Caractérisation et découpage de textes scientifiques pour la construction de systèmes de requête personnalisés

Nabil Ben Abdallah, Christine Michel, Sylvie Lainé-Cruzel

Caractérisation et découpage de textes scientifiques pour la construction de systèmes de requête personnalisés.

Nabil Ben Abdallah[1], Christine Michel[2], Sylvie Lainé-Cruzel[2]

[1] **Laboratoire CERSI**
ENSSIB
17-21, Bd du 11 novembre 1918
69622 VILLEURBANNE CEDEX
tel: 04 72 44 28 30
abdallah@enssibhp.enssib.fr

[2] **Laboratoire RECODOC**,
Bat 721, Université Claude Bernard LYON I
43, Bd du 11 Novembre 1918
69622 VILLEURBANNE CEDEX
tel: 04 72 43 13 91
michel@recodoc.univ-lyon1.fr
slaine@cismsun.univ-lyon1.fr

Résumé

L'augmentation continuelle de la masse d'information à consulter rend de plus en plus pénible la recherche de l'information pertinente, ceci est d'autant plus vrai lorsque l'on consulte des bases de données en texte intégral. Le sens d'un texte étant donné, non seulement par son contenu mais aussi par sa structure, l'idée dominante de Profil-doc est que les parties de document auront un usage différencié a priori suivant le besoin de l'utilisateur. De ce fait, nous nous sommes attachés à implémenter un système alliant une bonne description du document et une caractérisation de l'usager. Ainsi, en fonction de l'usager et de son profil, le système, pour une requête particulière, n'utilisera que certaines parties du document initial. Nous appellerons unités documentaires les parties du texte exploitées en fonction de la structuration du document.

Dans la première partie de cet article nous présenterons en détail les propriétés qui nous permettent de décrire les documents, dans la seconde nous présenterons plus en détail le processus d'interrogation du système.

Introduction

D'une manière très schématique, dans une opération de recherche documentaire classique, l'utilisateur se contente de formuler une requête, puis le système apparie les mots de la requête avec ceux du dictionnaire qu'il possède et génère ainsi une réponse. Dans les systèmes référentiels, la structure de la base assure un certain tri au niveau de la réponse. Au contraire, dans le cas des systèmes documentaires en texte intégral, il est toujours possible de trouver des documents contenant un des termes de la question, cela ne veut cependant pas dire qu'ils seront vraiment pertinents pour l'utilisateur. Si ces systèmes savent presque toujours, proposer une réponse à la demande de l'utilisateur, ils ne répondent que partiellement à ses besoins. *« C'est depuis longtemps une évidence que les volumes d'informations vont croissant et que parallèlement va croissant la masse d'information à consulter pour trouver une information pertinente .../... On peut faire un constat simple : si le bruit et le silence sont toujours à peu près les mêmes, par exemple de 50%, un utilisateur qui reçoit dix documents en réponse à une question, en trouvera cinq pertinents. Un utilisateur qui obtiendra cent documents, en trouvera sans doute cinquante pertinents, mais aussi cinquante hors sujet. Le facteur bruit devient une gêne très réelle pour l'utilisateur dès que le volume des réponses dépasse un certain seuil «tolérable» »* [Lain 94]. Certains avancent que, le problème vient du fait que ces systèmes documentaires n'atteignent pas le sens de la question ou le sens du texte. Des chercheurs se sont donc sérieusement penchés sur des méthodes permettant de faire une analyse plus poussée du texte. Ils ont implémenté des analyseurs morpho-syntaxiques, basés sur les règles syntaxiques, mais ils se sont rendus compte que ces derniers ne "comprenaient" qu'une partie du contenu lorsque la langue et la syntaxe étaient correctes. Ils ignoraient totalement les sous-entendus, les expressions idiomatiques, etc., toutes ces formules particulières qui sont incompréhensibles sans l'expérience et la culture du lecteur[1].

Pour pallier aux limites de l'« indexation » et avoir une meilleure connaissance du fonds, les systèmes documentaires traditionnels et automatiques ont tenté de décrire les documents par des critères externes à leurs contenus. Ainsi en bibliothéconomie classique, la

[1] Ce champ sémantique est souvent appellé contexte global.

dimension d'un ouvrage, son nombre de pages, ..., sont autant de critères supplémentaires permettant de gérer le fonds, mais il est rare qu'un utilisateur se serve de ces critères pour sélectionner des documents. Grâce aux systèmes de gestion de fichiers ou aux systèmes de gestion de bases de données, la recherche d'une notice par l'ensemble des champs (zones) la décrivant est devenue possible; des champs définissant des caractéristiques externes au contenu ont ainsi pu être rajoutés : le pays et le champ disciplinaire de l'auteur, le nom du laboratoire, etc.

Nous avons choisi de travailler avec la documentation scientifique et technique, nous considèrerons donc une documentation produite essentiellement par des disciplines appartenant aux sciences de type formel pur ou empirico-formel. La dénomination technique peut refléter soit l'appartenance à des techniques soit à des disciplines dont les recherches conduisent à des applications techniques. Les documents produits par des disciplines appartenant aux sciences de type herméneutique[2] pourraient faire l'objet de notre étude si leurs contenus déductifs, expérimentaux ou évaluatifs portent sur des applications des sciences empirico-formelles ou des techniques.

