Etat de l'art de la recherche en informatique documentaire : la representation des documents et l'acces a l'information

Roland Dachelet

INRIA

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Rapports de Recherche

N° 1201

Programme 8
Communication Homme-Machine

ETAT DE L'ART DE LA RECHERCHE EN INFORMATIQUE DOCUMENTAIRE: LA REPRESENTATION DES DOCUMENTS ET L'ACCES A L'INFORMATION

36 p (219)

Roland DACHELET

Avril 1990

Programme 8

Communication homme-machine

Etat de l'art de la recherche en informatique documentaire : la représentation des documents et l'accès à l'information

State of the art in information retrieval : research trends in document representation and access

Roland DACHELET

Avril 1990

* Etude réalisée grâce au concours du Ministère de la Recherche et de la Technologie (Délégation à l'Information Scientifique et Technique - DIST).

RESUME

Ce survol des recherches en informatique documentaire aborde essentiellement deux thèmes : la représentation du contenu des documents d'une part, l'accès à l'information d'autre part.
Les documents envisagés sont essentiellement les documents textuels.
Les deux approches traditionnelles de la représentation du contenu des documents sont confrontées: statistiques d'un côté, linguistico-conceptuelles de l'autre. Il apparaît que cette opposition d'approche recouvre en partie une dichotomie qui oppose sens et valeur. Les techniques de représentation du contenu et les avancées réalisées ces dernières années pour chacune des deux approches sont décrites. Il apparaît d'autre part que l'évolution de la recherche sur les modes de représentation du contenu va de pair avec une diversification des types de documents électroniques : textes, "information formats", bases de connaissances, hypertextes et hypermedia.
S'agissant de l'accès à l'information, on montre que les dernières années ont vu la problématique de l'appariement d'une requête à un ensemble de documents, dont on décrit les progrès techniques, se positionner progressivement à l'intérieur d'un cadre plus vaste, celui de la satisfaction du besoin d'information de l'utilisateur. Cela implique deux choses: d'abord prendre en compte les caractéristiques propres de l'utilisateur et au-delà, renverser la perspective en considérant qu'une base de données ne fait que stocker des données, données dont c'est à l'utilisateur lui-même de décider lesquelles constituent les informations qu'il recherche.
On est donc passé de l'implémentation d'interfaces conviviales à l'implémentation de systèmes dits hybrides ou multi-experts dont on caractérise les techniques de mise en oeuvre.

Mots-clés : Information documentaire - état de l'art - recherche assistée - science cognitive - représentation de document.

SUMMARY

This overview of information retrieval research deals essentially with two topics: content representation of documents on one hand, access to information on the other.
Documents concerned by the study are essentially textual documents.
Statistical and linguistico-conceptual approaches to content representation of documents are contrasted. This contrast appears to be partly paralleled by a dichotomy which opposes meaning and value. Techniques and achievements for content representation arrived at over the last years for each of the two approaches are described. Progress of research concerning modes of content representation is accompanied by an increasing diversification of electronic documents types : texts, "information formats", knowledge bases, hypertexts and hypermedia.
As for information access, last years are shown to bear witness of the fact that the problem of matching a query with documents, matching the technical advances of which are described, has been progressively moved into a more general framework: user information need fulfilment. This has two implications: first the need to take into account user (s) distinctive features, then, and beyond, the need to change perspective by considering that a database contains only data, the informational status of which is to be decided upon by the user.
Implementation of friendly interfaces has given way to the implementation of so-called hybrid or multi-expert systems the operating techniques of which are characterized.

Key-words : Information retrieval - state of the art - assisted retrieval - cognitive science - text surrogates.

SOMMAIRE

ETAT DE L'ART ET PERSPECTIVES

DE LA RECHERCHE EN INFORMATIQUE DOCUMENTAIRE*

L'exposé qui suit est structuré autour de deux thèmes. Le premier concerne les moyens de représentation du contenu des documents d'une base de données. Le travail qui s'accomplit en ce domaine est un travail de fond dont les résultats progressent lentement. Notre second thème : la recherche d'information, semble par contraste, être l'objet d'une activité beaucoup plus intense et quasiment monopoliser les efforts de la recherche.

I LA REPRESENTATION DES DOCUMENTS

Les documents dont il s'agit ici sont des documents textuels. Pourtant la nécessité de prendre en compte des documents multi-media se fait de plus en plus pressante en raison notamment de la complexité croissante des systèmes bureautiques. Les stations de travail mettent à la disposition des utilisateurs des outils de création et d'édition de documents multi-media. Par ailleurs, se multiplient les dispositifs de lecture optique qui eux aussi accroissent les capacités multi-media de ces systèmes. Cependant, on sait mal aujourd'hui, représenter le contenu de documents non textuels, images, séquences sonores, autrement que par des descriptions textuelles.

Les progrès en ce domaine sont pourtant nécessaires car il ne fait pas de doute que la bureautique est un type d'application qui va focaliser, si ce n'est déjà fait, une part majeure des efforts de l'informatique documentaire. Les problèmes posés sont stimulants : documents multi-media, absence d'intermédiaire documentaliste, intégration dans les outils de bureau existants. Les projets européens MULTOS [CONST87], [TSIC83] et EXPRIM [CREH86], [HALIN88] financés dans le cadre d'ESPRIT visent expressément ce genre d'application.
Pour l'heure, il s'agit de progresser dans le traitement des textes eux-mêmes.

---

* Mes remerciements vont à A. Bisseret, P. Falzon, J.C. Le Moal et A.M. Vercoustre.

I.1 REPRESENTER UN TEXTE

I.1.1 Les mots

Les techniques d'indexation automatique classiques voient la représentation des documents comme constituée de la totalité des formes que ceux-ci contiennent. De cet ensemble sont retranchés les mots grammaticaux trop fréquents et sans poids sémantique. Par forme il faut entendre séquence de caractères délimitée par des blancs ou des signes de ponctuation.

Les tenants de cette approche se sont bien vite rendu compte que des formes voisines, les formes conjuguées d'un verbe par exemple, ont entre elles une intersection formelle et sémantique qui est celle du mot. D'où la mise au point d'algorithmes d'élimination des séquences terminales. Toutefois, la notion de séquence terminale n'est qu'une approximation de la notion de suffixe. Pour qu'une séquence terminale soit un suffixe, il faut reconnaître un mot. Pour cela, il faut disposer d'un dictionnaire qui est le répertoire des mots, de leur signification et des flexions dont chacun d'eux peut être affecté. Ce sont là les premiers pas d'un traitement linguistique des textes.

Aujourd'hui on peut considérer comme acquis et maitrisé ce type de traitement même s'il n'est pas toujours mis en œuvre dans les systèmes commerciaux, tant s'en faut.

I.1.2 Les termes

Ce qui est toujours objet de recherche, c'est l'extraction automatique de séquences de mots. On sait que dans les domaines techniques surtout, les unités sémantiquement chargées sont les termes, c'est-à-dire des unités syntagmatiques dont la taille est supérieure au mot.

Deux types de techniques sont mises en œuvre pour effectuer ce genre de traitement : des techniques statistiques d'une part et des techniques linguistiques d'autre part.

Les techniques statistiques repèrent dans les textes les "mots" fréquemment co-occurents pour élaborer des "cartes de termes associés" (en anglais : "term association maps"). Le diagnostic porté sur l'utilisation de ces méthodes par [SALT86a] est modérément positif : d'une part, les mots ainsi associés ne présenteraient pas souvent

une unité de sens, d'autre part, l'augmentation de la précision[1] des documents rappelés lors d'une recherche ne serait pas concluante.

La mise en œuvre des "termes associés" peut conduire à une amélioration du taux de rappel[2] en accroissant les possibilités d'appariement entre les termes des requêtes et les termes affectés aux documents... "Malheureusement l'expérimentation montre que seules 20% des associations entre paires de mots élaborées automatiquement sont sémantiquement légitimes .... Globalement, le gain en précision résultant du processus de construction de syntagme n'est, en moyenne, pour la collection CACM, que de 8% pour les cinq valeurs du taux de rappel" [SALT86a]. La figure 1 ci-après donne le détail des résultats obtenus sur deux collections de documents [SALT86a].

Taux de Rappel	Collection 1 (CACM) 3204 docs, 52 requêtes			Collection 2 1460 docs, 76 requêtes		
	uni-termes	uni-termes plus syntagmes		uni-termes	uni-termes plus syntagmes	
.1	.5086	.5580	+10%	.4919	.4813	-2%
.3	.3672	.4065	+11%	.3118	.3158	+1%
.5	.2398	.2835	+18%	.2320	.2291	-1%
.7	.1462	.1466	0%	.1504	.1463	-3%
.9	.0711	.0704	- 1%	.0739	.0717	-3%
Amélioration Moyenne			+7,6%			-1,6%

Figure 1. Efficacité de la technique des mots associés. (tiré de [SALT86a])

Les techniques linguistiques quant à elles, se fondent sur un traitement morphologique et syntaxique pour extraire des unités syntagmatiques syntaxiquement et sémantiquement valides. Une grande part des travaux du CRISS est consacrée à cette entreprise [ROUAU87], [ANTO88]; Voir aussi [LANC88]. Les problèmes à résoudre sont : la profondeur de l'analyse syntaxique à mettre en œuvre, la recherche des sous-unités syntagmatiques sémantiquement pertinentes pour représenter le document - les descripteurs - et l'expansion du vocabulaire d'indexation par paraphrasage.

---

[1] Le taux de précision est défini comme le rapport du nombre de documents pertinents retrouvés sur le nombre de documents retrouvés.

[2] Le taux de rappel est défini comme le rapport du nombre de documents pertinents retrouvés sur le nombre de documents pertinents de la collection.

En première approximation, on peut dire que les techniques linguistiques substituent à la problématique de l'indexation par la forme des techniques statistiques, celle de l'indexation par le sens.

I.1.3 Le sens

La représentation du sens des documents textuels est le problème clé de l'indexation.

On assiste à plusieurs types de tentatives dans l'univers documentaire.

Les tentatives strictement linguistiques s'opposent à celles qui font appel, pour représenter le sens, à des objets de type conceptuel dont la validité n'est pas d'ordre linguistique; ce sont les modes de représentation des connaissances de l'intelligence artificielle. Autrement dit on peut opposer l'approche linguistique de la représentation du sens à l'approche intelligence artificielle.

Parmi les approches linguistiques, mentionnons tout d'abord l'approche pratiquée depuis près de 20 ans par N.Sager [SAG78] et connue sous le nom de Linguistic String Project. Les textes sont ici des textes scientifiques et techniques, et en particulier des compte-rendus d'hospitalisation. Les traitements linguistiques aboutissent à pouvoir convertir les énoncés en instances d'un nombre réduit de schémas d'énoncé. Ceux-ci donnent lieu à une formalisation sous forme de tables relationnelles appelées "information formats" (formats d'information). La faculté de pouvoir réduire la variété des énoncés à un petit nombre de "formats d'information" est due à un certain nombre de caractéristiques des textes techniques : homogénéité sémantique des classes d'équivalence syntaxique et récurrence d'un petit nombre de schémas canoniques d'énoncés.
Cette approche est possible sur tout document homogène, pas seulement des compte-rendus d'hospitalisation. Elle a par exemple été mise en œuvre sur des dépêches boursières, des bulletins météorologiques (dans une application de traduction semi-automatique). C'est l'approche dite des sous-langages [KITT82].
Une autre approche linguistique est celle qui consiste à représenter la sémantique des énoncés à l'aide de la logique. On procède aux analyses lexicale, morphologique et syntaxique. La représentation syntaxique permet de dériver une représentation sémantique sous forme logique. Les efforts entrepris ici sont ceux de l'informatique linguistique. (Voir à ce sujet [COUL86]). Il règne beaucoup d'incertitudes sur les modèles logiques à adopter. La sémantique des langues naturelles est un champ actif de la recherche en linguistique. Mentionnons entre autres les travaux suivants : [KAYS84],

[ZARR88a], [HOBB82]. Il ne semble pas raisonnable à l'heure actuelle d'escompter des résultats lorsque les textes sont longs et nombreux.

De l'approche linguistique, nous passons insensiblement à l'approche intelligence artificielle. La représentation et la modélisation des connaissances est l'objet de cette discipline. A côté de la logique comme mode de représentation des connaissances [SIMM87], d'autres formalismes ont été élaborés : frames, réseaux sémantiques (ASK de Belkin), scripts [SCHAN81]. Les bases de connaissance obtenues sont constituées de concepts entre lesquels divers types de relation sans statut linguistique (- c'est le cas par exemple des relations ISA ou PART OF-) sont établies.(Pour une revue remarquable, voir [KAYS84]). Si la traduction des phrases d'un texte en forme logique est chose difficile, l'extraction automatique de la représentation d'une phrase dans les termes d'une Base de Connaissances l'est tout autant. Elle est certainement inenvisageable actuellement comme technique opérationnelle de l'informatique documentaire à cause de l'ordre de grandeur des textes à traiter. Jusqu'ici, les textes que l'intelligence artificielle a réussi à traiter sont courts et traitent de domaines restreints. Les systèmes concernés sont souvent appelés systèmes questions-réponses.

Cependant, une tentative intéressante est à mentionner en ce domaine. C'est celle de De Jong et de son système FRUMP [SCHAN81]. FRUMP recherche dans les textes - des dépêches d'agence de presse - la manifestation de "canevas de scripts" ("sketchy scripts") d'événements qu'il a déjà en mémoire. Un tel système est viable parce que l'analyse linguistique est relativement pauvre et que la représentation à obtenir est déjà en partie connue. On pourrait dire que les "canevas de scripts" constituent de nouveaux types d'objet d'indexation plus élaborés que les mots-clés; ce sont des configurations de mots-clés. Bien entendu les traitements dont sont passibles les scripts ont un pouvoir sans commune mesure avec celui dont sont passibles les mots-clés. Ce pouvoir, c'est celui de l'inférence et donc la possibilité de répondre à des questions sur le sens sans que soit stockée au préalable de façon explicite ladite réponse ou que soit déclenché un mécanisme de paraphrasage. L'entreprise n'est viable, là encore que sur des textes restreints quant au domaine couvert.

A mi-chemin entre l'approche Sager et celle de De Jong, se situe le système HAVANE-CALIN de l'IRISA [BOSC86]. Il s'agit ici de gérer des petites annonces de logement. Le système est capable d'extraire du texte de la petite annonce une représentation qui permettra de répondre à une requête émise par un utilisateur. Le système peut traiter des textes à ordre libre et s'appuie sur une grammaire et une représentation sémantique issues d'une analyse linguistique poussée des textes à traiter.

On le voit, la représentation de textes par des structures plus élaborées que les simples mots-clés n'apparaît aujourd'hui viable que sur des textes techniques, c'est-à-dire des textes traitant de domaines restreints. Prendre sérieusement en compte les caractéristiques langagières des textes à traiter est une des clés du succès en documentation automatique. Les systèmes dont il vient d'être question ouvrent à notre avis de nouvelles perspectives en informatique documentaire.

Nous nous sommes intéressé jusqu'ici aux moyens de représenter le contenu sémantique des textes en considérant implicitement que la représentation du contenu sémantique d'une base de données textuelle devait être la somme des représentations individuelles de chacun des textes qui la constitue. Or ceci est loin d'être évident. Le développement historique de l'informatique documentaire nous invite à considérer les choses autrement.

I.2 REPRESENTER UNE BASE DE TEXTES

En effet, que ce soit en documentation manuelle ou en documentation automatique, les outils de la représentation du contenu considèrent la **représentation d'un document particulier comme une fonction de l'ensemble des documents contenus dans la base.**

Il en est ainsi du thesaurus. [SALT86a] considère que la fonction du thesaurus est de normaliser le vocabulaire des documents. A notre avis, la normalisation n'est qu'un effet de l'objectif plus général qui est de représenter les documents non pas de façon atomique, un par un, mais en les situant les uns par rapport aux autres. **Le thesaurus est un outil forgé pour la représentation de la base.** Il caractérise lexicalement la sémantique de la base. La sémantique de chacun des documents est une valeur particulière de cette base sémantique globale.
Il est même des cas où l'élaboration du thesaurus répond au souci de constituer une passerelle entre l'univers des documents et l'univers des utilisateurs de la base. C'est une sorte de compromis entre les lecteurs et les textes, une sorte de langage pivot. On est loin de la seule normalisation, nécessairement réductrice, artificielle et arbitraire.
En tout état de cause, le caractère fini de la liste qu'est avant tout un thesaurus donne à penser que les valeurs sémantiques de chacun de ses éléments se délimitent les unes par rapport aux autres. C'est bien la conception structuraliste du sens. Le sens d'un mot-clé est bien une fonction de la base de documents et non de chacun des documents auxquels ce mot-clé est affecté.

Les approches statistiques ont bien ce mérite de prendre en compte la totalité des documents de la base pour représenter le contenu de chacun d'eux. En effet, ces approches éliminent totalement le sens des termes d'indexation pour ne considérer que leur valeur. Chacun des "mots" du répertoire général des mots extraits de l'ensemble des documents se voit assigner une valeur, un poids représentant la valeur discriminative de chaque mot. Un terme porte d'autant plus d'information qu'il est fréquent dans un petit nombre de textes et rare dans le reste de la collection. La mesure retenue en général pour assigner le poids de chaque mot est : tf * idf où tf=fréquence du terme dans le document et idf, l'inverse de la fréquence relative du terme dans la collection.[3]. Cette mesure est assez proche du rapport signal/bruit de la théorie de l'information. "Le ratio signal - bruit...présente des propriétés dans une certaine mesure semblables à celles du facteur de la fréquence relative inverse..." [SALT89 p. 281].

Le débat entre indexation manuelle et indexation automatique est périodiquement relancé. Un récent article [BLAI85] a émis à l'encontre de l'indexation automatique les critiques suivantes issues d'une expérimentation menée sur une grande quantité de documents :

- performances médiocres lors de la recherche d'informations (taux de rappel[4] = 20%; taux de pertinence[5] = 80%),
- difficulté à améliorer ces performances car une augmentation du taux de rappel aboutit à retrouver un trop grand nombre de documents ce qui rend l'affinement ultérieur de la requête ingérable,
- tâches de vérification orthographique ou typographique beaucoup plus lourdes qu'en indexation manuelle parce que les représentants du texte (=les mots) sont beaucoup plus nombreux.

[SALT86b] montre que ni la première ni la seconde critique ne sont fondées et reprend l'énoncé des inconvénients de l'indexation manuelle énumérés par [CLEV77] :

- 60% seulement de mots-clés sont communs à deux sujets ou groupes de sujets chargés d'élaborer un thesaurus pour une même base.
- 30 % seulement de mots-clés sont communs à deux sujets ou groupes de sujets chargés d'indexer un même document à l'aide du même thesaurus
- et enfin, 40 % seulement de résultats sont communs à deux sujets ou groupes de sujets chargés de conduire la même recherche sur la même base.

---

[3] idf est égale à log N/n avec N= nombre de documents de la collection et n= nombre de documents contenant le terme.

[4] Rappelons que le taux de rappel est égal à: nombre de documents pertinents retrouvés / nombre de documents pertinents de la collection

[5] Rappelons que le taux de précision est égal à: nombre de documents pertinents retrouvés / nombre de documents retrouvés

A ce jour, d'après [SALT86b], l'évaluation la plus complète des deux méthodes de représentation du contenu est [CLEV66].

I.3 LA REPRESENTATION DU SENS : SENS COMMUN VS VALEUR

On le voit, à ce jour, la représentation du contenu des documents d'une base de données textuelles est partagée, et le demeure, entre deux conceptions du sens : l'une, l'approche statistique, est indissociable de la notion de valeur, l'autre, qu'elle soit traditionnelle (mots-clés) ou plus récente (intelligence artificielle) met en jeu des concepts, concepts appartenant à une sémantique générale dont on voit mal comment elle pourrait reposer sur autre chose que le sens commun. Entre 1970 et 1980, les progrès de la recherche ont surtout concerné l'approche statistique et probabiliste. La situation est en train de changer. Ainsi que le dit [CROF87 p. 390], les techniques statistiques ayant atteint leur plafond de performance, les chercheurs explorent aujourd'hui l'application des techniques de traitement automatique du langage et de représentation des connaissances. Les années 80 sont donc marquées par le développement des recherches concernant l'utilisation des résultats de l'informatique linguistique, de l'intelligence artificielle et, au-delà, de la recherche cognitive sur les apports de laquelle nous nous attarderons davantage dans la partie suivante.

I.4 LES DOCUMENTS : TEXTES, "INFORMATION FORMATS", BASES DE CONNAISSANCES, HYPERTEXTES

Il est un type d'impact de ces techniques sur lequel nous voudrions insister, c'est celui qui est exercé sur la nature des données elles-mêmes. En effet, le fait qu'il soit si difficile, comme on l'a vu en I.1.3, d'extraire automatiquement les éléments d'une base de connaissances à partir de textes conduit :

- à constituer directement des bases de connaissances sans passer par une expression textuelle de celles-ci,
- à envisager des traitements spécifiques à telle ou telle catégorie de textes,
- à structurer la collection de textes.

Quelques mots sur chacune de ces trois tendances.

Lorsque le domaine de savoir est très circonscrit et formalisé, il devient envisageable d'exprimer les connaissances directement dans le formalisme de la base de connaissances d'un système expert par exemple. C'est le cas par exemple du projet GRANT [COHE87] où des connaissances ayant trait à la nature de projets susceptibles d'être financés par tel

ou tel organisme sont exprimées sans médiation textuelle. C'est le cas aussi de [ZARR88a] qui dénomme la base de connaissances manipulée,"méta-document".. Les exemples se multiplient.

La présence de sous-langages conduit à des traitements spécifiques qui permettent de représenter le contenu des textes comme une collection de tuples ou de frames dont l'obtention est, toute mesure gardée, relativement aisée [SAG78], [KITT82], [BOSC86]. Le système FRUMP illustre la possibilité d'un traitement frustre mais sélectif des textes aboutissant à une représentation sophistiquée et donc riche en capacité inférentielle.

Les systèmes d'hypertexte et d'hypermedia [CONK87] auxquels nous consacrerons ultérieurement un développement particulier sont les vecteurs d'environnements d'information nouveaux dont la richesse se laisse déjà apercevoir au travers d'un certain nombre d'applications : encyclopédies, bornes d'information par exemple.

Enfin, les applications bibliographiques qui ont été le moteur des recherches se voient relayées dans ce rôle par l'importance prise par les applications bureautiques.

II ACCES A L'INFORMATION

Dans cette partie, nous ne nous intéresserons pas aux architectures matérielles dédiées aux traitements documentaires. Voir par exemple à ce propos [JIME88].

Quant aux techniques d'implémentation, disons simplement que de nouvelles techniques d'accès aux fichiers tendent à entrer en compétition avec les traditionnelles méthodes d'accès par fichiers inverses; ce sont les fichiers signés. La représentation signée d'un document textuel est une chaîne de bits de longueur fixe dont chacun des bits est positionné en calculant les valeurs d'une fonction de hashing sur les mots-clés du document en question. La recherche se fait d'abord en comparant la signature de la requête avec les signatures des documents de la base; elle s'affine ensuite en comparant les caractères eux-mêmes de la requête avec ceux de l'ensemble des documents dont la signature s'appariait avec celle de la requête [SALT89], [BERT88], [CHRI84], [FALOU87]. Un attrait majeur de cette technique, valable en particulier pour les CD-ROM, est que l'ajout d'un document ne se traduit pas par une réorganisation complète du fichier d'indexation. [POGU87] examine une implémentation parallèle de cette technique.

Nous nous intéresserons dans cette partie à l'évolution et aux tendances manifestées dans l'appariement entre requêtes et représentants des textes d'une base de données. Ensuite, nous essaierons de montrer de quelles évolutions, les outils de l'appariement, c'est-à-dire les interfaces, ont fait l'objet ces dernières années. Nous consacrons enfin quelques pages à certains systèmes expérimentaux qui nous semblent exemplaires des tendances actuelles de la recherche.

II.1 LES MODES D'APPARIEMENT ENTRE REQUETES ET DOCUMENTS

II.1.1 Apparier des formes

II.1.1.1 Appariement approché I : booléen étendu

La technique traditionnelle de recherche d'information est la recherche des solutions d'une équation booléenne de descripteurs.

C'est la technique qui prévaut très largement dans les systèmes commerciaux opérationnels. [BELK87] trouve à la pérennité de cette situation les raisons suivantes :

- l'investissement réalisé dans ces systèmes est si considérable que les modifier ne serait pas économiquement viable,

- les techniques alternatives n'ont pas été testées dans des environnements en grandeur réelle,
- les résultats obtenus par des techniques alternatives ne sont pas suffisamment supérieurs, même au niveau expérimental, pour justifier les changements,
- les structures des requêtes booléennes expriment des aspects importants des besoins des utilisateurs.

Les inconvénients de cette technique sont bien connus [BELK87] :

- ne sont pas retrouvés de nombreux textes pertinents dont la représentation ne correspond qu'approximativement à la requête,
- les textes retrouvés ne sont pas ordonnés selon leur degré de pertinence,
- l'importance relative des concepts à l'intérieur de la requête ou à l'intérieur du texte n'est pas prise en compte,
- la formulation des requêtes est complexe,
- le résultat est fonction des deux représentations comparées qui doivent faire appel au même vocabulaire.

La rigidité de cette technique avait déjà été assouplie par l'utilisation des jokers et autres masques ainsi que par la mise en œuvre plus ou moins automatique du thesaurus pour expanser une requête.

Les techniques statistiques ne présentent pas ces inconvénients : l'appariement entre requête et documents peut être plus ou moins strict, l'importance relative des documents et des concepts est prise en compte, des règles rationnelles sont mises en œuvre pour ordonner les documents retrouvés, les résultats sont meilleurs ou au moins comparables.

C'est pour remédier aux inconvénients des techniques booléennes de façon économiquement viable qu'une technique dite "booléenne étendue" a été mise au point [SALT83], [SALT85]. C'est un cas particulier des techniques statistiques et probabilistes[6]. "It is known that the extended boolean system provides superior retrieval effectiveness" [SALTON86a p. 8]. Le principe est d'une part, de conférer aux termes de recherche de l'équation booléenne des poids et d'autre part, d'interpréter les opérateurs de l'équation comme des distances entre requêtes et documents.

---

[6] Notre propos étant de confronter techniques d'appariement strict et techniques d'appariement approché, nous ne rentrons pas dans le détail des méthodes statistiques. Nous ne nous attardons pas non plus sur les différences qui existent entre modèle statistique et modèle probabiliste. Ce dernier modèle est au coeur du système SPIRIT [FLUH80]. Pour l'heure, nous nous contentons de citer [SALT86 p. 4].:."The probabilistic retrieval model is attractive because it provides a theoretical foundation for the retrieval operation which takes into account the notion of document relevance. ... The probabilistic model offers justification for various methods that had previously been used in automatic retrieval environments on an empirical basis.[for example the inverse document frequency (idf) term weighting system]."

Il s'agit là d'un compromis entre une technique d'appariement exact, la technique booléenne, et une technique d'appariement approché, la technique statistique, qui apparaît industriellement crédible. La recherche a démontré la supériorité des techniques d'appariement approché. Aux techniques statistiques vectorielles exposées en détail par [SALT89] sont venues s'ajouter d'autres techniques d'appariement approché, les techniques de clusterisation.

II.1.1.2 Appariement approché 2 : clusters

L'hypothèse de clusterisation émise en premier lieu par [JARD71] est que des documents très semblables les uns aux autres sont susceptibles d'être davantage pertinents pour une même requête que des documents ne témoignant pas du même degré de similitude entre eux. La clusterisation est une technique de classification qui permet d'identifier et de constituer des groupes, des clusters d'objets, ici de documents. En confrontant une requête aux clusters - à dire vrai à un représentant abstrait de ceux-ci - plutôt qu'à chaque document individuellement, on retrouve plus facilement et plus rapidement les documents pertinents. Le degré de similitude inter-document est calculé sur la base des descripteurs des documents. Ces descripteurs peuvent avoir été attribués manuellement ou avoir été obtenus par indexation automatique. De nombreux travaux actuels portent sur les mérites comparés de diverses techniques de clusterisation [VOOR86], [WILL87], [PANY87], [ELHA87].

Il faut remarquer que le principe de ces techniques de classification est mis à profit dans l'univers des hypertextes. Utilisées conjointement avec les techniques de "relevance feedback" sur lesquelles nous reviendrons en II.2.2.1 plus loin, elles permettent de créer ce qu'il est convenu d'appeler dans cet univers là, des liens dynamiques entre documents [DEBI88].

L'utilisation conjointe des techniques statistiques et des techniques de clusterisation est souhaitable car toutes deux permettent de retrouver des ensembles de documents qui peuvent être différents.
"La version actuelle du système [I3R] met en œuvre deux stratégies statistiques de recherche de l'information. La stratégie principale repose sur le modèle probabiliste... La seconde s'appuie sur les clusters. La justification pour la mise en œuvre de cette stratégie supplémentaire est fournie par des expériences qui ont montré que les deux stratégies tendent à retrouver des ensembles différents de documents et, que, en particulier, la stratégie fondée sur les clusters parvient à retrouver des documents là où la recherche probabiliste échoue" [CROF87 p. 391].

II.1.2 Apparier du sens

Jusqu'ici, nous avons considéré les techniques de recherche d'information qui s'inscrivaient dans ce que nous avons appelé, en I.3, le paradigme de la valeur. Quelles sont les techniques de recherche d'information lorsque la base de données est, ou est représentée par, une base de connaissances ?

II.1.2.1 Paraphrasage

Tout d'abord, il faut constater que la plupart des systèmes concernés ont une composante de traitement du langage naturel. On ne s'étonnera donc pas de les voir tirer parti de leurs capacités en ce domaine. C'est la fonctionnalité de paraphrasage qui est surtout invoquée tant pour ce qui est des requêtes que pour ce qui est des éléments d'indexation. Les paraphrases générées par transformation lexicale et/ou syntaxique démultiplient les possibilités d'appariement formel entre requêtes et documents. S'appuyant sur des connaissances purement linguistiques, donc indépendantes du domaine d'application, les transformations sont générales. Elles ont par ailleurs la propriété de ne pas modifier le sens des objets sur lesquels elles s'appliquent. Les transformations paraphrastiques sont un des piliers de la linguistique moderne, ce qui leur confère une validité scientifique sérieuse. Elles sont mises en œuvre depuis longtemps dans les systèmes question-réponse dont la problématique est proche, comme nous l'avons fait remarquer plus haut, de celle des systèmes documentaires [FOUR84]. Les systèmes décrits par [BASS86], [FLUH85], [DEBI88], [ROUAU87], y font appel.

Ces techniques de paraphrasage manipulent des représentations assez proches malgré tout, de la surface langagière et n'invoquent pas un niveau de représentation des connaissances plus profond d'ordre conceptuel. Aussi les systèmes dont il vient d'être question articulent-ils la gestion d'opérations langagières avec des opérations de manipulation des connaissances proprement dites. Cette articulation est théorisée par [COUL82] sous la dénomination de processus de compréhension à profondeur variable.

II.1.2.2 Inférence

Inférence est le nom donné par l'intelligence artificielle à la manipulation des connaissances.
Les opérations concrètes que ce nom recouvre dépendent du formalisme adopté pour représenter les connaissances. S'agissant d'un réseau sémantique, une opération dite d'"activation par proximité" (anglais : "spreading activation") est souvent utilisée en informatique documentaire. Une question est utilisée pour activer les parties du réseau

sémantique qui décrit le contenu des documents. Dans les réseaux simples, les nœuds représentent les termes d'indexation qui sont connectés aux documents. Dans les réseaux plus riches en connaissances, les nœuds et les arcs représentent des concepts du domaine, les relations qu'ils entretiennent entre eux et les documents sur lesquels ils pointent. Une fois les nœuds de départ activés, l'activation se propage vers les autres nœuds en suivant les liens établis et en respectant certaines contraintes. Dans le système GRANT décrit par [COHE87], les contraintes sont de trois types : contrainte de distance (l'activation cesse au 5ème nœud), contrainte de branchement (l'activation est interrompue aux nœuds dont sont issus un trop grand nombre d'arcs), contrainte de chemin (l'activation est sensible à des "positive or négative path endorsements" qui représentent des meta-connaissances sur les cheminements privilégiés dans le réseau en fonction du domaine représenté). Selon les auteurs, l'"activation par proximité sous contrainte" obtient, avec des scores raisonnables de rappel et de précision, [des résultats], étayés sur des relations sémantiques, qui n'auraient pu être obtenus au moyen d'une simple recherche par mots-clés" [COHE87 p. 267].

II.1.3 Le néo-connexionisme

La recherche en intelligence artificielle et au-delà en informatique générale est traversée depuis quelques années par un courant "sub-symbolique". Ce courant appelé néo-connexioniste explore la simulation de réseaux neuronaux par machine [RUM86]. Embryonnaire encore, il inspire quelques recherches en informatique documentaire, par exemple [BRAC88]. Il est vraisemblable pourtant que celles-ci connaîtront un développement important parce qu'une certaine convergence existe avec les traditionnelles méthodes statistiques et les techniques de "relevance feedback" esquissées en II.2.2.1 plus loin.

II.2 LES INTERFACES

II.2.1 Les interfaces "conviviales"

Jusqu'aux années 80, la recherche a porté essentiellement sur l'ergonomie des interfaces, l'objectif étant de rendre l'interaction de l'utilisateur avec le système documentaire plus souple, plus agréable. La problématique générale était bien de mettre en œuvre un outil, le plus puissant possible, permettant de projeter l'utilisateur, ou plutôt sa requête, sur la base ou les bases. Les interfaces graphiques, les interfaces en langage naturel répondaient à ce souci.

Les interfaces en langage naturel se sont sophistiquées en incorporant un niveau de compréhension qui leur permet d'interpréter le **sens** de la question formulée en langage naturel, de sélectionner la base appropriée, (souvent, en effet, les interfaces en question sont mises en œuvre dans un contexte multi-base), de construire une première stratégie de recherche qui est ensuite modifiée en fonction des résultats obtenus (généralement en fonction du nombre de documents retrouvés). Le système EURISKO [BART87], [BART88], qui met en œuvre des stratégies de planification très élaborées, est un exemple de ce type de système.

II.2.2 Les systèmes multi-experts ou hybrides : les aides à l'activité de recherche

Une deuxième génération de systèmes intelligents de recherche d'informations est en train de voir le jour. Ces systèmes prennent davantage en compte **la situation de recherche d'informations**, tirant en cela partie des recherches cognitives qui conduisent à concevoir **des systèmes qui modélisent les comportements** observés **de l'utilisateur**, ici le demandeur d'information et/ou le documentaliste [DANI86]. Dans le système GRUNDY [RICH79], par exemple, la modélisation des utilisateurs résulte d'un calcul sur des stéréotypes déclenché lors des énoncés que produit l'utilisateur. Tous ces systèmes sont nécessairement complexes. Ils incorporent des types de connaissances multiples dépassant très largement les seules connaissances du domaine invoquées dans les systèmes de première génération. Ces systèmes sont dénommés systèmes multi-experts ou encore systèmes hybrides.

Avant d'exposer plus en détail les types d'expertise incorporés dans les systèmes, nous voudrions attirer l'attention sur le fait qu'un certain nombre des modules de ces systèmes implémentent des techniques éprouvées. Ainsi, les techniques statistiques ou probabilistes sont bien entendu mises à contribution - mais pas seules - voir à ce sujet l'exemple déjà mentionné de I3R [CROF87] qui utilise aussi les techniques de clusterisation.

II. 2. 2. 1 Mettre en œuvre ensemble des outils et des techniques éprouvés

D'autres techniques bien éprouvées et qui, de plus ont l'avantage de correspondre à des situations concrètes de recherche d'information sont également mises en œuvre : relevance feedback, co-citation, browsing, cluster de termes.

[PEDE87] distingue trois types de besoins des utilisateurs :

- besoins de vérification : "l'utilisateur veut vérifier ou retrouver de l'information sur des éléments d'information aux caractéristiques connues".
- besoins conscients concernant un sujet : "L'utilisateur veut clarifier, passer en revue ou approfondir certains aspects d'un sujet bien connu".
- besoins flous concernant un sujet : "L'utilisateur veut explorer de nouveaux concepts sur des sujets non connus".

Nous allons à présent passer en revue les techniques évoquées plus haut en montrant comment chacune permet de répondre à l'un ou l'autre des types de besoins énumérés par [PEDE87].

"Browsing"

C'est le troisième type de besoin que visent à satisfaire les techniques de **browsing**, de **navigation**, techniques qui constituent le mode typique d'accès à l'information dans les hypertextes.

Oddy [ODDY77] fonde son programme THOMAS sur cette situation de recherche d'information dans laquelle un utilisateur n'est pas à même de formuler une requête précise mais sera capable de reconnaître ce qu'il cherchait en le voyant. Le programme épargne à l'utilisateur qui, au démarrage de sa recherche n'est pas à même d'exprimer précisément son besoin, la nécessité de formuler d'entrée de jeu une requête précise et lui donne la possibilité, lorsqu'il découvre peu à peu qu'il sait ce qu'il veut, de formuler celle-ci. Les éléments de la réalisation sont décrits ainsi : "[Le programme THOMAS] se forge une représentation du sujet d'intérêt de l'utilisateur, représentation issue du modèle du monde qu'il détient et sélectionne les références à afficher selon l'état de cette représentation qu'il modifie au fur et à mesure des réactions de l'utilisateur aux références affichées" [ODDY77 p. 5].

Le programme essaie donc de se comporter comme un documentaliste ou un bibliothécaire dans son interaction avec un usager, lequel a, comme dans la réalité, un rôle important à jouer. "The job of the mechanical part of the system is to create a helpful framework within which **the user can make problem-solving decisions**" [ODDY77 p. 12].

Bien qu'ancien, ce système est inspiré par une conception qui le démarque très nettement des systèmes dits conviviaux dont les interfaces intelligentes ne constituent après tout qu'un avatar. "Much of the literature on interactive retrieval systems... is primarily concerned with man-machine interface engineering : for example how to reduce the effort and training required of users and increase the computer's tolerance to their errors, and

how to make users aware of the system's vocabulary or information structures. Software to achieve such ends acts as an interface between a conventional search program... and the user of a terminal. [ ]. **A user's browsing, or exploration of the system is normally controlled by the interface, whereas in the approach I [am describing] it is an integral part of the reference retrieval process**" [ODDY77 p. 3].

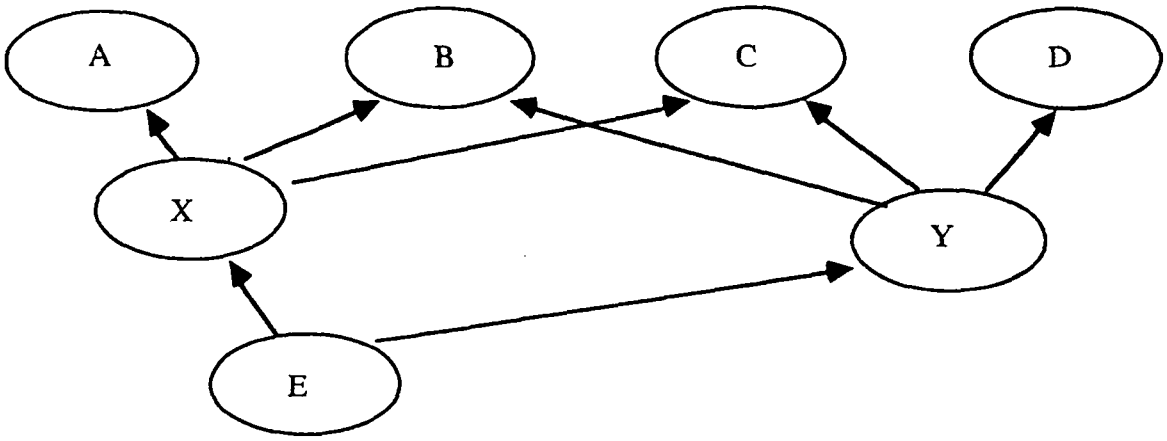
I3R (Intelligent intermediary for information retrieval), un prototype de recherche de système hybride parmi les plus en pointe, comporte un module de browsing. "The I3R browsing facility provides access to the documents and other information in the database, along with commands to navigate along links to related information." [CROF87 p. 392].

"Term clustering"

Un autre type de technique permettant l'exploration d'une base de données est la mise en œuvre de ce que [TURN85], [MICH85], [CALL83] appelle **co-word analysis** et [SALT89], **term clustering**. Il s'agit de rapprocher, les uns des autres, des descripteurs, sur la base de leurs fréquences de co-occurrence. Les ensembles de descripteurs constitués ainsi sur des bases purement statistiques (techniques de classification automatique) constituent des clusters. [SALT89] examine l'utilisation de ces clusters pour constituer automatiquement des thesaurus alors que [MICH85] envisage d'en tirer parti pour donner à l'utilisateur des moyens d'appréhender les thèmes dont traite une base et éventuellement lui permettre de positionner la valeur de tel ou tel descripteur par rapport à d'autres.

Liens de citation

Une recherche d'information, expression des besoins 2 et 3 ci-dessus, peut être guidée par les **liens de citation** bibliographique existants entre les documents qui composent la base de données bibliographique. En effet, les liens de citation entre documents expriment des liens de similitude de sujet entre ces documents. Des mesures de ces liens ont été proposées. On a distingué le "couplage bibliographique" et "la co-citation". Dans le schéma ci-après (figure 2), les documents X et Y ont deux unités de couplage bibliographique parce qu'ils citent tous deux les deux documents B et C alors que la valeur de co-citation entre X et Y est de un parce qu'un seul document, le document E les cite tous deux. Il a été proposé [AMSL72] d'intégrer ces deux dimensions en une seule "mesure de similitude de sujet" qui prend en compte à la fois les descendants d'un document, c'est-à-dire ceux qui le citent et les ascendants de ce document, c'est-à-dire ceux qu'il cite. Cette mesure unique s'est révélée supérieure à chacune des deux autres dans les tests d'accès à l'information [BITCH80].



LIENS DE CITATION ENTRE DOCUMENTS

Figure 2

I3R [CROF87], à nouveau, exploite ce type d'information.

"Relevance feedback"

Enfin, parmi les techniques pré-existantes à l'introduction de l'intelligence artificielle, et compatibles avec elle, la mise en œuvre du "**relevance feedback**" est certainement celle dont l'utilité est unanimement admise. Elle constitue bien une stratégie naturelle de recherche d'information.

Dans un système statistique, le "relevance feedback" est effectué de la manière suivante [SALT86a p. 6] : les résultats d'une première recherche sont utilisés automatiquement pour reformuler la requête en accroissant les poids des termes de la requête qui sont présents dans les documents retrouvés jugés pertinents par l'utilisateur, et à l'inverse, en diminuant les poids des termes de la requête qui sont également présents dans les documents non pertinents retrouvés. La modification des poids peut s'accompagner d'une expansion de la requête à laquelle on adjoint des termes caractérisant les documents pertinents retrouvés.

L'augmentation de la précision est spectaculaire (figure 3).

	Coll.CACM 3204 docs, 52 requêtes					Coll. CISI 1460 docs, 76 requêtes			
Taux de Rappel	initial td*idf	1er feedback sans expansion		itération avec expansion		initial td*idf	1er feedback sans expansion		itération avec expansion
.1	.3641	.3546	(-3)	.5338	(+47) .2476	.3631	(+47)	.3654	(+48)
.3	.2038	.2633	(=29)	.3579	(+76) .1500	.2369	(+58)	.2899	(+93)
.5	.1329	.1743	(+31)	.2804	(+111) .1209	.1729	(+43)	.1647	(+36)
.7	.0732	.1038	(+42)	.1621	(+121) .0753	.1154	(+53)	.1097	(+46)
.9	.0405	.0542	(+35)	.0827	(+104) .0505	.0644	(+28)	.0611	(+21)
Amél. moyen.	+26,6%			+91,8%		+45,8%		+48,8%	

Figure 3. Impact du "Relevance feedback" (tiré de [SALT86a p. 6])

Ainsi donc, la mise en œuvre d'une seule itération de feedback conduit à une augmentation moyenne de la précision de 90% dans un cas et de 50% dans l'autre.

Tous les systèmes hybrides de la génération actuelle ont recours à une forme ou une autre de "relevance feedback".

DIALECT [BASS86] conduit cette opération automatiquement en mettant en œuvre des outils d'analyse linguistique. "La question de l'utilisateur, écrite en langage naturel, est analysée puis utilisée pour extraire un premier noyau de 'zones de texte' très pertinentes. Ces zones sont à leur tour analysées et exploitées en vue d'enrichir la question. Le système s'appuie essentiellement sur des procédures d'analyse distributionnelle permettant de repérer des régularités (syntaxiques) formelles. Cette stratégie générale et autonome est relancée automatiquement jusqu'à l'obtention d'une condition d'arrêt".

Dans SPIRIT [ANDR81], [FLUH80], c'est l'utilisateur qui sélectionne parmi les textes retrouvés, les parties de ceux-ci qui doivent servir de nouveaux points de départ pour la recherche. Il faut remarquer à nouveau que le relevance feedback "établit une liaison dynamique entre des textes ou des parties de texte, liaison très voisine de la notion d'hypertexte dynamique" [DEBI88 p. 351].

La relance d'une recherche s'accompagne donc d'une reformulation de la requête initiale. Le relevance feedback tel qu'il est décrit ci-dessus constitue une façon de reformuler cette requête. A celle-ci s'ajoutent les autres techniques de reformulation : invocation du

thesaurus ou de la base de connaissances sur le domaine, transformations linguistiques morphologiques (lemmatisation), syntagmatiques (troncature ou au contraire constitution de syntagmes) [DEBI88], utilisation des connaissances sémantiques contenues dans un dictionnaire général de langue [FOX80].

Les quatre mécanismes que nous venons d'évoquer -et qui font toujours l'objet d'améliorations - sont en tout ou partie utilisés dans les systèmes de la dernière génération.

II.2.2.2 Quelques prototypes de systèmes hybrides

Aux Etats Unis, les représentants les plus avancés de cette génération de systèmes sont CODER [FOX87] et I3R [CROF87],.tous deux développés par d'anciens élèves de G. Salton qui a pourtant consacré toutes ses recherches à l'approche statistique en informatique documentaire.

En France, c'est le système IOTA, développé à Grenoble, qui apparaît s'inscrire le mieux dans ce courant [CHIA86], [CHIA87], [DEFU86].

IOTA

L'objectif de IOTA est de modéliser le comportement d'un documentaliste réalisant des interrogations. C'est un système-expert en recherche d'information dont les fonctionnalités sont spécifiées à partir de la tâche à modéliser.

Le processus général de traitement d'une requête est le suivant :

- analyse de la requête primaire exprimée en langage naturel dont on extrait les syntagmes nominaux pour élaborer une expression booléenne.
- appariement entre syntagmes nominaux de la requête et termes d'indexation par transformations linguistiques (et non pas sémantiques) pour élaborer la requête finale.
- évaluation du niveau de dégradation de la requête primaire en la comparant avec la requête finale élaborée au pas précédent. Il en découle l'évaluation du niveau de l'utilisateur.
- interprétation de la requête finale : obtention des références. Dans IOTA, les termes d'indexation sont affectés de poids calculés statistiquement.
- évaluation du résultat; c'est avec les étapes suivantes l'étape la plus "cognitive" du système. Nous y revenons plus loin.

- reformulation de la précédente requête finale à partir de l'évaluation. On remplace les concepts par d'autres, tirés du thesaurus, et/ou on modifie les opérateurs de l'équation booléenne.
- apprentissage des inférences réalisées.

Selon les auteurs, les deux aspects les plus intéressants du système sont : l'élaboration de la typologie des utilisateurs et l'évaluation des réponses (qui est liée à la précédente).

Un utilisateur est caractérisé comme spécialiste, moyen ou débutant. Cette caractérisation est fonction de l'expertise de l'utilisateur par rapport au domaine (comparaison de ses concepts avec ceux du thesaurus en généricité/spécificité) et fonction, d'autre part, de son expertise dans la connaissance du système (niveau de dégradation de sa requête primaire pour élaborer la requête finale).

Les réponses sont évaluées en fonction de l'utilisateur qui a soumis la requête - pour un débutant, une réponse très spécifique n'est pas particulièrement bonne -, en fonction de la pertinence des réponses - rappelons que les descripteurs sont affectés de poids -, en fonction du nombre de références retrouvées, et enfin en fonction du degré de dégradation de la requête.

On l'aura remarqué, l'objectif du système est avant tout de satisfaire le besoin d'information de l'utilisateur. La réponse est bonne si elle retrouve les documents pertinents certes, mais aussi et tout autant, si ces documents sont adaptés à l'usager. C'est un tel objectif que [CROF87] assigne aux systèmes documentaires :"the goal of a document retrieval system is to retrieve documents that are **relevant** to a **particular user**'s query" [CROF87 p. 389].
Satisfaire cet objectif implique la nécessité de faire coopérer des expertises multiples.

CODER [FOX87]

Ainsi CODER (Composite Document Expert/Extended/Effective Retrieval) met en œuvre des experts, pour la construction des modèles de l'utilisateur et de la requête, pour l'indexation, pour la recherche, pour le browsing, pour les explications données à l'utilisateur, et pour la gestion du thesaurus. Les ressources à utiliser sont multiples. "The plan for CODER was to ultimately use logic programming, blackboard-based expert systems methods, natural language processing, several knowledge representation schemes, heuristics for applying search methods, planning of resource utilization, temporal reasoning, and user modeling" [FOX87 p. 350].

La coopération entre ces différents experts est assurée par la technique du tableau noir. Cette structure souple de contrôle est d'autant plus justifiée, même si elle est loin d'être exempte de problèmes, que le nombre d'experts est grand.

I3R [CROF87]

Ces experts sont au nombre de six dans I3R! Le maître d'œuvre du système les caractérise ainsi :

Le "user model builder" rassemble des informations sur l'utilisateur pour déterminer si un stéréotype particulier s'applique. Les stéréotypes conditionnent le type d'interaction, les objectifs de la session. L'identification d'un stéréotype est obtenue à partir d'informations recueillies auprès de l'utilisateur, par exemple son intérêt pour un taux de rappel élevé. L'autre fonction majeure de cet expert est d'acquérir des connaissances sur le domaine d'intérêt de l'utilisateur.

Le "request model builder" construit un modèle du besoin d'information de l'utilisateur. L'essentiel de cette tâche est d'obtenir une requête initiale de la part de l'utilisateur et d'opérer une indexation statistique simple sur celle-ci pour obtenir les termes d'indexation avec leurs poids. Les requêtes peuvent être exprimées en langage naturel, sous forme de mots reliés par des opérateurs booléens, ou sous la forme de mots et de syntagmes auxquels sont associés des poids. Cet expert recueille également les jugements de pertinence sur les documents rappelés ainsi que des informations sur le contenu des documents pertinents.

Le "domain knowledge expert" utilise et les connaissances du domaine tirées du modèle utilisateur et la base de connaissances pour inférer les concepts reliés à la requête initiale. Ces concepts sont soumis à l'utilisateur pour évaluation et éventuelle inclusion dans le "request model".

Le "search controller" choisit et exécute les stratégies formelles de recherche. Ces stratégies sont fondées sur le modèle probabiliste et sur les techniques de clusterisation. Les documents retrouvés sont placés dans le "request model".

Le "browsing expert" procure à l'utilisateur une méthode informelle de trouver des documents pertinents en naviguant dans la base de connaissances. Le processus de browsing peut partir d'un document, d'un auteur, d'un terme d'indexation et suivre les liens vers d'autres objets de la base de connaissances.

L'"explainer" a pour tâche de fournir des explications à l'utilisateur sur sa demande.

La base de connaissances contient les entités : les documents, les termes, les concepts, les auteurs, les utilisateurs et les sessions. Les entités possèdent des propriétés : pour un document, son titre et son résumé, par exemple. Il existe des relations entre entités : statistiques entre documents et entre termes (mesures de similitude), bibliographiques entre documents et auteurs, sémantiques entre concepts; Entre concepts et utilisateurs, les relations expriment la connaissance qu'a l'utilisateur du domaine.

Les experts peuvent agir indépendamment et communiquer par l'intermédiaire du tableau. Un contrôleur règle l'activation des experts. Il utilise un plan et un agenda. L'agenda contient les actions que les divers experts peuvent entreprendre dans telle ou telle situation. Le plan permet de règler les priorités dans l'activation des experts. La communication avec l'utilisateur se fait par la médiation d'un "interface manager".

Le traitement d'une requête se fait en trois temps : formulation et affinage de la requête, recherche puis évaluation. Lors de cette dernière phase, l'utilisateur examine les documents rappelés, identifie ceux ou les parties de ceux, telles les syntagmes, qu'il estime importants. Cette information est utilisée pour mettre à jour le "request model".

Nous avons développé assez longuement les traits de ce système. Ce prototype, de par ses ambitions cognitives au service desquelles il met toute une gamme d'outils et de techniques, incarne de façon privilégiée, à notre avis, les avancées actuelles de la recherche en informatique documentaire. Est-il trop ambitieux?

III CONCLUSION

Ce survol des recherches en informatique documentaire a abordé essentiellement deux thèmes : la représentation du contenu des documents d'une part, l'accès à l'information d'autre part.
Les documents que nous avons envisagés sont essentiellement les documents textuels.

Les deux approches traditionnelles de la représentation du contenu des documents ont été confrontées : statistiques d'un côté, linguistico-conceptuelles de l'autre. Il est apparu que cette opposition d'approche recouvrait en partie une dichotomie qui opposait sens et valeur. Nous avons tenté d'établir que toute représentation du sens d'un document particulier d'une base était une fonction de l'ensemble des documents qui constituaient ladite base et que, par conséquent, le concept de valeur constituait une dimension du sens. C'est ce que semble ignorer l'approche linguistico-conceptuelle. L'approche statistique semble, elle, de fait, ne conférer un statut représentationnel qu'à cette seule valeur.

Il nous est apparu que l'évolution de la recherche sur les modes de représentation du contenu allait de pair avec une diversification des types de documents électroniques : textes, "information formats", bases de connaissances, hypertextes et hypermedia.

S'agissant de l'accès à l'information, les dernières années ont vu la problématique de l'appariement d'une requête à un ensemble de documents se positionner progressivement à l'intérieur d'un cadre plus vaste, celui de la satisfaction du besoin d'information de l'utilisateur. La recherche d'information est devenu alors un problème de recherche cognitive.

Nous avons d'abord passé en revue les techniques d'appariement. Nous avons distingué celles qui visaient à apparier des formes et celles qui visaient à apparier des sens. Les avancées réalisées pour les premières sont le booléen étendu et la clusterisation essentiellement. Les secondes ont vu la mise en œuvre de processus de compréhension à profondeur variable : paraphrasage et inférence.
Nous avons ensuite montré comment l'on était passé, par la prise en compte des recherches cognitives, des interfaces conviviales à des systèmes hybrides ou multi-experts.

Ces derniers font coopérer des outils, pour certains, déjà éprouvés : "relevance feedback", liens de co-citations, "browsing", "co-word analysis".

Mais l'objectif qu'ils visent à satisfaire a évolué. Au lieu d'aider à projeter une requête anonyme sur une base de données, les prototypes actuels se donnent pour objectif de satisfaire le besoin d'information d'un utilisateur. Cela implique deux choses : nécessité de prendre en compte les caractéristiques propres de l'utilisateur d'une part, et d'autre part, nécessité de renverser la perspective en considérant qu'une base de données ne fait que stocker des données, données dont c'est à l'utilisateur lui-même de décider lesquelles constituent les informations qu'il recherche.

On a renoncé à l'espoir de fournir à l'utilisateur **LA** réponse à **LA** question posée. **Une réponse, c'est aujourd'hui un document ou un ensemble de documents que l'utilisateur estimera pertinent** (s). Cela se traduit par des systèmes qui visent moins à se substituer à lui qu'à mettre à sa disposition toute une panoplie d'outils, au premier rang desquels les moyens d'une interactivité élaborée.

BIBLIOGRAPHIE

[AMSL72] Amsler, R. (1972). Applications of citation-based automatic classification. Internal technical report n° 72-12. Linguistics research center, the university of Texas at Austin, Austin, TX.

[ANDR81] Andreewsky, A., Binquet, J. P., Debili, F., Fluhr, C. & Pouderoux, B. (1981). Linguistic and statistical processing of texts and its application in the field of legal documentation. 6th symposium on legal data processing in Europe. Council of Europe, Thessaloniki.

[ANTO88] Antoniadis, G., Lallich-Boidin, G., Polity, Y. & Rouault, J.A. (1988). French text recognition model for information retrieval systems. In Y. Chiaramella, (Ed.), Proceedings of the 11th international conference on research and development in information retrieval, (pp 67-84). June 13-15, Grenoble, France. Grenoble : Presses universitaires.

[BART87] Barthes, C., Frontin, J. & Glize, P. (1987). EURISKO : an artificial intelligence tool for automatic online information retrieval In 11th Online information, London, 8-10 December, (pp. 431-442). Oxford : Learned information.

[BART88] Barthes C. & Glize, P. (1988). Planning in an expert system for automated information retrieval. In Y. Chiaramella, (Ed.), Proceedings of the 11th international conference on research and development in information retrieval, June 13-15. Grenoble, France. (pp. 535-550). Grenoble : Presses universitaires.

[BASS86] Bassano, J.C. (1986). DIALECT: un système expert pour la recherche documentaire. In Actes des 6èmes Journées systèmes-experts, Avignon, (pp. 1327-1350).

[BASS87] Bassano, J.C. (1987). Systèmes experts et systèmes documentaires intelligents - état de l'art et perspectives In Actes des 7èmes journées systèmes experts. Avignon. (pp. 491-510)

[BELK87] Belkin, N.J. & Croft, W.B. (1987) Retrieval techniques. In M. E. Williams, (Ed.), Annual review of information science and technology, 22. 109-145, Elsevier

[BERT88] Bertino, E., Gibbs, S. & Rabitti, F. (1988). Document query processing strategies : cost evaluation and heuristics. *SIGOIS Bulletin* 9(23), 169-181

[BITCH80] Bitcheler, J. & Eaton, E.A. (1980). The combined use of bibliographic coupling and cocitation for document retrieval. *Journal of the American society for information science* 31(4), 278-282.

[BLAI85] Blair, D.C. & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM* 28(3), 289-299.

[BOSC86] Bosc, P., Courant, M. & Robin, S. (1986). CALIN - A user interface based on a simple natural language. In F. Rabitti, (Ed.), Proceedings of the ACM Conference on research and development in information retrieval, September 8-10, Pisa, Italy.

[BRAC88] Brachman, R.J. & McGuinness, D.L. (1988). Knowledge representation, connectionism, and conceptual retrieval. In Y. Chiaramella, (Ed.), Proceedings of the 11th international conference on research and development in information retrieval, June 13-15, Grenoble, France. (pp. 161-174). Grenoble : Presses universitaires.

[CALL83] Callon, M., Courtial, J.P., Turner, W.A. & Bauin, S. (1983). From translation to problematic networks : an introduction to co-word analysis, *Informations en sciences sociales* 22(2), 191-235.

[CHIA86] Chiaramella, Y., Defude, B., Bruandet, M.F. & Kerkouba, D. (1986). IOTA: a full text information retrieval system. In F. Rabitti, (Ed.), Proceedings of the ACM Conference on research and development in information retrieval, September 8-10, Pisa, Italy, (pp. 207-213).

[CHIA87] Chiaramella, Y. & Defude, B. (1987). IOTA : un prototype de système expert en recherche d'informations, 7èmes journées systèmes experts, Avignon, (pp. 511-528).

[CHRI84] Christodoulakis, S. & Faloutsos, Ch. (1984). Signature files : an access method for documents and its analytical performance evaluation. *ACM transactions on office information systems* 2(4), 267-288.

[CLEV66] Cleverdon, C.W. & Keen, E.M. (1984). Aslib-Cranfield research project. Vol. 2. Test results. Cranfield institute of technology, Cranfield, England.

[CLEV77] Cleverdon, C.W. (1977). A computer evaluation of searching by controlled language and natural language in an experimental NASA database. Rep. ESA 1/432, European Space Agency, Frascati, Italie.

[COHE87] Cohen, P.R. & Kjeldsen, R. (1987). Information retrieval by constrained spreading activation in semantic networks. *Information processing and management*, 23(4), 255-268.

[CONK87] Conklin J. (1987). Hypertext : an introduction and survey. *IEEE Computer*, 20(9), 17-41.

[CONST87] Constantopoulos, P., Yeorgaroudakis, Y., Konstantas, D., Kreplin, K., Eirund, H., Fitas, A., Savino, P., Converti, A., Martino, L., Rabitti, F., Bertino, E., Thanos, C. & Beetstra, T. (1987). Office document retrieval in MULTOS. In ESPRIT'86, Results and Achievements (pp. 563-574), North Holland.

[COUL86] Coulon, D. & Kayser, D. (1986). Informatique et langage naturel: présentation générale des méthodes d'interprétation des textes écrits. *Techniques et science informatiques* 5(2), 103-128.

[COUL82] Coulon, D. & Kayser, D. (1982). La compréhension, un processus à profondeur variable. *Bulletin de psychologie*, 35(356), 815-823.

[CREH86] Créhange, M., David, J.M., Foucault, O., Halin, G. & Thiery, O. (1986). Les structures de données dans le projet EXPRIM. Congrès Inforsid.

[CROF87] Croft, W.B., & Thompson, R.H., (1987). I3R : A new approach to the design of document retrieval systems. *Journal of the American society for information science.*, 38(6), 389-404.

[CROF88] Croft, W.B., Lucia, T.J. & Cohen, P.R. (1988). Retrieving documents by plausible inference : a preliminary study. In Y. Chiaramella, (Ed.) Proceedings of the 11th international conference on research and development in information retrieval June 13-15, Grenoble, France. (pp. 481-494),Grenoble : Presses universitaires.

[DANI86] Daniels, P.J. (1986). Cognitive models in information retrieval : an evaluative review. *Journal of documentation* 42(4), 272-304.

[DEBI88] Debili, F., Fluhr, C. & Radasda, P. (1988). About reformulation in IRS. In RIAO 88, User-oriented content-based text and image handling, March, (pp. 21-24), MIT, