. Science 39(3): 177-196; 19889-

SAVARD R. - Principe directeurs pour l'enseignement du marketing dans la formation des bibliothécaires documentalistes archivistes . Paris: UNESCO, 1988

SCHAUBLE P. - Improving the effectiveness of systems retrieval by information structure. Information Processing Management. 25(4): 363-376; 1989

PITRAT J. - Textes ordinateurs et compréhension. Paris: Edition Eyrollles, 1985

TAKASHI M. - An approch toward functional text structure analysis of scientific and technical document. Information Processing and Mnagement.17(6): 329-333,1981

UMBERTO E. - Note sur la sémiotique de la réception. Actes Sémiotiques.9(81); 1987

WHITE D. A.- Information use and needs manufacturing organisation : organizational factors in information behaviour. International Journal of Information Management. 6: 157-170;1986


9597224