Une étude approfondie[3] sur certain nombre de textes, livres, thèses, articles de revues scientifiques, a montré qu'on pouvait trouver, pour chacun d'eux, une structure générique facilement identifiable. En effet, dans la majorité des cas, un texte (article, conférence, rapport, ouvrage, etc.) a une *structure générale*, il forme une unité car il est construit pour faire passer un message : résultats de synthèse, nouvelles pistes de recherche, etc. Cette unité matérielle et intellectuelle est le résultat d'un lien parfaitement établi entre ses différentes parties, celles-ci pouvant former à leur tour des unités indépendantes remplissant une fonction bien déterminée. Ainsi, par exemple, la bibliographie est utilisée généralement pour étayer les propos cités dans les différentes parties du texte et pour donner au lecteur une idée plus ou moins exhaustive de tout ce qui a été écrit sur le sujet traité, ce qui représente d'une certaine manière le contexte du texte. Cette constatation, nous a conduit à admettre que "l'éclatement" du document en unités documentaires nous permet, tout en préservant l'unité globale du document (le lien entre l'unité documentaire et le document auquel elle appartient), de présenter à l'usager une information plus affinée et plus facile à saisir.

Mais cette structuration de document n'est pas unique; en effet, on peut aussi considérer *les différents types de textes* (publicitaires, scientifiques), *le mode d'organisation du discours* (narratif, argumentatif, etc.) ou même encore *la structure physique* (attributs typographiques, polices, espaces, etc.) comme des caractéristiques propres à discriminer une fraction du document.

Le projet Profil-doc [Lain 96] utilise ces différentes structures pour décrire les documents en unités documentaires, au sein d'un système documentaire en texte intégral. Chacune des unités est alors accessible par des index bien sûr, mais aussi par ses propriétés. Le découpage est basé sur la fonction remplie par ces parties du document et non sur leur contenu. Au niveau de l'utilisateur, ces propriétés seront autant d'outils supplémentaires utilisables lors de la requête, pour sélectionner l'information « pertinente ». En effet, on peut remarquer que l'utilisateur, face à un système en texte intégral qui lui fournit généralement trop d'information, va développer une stratégie de recherche empirique. Il va par exemple se limiter à certaines bases de données, selon la discipline ou le type de revues répertoriées, ou encore, selon la langue, pays ou année. Toutes ces stratégies ont deux caractéristiques : elles portent sur des critères (la forme, le support, le style, ...) autres que le contenu du document, elles sont très fortement individualisées et permettent une personnalisation de la recherche [Lai96].

Dans cette optique, le système Profil-doc veut aller plus loin que l'utilisation simple de ces critères pour la description et sélection des documents. En effet, ces propriétés nous permettront de sélectionner un corpus "personnalisé" suivant les caractéristiques de l'utilisateur, corpus sur lequel portera la question. En d'autres termes, ces propriétés, appariées avec le profil de l'utilisateur, nous permettent de présélectionner un ensemble de documents. **Donc, il est clair que la manière selon laquelle sont attribuées les propriétés**

[2] L'"objet" de ces sciences sont des phénomènes de comportement humain et de ce fait, il serait difficile d'établir, à partir de la simple observation, des représentations formelles (modèles ou schémas) qui pourraient être utilisées pour vérifier des propriétés empiriques directement observables. En effet, ces phénomènes sont très imprévisibles et essentiellement contextuels.

[3] Norme 5963 : Méthode d'analyse des documents - Norme ISO 2145 : Numérotation des divisions et subdivisions dans les documents écrits - Norme ISO 8613 : Architecture du document.
Norme ISO 7144 : Présentation des thèses et des documents assimilés.
Norme ISO 5966 : Présentation des documents scientifiques et techniques.

diffère complètement des processus bien connus de l'indexation.

Dans la première partie de cet article nous présenterons les propriétés qui vont nous permettre de découper et caractériser les unités documentaires. Dans la seconde partie, nous présenterons en détail le processus d'interrogation du système Profil-doc.

1- Le principe

Ce système de sélection s'appuie sur trois composantes fondamentales :

- un découpage des documents en unités documentaires,
- une caractérisation du profil de l'utilisateur,
- un système d'aiguillage.

Nous venons de voir, l'utilité du découpage et de la caractérisation des documents, nous verrons plus en détail la validation théorique du choix de chacune des propriétés.

Le «profil» de l'utilisateur est défini par diverses caractéristiques : son niveau éducationnel, son champ disciplinaire, le type de recherche souhaitée (recherche exhaustive, pointue, etc.), la situation de la recherche (réalisation d'un projet, mise à jour des connaissances, etc.). Cette caractérisation nous permet de cerner ses besoins informationnels.

Le système d'aiguillage est le coeur du processus, en effet, c'est cette fonction qui va nous permettre de définir l'ensemble des propriétés des unités documentaires souhaitables en fonction d'un profil donné. Nous n'expliciterons pas en détail dans ce travail le processus d'aiguillage, une thèse [Bena 97] est en cours de réalisation sur le sujet.

2- Le découpage des documents

Les unités appartenant au même document *(des unités soeurs)* héritent des propriétés du document père et se distinguent par des propriétés qui leur sont propres. Nous ne passons pas par une étape de compréhension du contenu pour attribuer les propriétés, en effet, elles sont soit facilement repérables à l'intérieur du document ; c'est le cas pour les propriétés *champ disciplinaire de l'auteur* et *style de l'unité documentaire* (voir plus loin) ; soit repérées par certains marqueurs (linguistiques ou autres), c'est le cas de la propriété *forme discursive* de l'unité documentaire.

2-1- Les propriétés propres au document entier

2-1-1- L'environnement de production : champ disciplinaire, profession et communauté de l'auteur

L'environnement de production nous sert à caractériser la nature et la manière de traiter les sujets, et la nature de production d'information. En effet, des régularités sont observables sur la production des écrits dans les domaines scientifiques et techniques suivant *la communauté, le champ disciplinaire ou la profession de l'auteur*. L'utilisation de ces trois propriétés nous permettra de cerner plus finement, par la caractérisation de l'auteur, l'écrit en question et ainsi de présenter à l'utilisateur, selon son objectif, les types de documents les plus appropriés.

Champs disciplinaire	Profession	Communauté
♦Anthropologie ♦Astronomie-Astrophysique ♦... ♦Technologie métallurgique ♦Zoologie	♦étudiant ♦enseignant ♦enseignant chercheur ♦chercheur ♦spécialiste en communication (Journaliste, etc.) ♦ingénieur ♦administratif	♦étudiants ♦universitaires ♦grands industriels ♦PME-PMI ♦secteur public et parapublic ♦individus : les auteurs n'appartenant pas à une communauté donnée.

2-1-2- Le support de diffusion : Type de l'environnement éditorial, type d'article

De la même manière, le support de diffusion (la revue, le type d'article) caractérise le type d'information véhiculée, et nous renseigne sur son utilisation et son type de public privilégié.

Nous pouvons considérer trois grandes catégories d'écrits scientifiques et techniques : les écrits destinés à des chercheurs, les écrits destinés à un public plus large (les écrits de vulgarisation) et des écrits à caractère didactique. C'est en fonction de la nature de la communication et de la catégorie sociale des interlocuteurs que l'auteur d'un document choisit généralement son environnement éditorial. Le « type d'article » est largement lié à l'environnement éditorial, cependant, prenons l'exemple des revues primaires qui publient des travaux de recherches, les types d'articles trouvés sont alors beaucoup plus hétérogènes, et la fonction remplie par chacun diffère. Le type d'article doit alors être caractérisé car la fonction visée est complètement différent.

Type environnement éditorial	Type d'article
♦ thèses et mémoires de fin d'étude ♦ revues professionnelles ♦ presse grand public (revues de vulgarisation ou presse grand public) ♦ revues primaires ♦ divers ♦	♦ article original ♦ article de synthèse ♦ rapport de conférence ♦ résumé de conférence ♦ éditorial ♦ lettre à l'éditeur ♦ recension d'ouvrage ♦ article de vulgarisation grand public ♦ article de vulgarisation public averti ♦ écrits à caractère didactique ♦ article technique de métier

2-2- Les propriétés propres aux unités documentaires

Nous venons de passer en revue l'ensemble des propriétés « héritées » du document initial (*document père*); nous allons à présent présenter trois propriétés caractérisant individuellement chaque unité. Ces propriétés justifieront le processus de découpage du document père, car elle définissent **la fonction d'usage** que l'on associe au document.

Il existe différentes formes de structuration des documents, imposées soit par les règles de production des documents scientifiques et techniques, soit par les auteurs au moment de la rédaction. D'une façon générale, les documents courts tels que les articles publiés dans les revues primaires, possèdent habituellement une structure bien définie : résumé et mots clés, introduction, matériels et méthodes, résultats, discussion, conclusion et bibliographie. En revanche, les documents longs tels que les ouvrages, les thèses, les mémoires, etc. sont souvent structurés en parties, chapitres, paragraphes, sous-paragraphes. La lecture de ces documents est facilitée par une table des matières placée, suivant les éditions, à la fin ou au début du document et qui reprend fidèlement la structure logique du document.

Les avis sont partagés quand on essaie de définir ce qu'est la structure des documents. Par exemple, J. André et V. Quint [Andr 94] : «*Par structure logique, on entend donc l'organisation du document en entités telles que chapitres, sections, titres, paragraphes, notes, citations, etc. .../.... Elles ajoutent au contenu même un niveau de signification supplémentaire, qui permet au lecteur de se repérer dans le parcours du document* ». D. Malrieu [Malr] différencie : *«la superstructure textuelle (ensemble des titres, sous titres) fournit des informations sur l'organisation fonctionnelle du texte ; elle contient des informations sémantiques et typologiques : le titre délimite le type de questionnement source, et d'élaboration*[4]*, le domaine du discours .. »* et *« la superstructure textuelle qui est doublée d'une macrostructure, entendue comme ensemble des séquences fonctionnelles d'un texte, qui ont des marques linguistiques mais pas de marques graphiques...»*

Nous utilisons, dans ce qui suit, l'expression **«unité logique»** pour désigner un **«paragraphe» titré du**

[4]L'auteur considère que l'oganisation conventionnelle des textes est contrainte par le type de questionnement et le paradigme de résolution, qui définissent le type d'élaboration textuelle. Nous citons à titre d'exemple le plan IMRAD appliqué à l'ensembles des articles originaux en science de type empirico-formel et qui reflète bien un type de questionnement et un paradigme donné de résolution. (voir : Vers une définition des sciences)

document. Les paragraphes non titrés caractérisés par un alinéa, un saut de ligne (ou les deux) ne seront pas considérés comme des unités logiques. Nous pouvons éventuellement considérer les subdivisions de cette unité comme des unités logiques à part si elles se composent de paragraphes titrés.

2-2-1- Type de l'unité logique

Il nous reste à voir maintenant l'utilité de la structuration des documents pour l'auteur ainsi que pour le lecteur. Le premier a besoin de mettre en oeuvre un enchaînement explicite d'idées pour faire passer son message (résultats de recherche, etc.). L'écrit scientifique doit être en fait une traduction fidèle de la démarche adoptée. En fait ce n'est pas par hasard que la structuration des écrits des sciences de type empirico-formel diffère de celles de type herméneutique. Le chapitre, le paragraphe, etc. ne sont que des «macroformes» d'illustration de l'approche scientifique adoptée. D. Lewis cité par J. Heslot [Hesl 80] dans son analyse de la convention, parle de régularités d'écriture (en soulignant qu'il peut exister d'autres régularités d'écritures).

Le processus de lecture n'est pas naturel et régulier dans l'absolu, il change en fonction de la tâche à réaliser (ex. mise à jour des connaissances, réalisation d'une expérimentation, ...), de ses contraintes (temps disponible, ...), du document à consulter, en résumé de la situation de lecture. Le lecteur opère, à l'inverse de l'auteur, de la structure vers les idées véhiculées et c'est ainsi que la structure devient un outil de vérification de la cohérence d'un texte. Le lecteur averti, suivant son besoin, commence à repérer les parties du document qui l'intéressent puis au sein de chaque partie, il localise les passages dont il a besoin.

Notre hypothèse est qu'un document entier forme une cohésion puisqu'il est construit pour faire passer un message : résultat de synthèse, nouvelles pistes de recherche etc. Dans le contexte documentaire, il est intéressant d'exploiter la possibilité de consultation de parties isolées du document scientifique. C'est pour cette raison que nous avons pris la décision de découper le document (article, etc.) en parties «sémantiquement indépendantes». Le découpage est indépendent du support (article, ouvrage,), du type de l'article et de l'environnement éditorial. Chaque type d'unité documentaire remplit généralement une fonction donnée, toutes ne sont pas obigatoires, et certaines d'entre elles sont répétables à l'intérieur d'un même document. Nous avons reconnus les quatorze types suivants ainsi que leur caractère explicite ou non:

- résumé (explicite)
- introduction (explicite),
- description du contexte (implicite),
- description du thème (implicite),
- description de la méthode (implicite),
- environnement (implicite),
- développement (implicite),
- expérimentation (implicite),
- résultats (implicite),
- discussion (implicite),
- conclusion (explicite),
- bibliographie(explicite),
- table des matières (explicite),
- annexes (explicite).

Une unité explicite est désignée par l'auteur au moment de la rédaction, par exemple, l'unité «bibliographie» qui se distingue des autres unités par le style de son texte (normes de catalogage) et la nature de l'information véhiculée (information référentielle), est généralement placée à la fin du document et désignée par le titre «bibliographie». Le repérage des unités explicites ne devrait donc pas présenter de problèmes.

En revanche, les unités implicites ne portent pas forcément des titres tels que développement, environnement, etc.; et nécessitent, par conséquent, le repérage de quelques marqueurs[5] pour décider de quel type d'unité logique il s'agit. Nous ne pouvons en effet pas uniquement nous baser sur les titres qu'attribue l'auteur, car ils représentent le contenu de l'unité logique et non pas sa fonction informative. Nous entendons par fonction informative, ce que l'unité est sensée implicitement apporter comme information. Par exemple, la partie introduction d'un article original a, en règle générale, pour fonction d'indiquer pourquoi l'auteur traité le sujet, le contexte dans lequel il sera abordé, la démarche adoptée et les liens avec l'actualité de la recherche. Cette fonction est indépendante du sujet traité et du contenu.

En consultant l'introduction, le lecteur prévoit le type d'information qu'elle renferme et ceci, quelque soit le thème traité par l'article. Il sait qu'il n'y trouvera pas les détails de l'expérimentation.

[5] Il ne s'agit pas forcément des marqueurs linguistiques. Nous y reviendrons quand nous aurons à définir les différents types d'unités logiques.

Cette constatation reste valable pour d'autres parties de l'article qui ont, elles aussi, des fonctions informatives bien définies.

Mais si les unités explicites (introduction, résumé, etc.) ont des fonctions informationnelles connues de l'auteur et du lecteur ; on pourra alors parler de fonctions informationnelles «conventionnelles», il semble nécessaire de préciser les fonctions informationnelles que nous attribuons aux unités implicites.

Définitions des différents types d'unités logiques.

Résumé **:** Quel que soit le type du résumé (résumé d'auteur ou résumé documentaire), il doit permettre au lecteur de situer le sujet de l'étude, d'avoir une idée sur la méthode employée et les éléments nouveaux mis en évidence par l'étude, il doit être particulièrement informatif et d'une parfaite autonomie de lecture.

Introduction : Comme le résumé, il s'agit d'une unité explicite facilement repérable dans le document. Elle permet de faire l'état de la science d'où part la recherche en situant le sujet dans son contexte scientifique et à partir de ce bilan, l'auteur montrera en quoi l'imperfection des connaissances actuelles sur le sujet justifie sa propre recherche. Suivant l'environnement éditorial, l'auteur détaille parfois les hypothèses de son travail. Comme pour le résumé, l'introduction remplit une fonction informationnelle bien définie.

Description du contexte : Le contexte est l'ensemble des circonstances, d'ordre scientifique (état de l'art), économique ou même politique, dans lesquelles se produit l'étude présentée dans l'article ou l'ouvrage. En fonction de la nature du travail, l'auteur peut estimer que la partie introduction n'est pas suffisante pour décrire d'une façon exhaustive le contexte scientifique de son travail. L'unité logique « description du contexte » est déterminée par le repérage soit d'une dénomination explicite, «l'état de l'art », soit d'indices particuliers tels que la présence d'un grand nombre de citations.

Description du thème : La description de thème replace le sujet de l'étude dans une problématique de recherche large, dans le cas où l'introduction ne suffit pas à préciser les concepts et à établir le formalisme voulu. Nous remarquerons cette unité par des énoncés de type : nous présentons, dans cette section, le sujet ou le problème ..., et le fait que l'on soit dans une unité logique autre que l'introduction

Description de la méthode : Il s'agit de la méthode utilisée pour réaliser l'étude. Cette unité correspond, dans le plan IMRAD à la partie « matériel et méthode », quand l'auteur ne suit pas ce plan, on parle de «modèle d'étude », « méthode d'observation et évaluation », etc. Si l'auteur ne désigne pas l'unité description de la méthode par un titre explicite, il devient indispensable d'effectuer une lecture, le repérage automatique n'est pas envisageable pour l'instant.

Environnement : C'est l'unité logique qui présente le matériel utilisé généralement dans l'expérimentation

Expérimentation : La diversité de l'information à traiter nous amène à accorder un sens assez large à l'expérimentation., cela peut être tout aussi bien un travail de mesure ou des simulations en laboratoire, mais aussi une collecte de données (le questionnaire, l'observation, etc.).

Résultats : Dans le cas de sciences empirico-formelles, les résultats sont ceux de l'expérimentation si elle a eu lieu. Suivant la nature du travail, les résultats sont présentés soit sous forme globale visuelle, graphiques, courbes, ou bien récapitulés dans des tableaux de chiffres, ou bien simplement dans un texte (suite d'énoncés ou théorème dans le cas d'études théoriques).

Discussion : Outre le titre explicite, la partie discussion pourrait être déterminée par un texte argumentatif renfermant des données de la partie résultats et des renvois à des références bibliographiques.

Développement : Le terme développement peut désigner une partie du document présentant plus en détail un traitement ou un problème particulier, nous utiliserons la caractéristique « développement » par défaut, lorsque nous n'aurons pu attribuer une caractéristique plus pertinente pour la partie de document concernée.

Conclusion : L'unité logique « conclusion » est d'une manière générale facilement identifiable, un problème se pose cependant lorqu'elle est incluse dans la discussion.

Bibliographie : La rédaction de la bibliographie obéit généralement à des normes données de catalogage, ainsi son repérage dans le document ne devrait poser aucun problème.

Table des matières : L'auteur n'y reprend que la structure du corps du texte. Elle se distingue du reste du document par son style et elle est généralement titrée « sommaire » ou « table des matières ».

Annexes : Elle est utilisée comme une information complémentaire parfois indispensable à la compréhension de la recherche. Cette unité n'est pas

caractérisée par un style donné mais est généralement titrée « annexe ».

2-2-2- Forme discursive du texte

Dans le cadre du projet profil-doc, il est pertinent de décrire les unités documentaires par leur forme discursive. Marie-France EHRLITCH [Ehrl 94] établit un lien entre le type de texte, la tâche à réaliser et la nature de la représentation. Prenons l'exemple d'un usager qui est dans une situation d'interprétation des données, c'est le texte argumentatif qui répondrait le mieux à son besoin

Reste à déterminer comment effectuer une typologie des textes. Les travaux des linguistes sur la question se partagent entre le repérage des marqueurs linguistiques et la référence à la situation de communication.

Selon Beaugrande et Dressler cités par F. Rastier [Rast 91] : Les textes descriptifs «*servent à remplir des espaces de savoir dans lesquels les centres de contrôle sont des objets ou des situations*»; les textes narratifs sont «*ceux qui disposent dans un ordre séquentiel des actions et des événements*»; enfin les textes argumentatifs sont «*ceux qui favorisent comme vraie vs fausse ou positive vs négative l'acception ou l'évaluation d'idées ou de convictions déterminées*». Selon F. Rastier [Rast 91] «*La dominance d'une fonction permettrait certes de distinguer trois types de textes,.../... Mais en fait Beaugrande et Dressler sont contraints de faire appel à des connaissances encyclopédiques sur la situation de la communication : la fiche technique d'un projecteur de diapositives contient, par exemple, des éléments typiques de textes argumentatif mais relève principalement de la fonction descriptive parce qu'elle a pour but pour décrire l'usage et le maniement d'un appareil donné.*[6]»

Si Rastier a mis l'accent sur le contexte de communication pour «réfuter» la typologie établie par Beaugrande et Dressler ; Combettes [Comb 88], quant à lui, conteste le parallélisme « simple » entre une intention donnée et une marque linguistique. C'est une combinaison de faits de langue d'ordres différents qui peut déterminer un type donné de texte. En résumé, la forme discursive du texte est déterminée par la fonction informative du texte, les différents faits de langue utilisés et la situation de communication.

Sur le plan linguistique, en se limitant à l'information scientifique et technique, la relation argumentative est déterminée généralement par trois éléments : une assertion de départ (donnée, prémisse), une assertion d'arrivée (conclusion, résultat) et une ou plusieurs assertions de passage qui permet de passer de l'une à l'autre (preuve, argument). Un texte argumentatif, sera repéré par de nombreux connecteurs logiques

Patrick CHARAUDEAU [Char 92] définit sept relations logiques : la conjonction, la disjonction, la restriction, l'opposition, l'implication, l'explication et l'hypothèse. Chaque relation logique est exprimée par des connecteurs donnés, étant pour nous, des marqueurs caractérisant le texte argumentatif.

Relations logiques	Marqueurs
• Conjonction	• et, avec
• Disjonction	• ou
• Restriction	• mais, bien que, cependant, même si, en supposant que, admettons que, je concède que, néanmoins, il reste que, par ailleurs, pourtant
• Opposition	
• Implication	• pendant que, tandis que, alors que
• Explication	• si, tout ce qui, seul, à condition que, pour autant que, dans la mesure où, pourvu que, dans la mesure où
• Hypothèse	• si, parce que, de sorte que, pour, afin que, dans le but de, en vue de, de manière que, ainsi, donc, de sorte que, parce que, comme, car, sous prétexte que, à condition que, dès lors que, aussi • si, dans l'hypothèse où, au cas où, si jamais, si par hasard, supposer que, en admettant que, en l'absence de, à défaut de

Texte descriptif : Selon A. J. Greimas cité par J. M. Adam [Adam 92] : *«On appelle ... description, au niveau de l'organisation discursive, une séquence de surface que l'on oppose à un dialogue, récit, tableau, etc., en postulant implicitement que ses qualités formelles autorisent à la soumettre à l'analyse qualificative. Dans ce cas, description doit être considéré comme une dénomination provisoire d'un objet qui reste à définir».* Cette courte définition de la «description» nous laisse admettre que le processus descriptif diffère du narratif (récit), il représente une organisation discursive du texte qui permet d'identifier et de qualifier des êtres. L'accumulation de certaines marques linguistiques, comme l'emploi des temps qui ne traduisent pas une progression de l'action, la notation des lieux et l'emploi de certaines catégories grammaticales (adjectifs, etc.) pourrait caractériser un processus descriptif.

Texte Narratif : Pour P. Charaudeau [Char 92] : *«le narratif nous fait découvrir un monde qui est construit dans le déroulement même d'une succession d'actions qui s'influencent les unes les autres et se transforment dans un enchaînement progressif».* Cette manière de définir le processus narratif, nous permet d'admettre que des marques linguistiques comme les temps employés pourraient caractériser ce processus. Mais comme nous

[6] F. Rastier. Sens et textualité (p 45) . Hachette. 1989. 279 p.

l'avons déjà signalé, des indices linguistiques ne peuvent, à eux seuls, déterminer le mode d'organisation du discours.

2-2-3- Style du document (SD)

Elle sert à décrire l'unité documentaire par des *« attributs de formes »*. Ainsi, par exemple, une unité documentaire peut se distinguer d'une autre par la présence de résultats numériques. *(Dans une base de données la propriété "style de document" a un pouvoir discriminatoire assez important)*. On a les modalités suivantes :

Littéraire pur

Contenant "surtout" des données numériques

Formules de calcul ou équation

Schéma ou figure

3- La phase d'interrogation

Nous venons de décrire notre manière de structurer les documents et ainsi la structure du corpus d'interrogation. Nous allons à présent présenter le processus d'interrogation du système profil-doc. Reportons nous à la figure 1 et explicitons les cinq étapes qui composent l'interrogation.

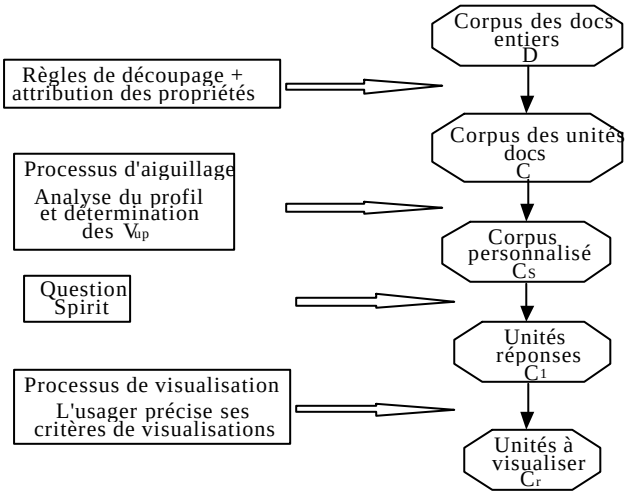


Figure 1 : Processus d'interrogation

3-1- Description du profil de l'usager

Dans un contexte documentaire, déterminer la connaissance d'un utilisateur sur un sujet permet d'améliorer considérablement sa recherche. Nous n'irons pas aussi loin dans l'étude de l'usager dans le cas du système profil-doc, seules nous intéressent les caractéristiques qui influencent la réalisation de la recherche d'information.

M.-F. Blanquet [Blan 94] a présenté comment les intermédiaires en information développent et utilisent les modèles usagers. Pour elle, le dialogue professionnel/usager comprend deux parties : l'une porte sur le sujet de la recherche en soi, la question; l'autre sur l'environnement spécifique de cette dernière, étant entendu qu'une même question à formulation identique peut entraîner des réponses fort différentes suivant le profil du demandeur. L'environnement de la question a précisément pour objectif de prendre la mesure des besoins de l'usager. **Une même question peut induire des réponses différentes en fonction d'un grand nombre de facteurs objectifs ou subjectifs : le niveau de l'utilisateur, sa profession, les langues comprises , l'utilisation précise pour l'information trouvée et tout un contexte personnel dont, d'une façon implicite, le professionnel de l'information tient compte pour effectuer sa recherche.**

Pour P. J. DANIELS [Dani 86] les modèles usagers existants appartiennent à deux grandes classes: les modèles quantitatifs empiriques et les modèles analytiques cognitifs. Les modèles quantitatifs empiriques sont des formalismes restreints d'une classe générale d'usagers. Le but essentiel des tels modèles est de corréler le comportement "externe" de l'usager avec les paramètres de conception du système. Ces modèles n'essayent pas de présenter les raisonnements, les croyances et les connaissances des usagers. Les modèles cognitifs empiriques tentent de modeler le comportement cognitif de l'usager d'une manière plus qualitative. En effet, de tels aspects peuvent inclure les connaissances sous-jacentes, le but, les croyances, le style d'interaction et la connaissance du système par l'utilisateur. Il admet que la modélisation de l'usager est une fonction constituée de cinq "sous-fonctions":

- USER: détermine le statut de l'usager

- UGOAL: détermine les buts de l'usager

- KNOW: les états de connaissances de l'usager dans un domaine

- IRS: détermine la familiarité de l'usager avec le système documentaire

- Back: détermine le "background" de l'usager

Dans ce même esprit, nous avons choisi les quatre caractéristiques suivantes: Niveau éducationnel, Champ disciplinaire, Etapes de recherche, Type de recherche. Lorsque l'usager «entre » sur le système il renseigne

donc à partir des listes suivantes, chaque caractéristique[7].

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

3-2- Mise en place d'une fonction d'aiguillage

La fonction d'aiguillage est vraiment le coeur du système, c'est elle qui va extraire les unités documentaires du corpus, en fonction du profil donc de l'usage fait par l'utilisateur. Nous ne la décrirons pas en détail dans cet article, elle fait en effet l'objet d'une thèse à soutenir prochainement [Bena 97] et nécessite, du fait de sa complexité, un article a part entière pour sa présentation.

Brièvement, nous nous sommes basés sur la littérature ainsi que sur une enquête que nous avons effectuée sur les usages et habitudes des chercheurs en SIC, sciences pharmaceutiques et sciences physiques, pour construire une matrice «profil-utilisateurs ». La projection des caractéristiques décrivant le profil de l'utilisateur, sur cette matrice, permet d'obtenir un ensemble de propriétés que doivent présenter les unités documentaires sur lesquelles portera la requête de l'utilisateur.

3-3- Sélection d'un sous corpus de la base

Les unités documentaires ne vont pas valider toutes les combinaisons des propriétés référencées précédemment. Une série de fonctions booléennes nous permet de combiner ces propriétés, pour effectivement extraire les unités. Le premier sous-corpus Cs, est extrait à partir de ces propriétés, c'est sur ce sous corpus que la requête de l'usager va porter.

3-4- Requête de l'usager

La requête de l'usager (ainsi que l'alimentation de la base) se fait à partir du logiciel documentaire SPIRIT[8]. Il permet la génération automatique de bases de données textuelles, et leurs interrogations en langage naturel. La réponse du système est présentée sous forme d'une liste triée de documents [Rada 88]. Cette valeur de pertinence est évaluée par un poids qui est calculé pour chacune de ces classes. Le poids de la classe est l'addition des poids de tous les éléments (mots), composant la question. La particularité de SPIRIT est de coupler deux analyses, linguistique et statistique, qui sélectionnent les concepts et les pondèrent.

3-5- Visualisation et navigation de l'usager

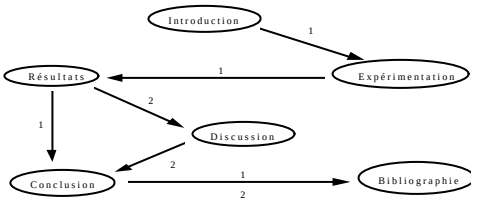
La requête de l'usager a permis de sélectionner une partie de Cs, C1 qui contient les unités de sens « pertinent ». Les parties du document à présenter à l'utilisateur ne sont pas nécessairement celles sur lesquelles la requête a été effectuée. C'est l'utilisateur, par le processus de navigation qui va choisir de visualiser telle ou telle partie du document entier.

Nous avons choisi de proposer une lecture de type navigationelle car la lecture de l'article scientifique s'y prête du fait de sa structure et de «l'indépendance sémantique» de ses différentes parties. En termes grossièrement simplifiés, la lecture navigationnelle traduit une démarche «naturelle» de compréhension. Ainsi un chemin donné de lecture reflète, généralement, un sens voulu par l'utilisateur (le lecteur).

Nous proposons ici trois scénarios de navigation qui ne sont pas limitatifs.

Dans le premier scénario nous proposons à l'utilisateur de naviguer dans les unités d'un même document suivant un ordre prédéfini par les habitudes de lecture des chercheurs mise en évidence dans le questionnaire.

Exemple:



[7] La définition des modalités s'est fait en suivant les résultats d'un questionnaire que nous avons effectue aupres de chercheurs en SIC, sciences pharmaceutiques, et sciences physiques

[8] développé par la société TGID

Le second scénario consisterait à définir des chemins de navigation en fonction des propriétés de l'unité d'appel, en présentant en premier par exemple les unités validant des propriétés type de l'unité logique, forme discursive et/ou style, puis les unités validant l'environnement de production puis celle validant le support de diffusion.

Le dernier type de scénario suit la lecture séquentielle de l'article, en présentant les unités, lorsqu'elles sont extraites du même document, selon leur ordre dans le document initial.

Conclusion

Si nous avons trouvé de nombreuses réflexions de linguistes et documentalistes sur la caractérisation de l'information scientifique et technique, par contre nous manquons de données sur l'exploitation qu'en font les utilisateurs. Seules ces dernières seraient à même d'évaluer la pertinence du découpage et la validité des indices que nous avons retenus pour caractériser les parties de discours.

Pour avancer sur ces questions, il semble nécessaire à l'heure actuelle, de chercher systématiquement et de s'appuyer sur des données, de type questionnaire pour mieux cerner les « stratégies personnalisées » ; et d'effectuer des expérimentations et des évaluations sur le prototype déjà établi. C'est le projet de notre travail actuel.

Bibliographie

[Adam 92] Adam Jean-Michel, Petitjean André. Le texte descriptif. Nathan. 1992. 239 p.

[Andr 94] André J. ; Quint V. Structure et modèle de documents in le document électronique : cours de l'INRIA. 1990.

[Bena 97] N. Ben Abdallah. Analyse et structuration de documents scientifiques pour un accès personnalisé à l'information utile : vers un système d'information évolué. (Thèse à soutenir en mai 97)

[Blan 94] Blanquet Marie-France. Intelligence artificielle et système d'information. ESF. 1994. 269 p.

[Char 92] Charaudeau Patrick. Grammaire du sens et de l'expression. Hachette. 1992 .927p.

[Comb 88] Combettes Bernard. Le texte informatif. De Boeck Université. 1988. 140 p.

[Dani 86] Daniels P.J. Cognitive models in information retrieval an evaluative review. Journal of documentation , Vol. 42, N°4, Decembre 1986, pp. 272-304.

[Ehrl 94] Ehrlitch Marie-France. Mémoire et compréhension du langage. 1994. 350 p.

[Lain 94] Lainé-Cruzel Sylvie. Vers de nouveaux systèmes d'information prenant en compte le profil des utilisateurs. Documentaliste. Sciences de l'information - 1994 - 31 (3) - pp 143-147.

[Lain 96] Lainé-Cruzel Sylvie, Lafouge Thierry, Lardy Jean-Pierre, Ben Abdallah Nabil. Improving information retrieval by combining user profile and document segmentation. Information Processing and management -1996- vol 32 n 3 - pp 305-315.

[Malr] Malrieu Denise. Genre textuel, surlignage et marques linguistiques d'importance.

[Rast 91] Rastier François. Sémantique et recherche cognitive. PUF. 1991. 262 p.

[Rada 88] Radasoa H. Méthodes d'amélioration de la pertinence des réponses dans un système de bases de données textuelles.Thèse. Université Paris Sud. Centre d'Orsay - 28 Novembre 1988 - pp 156.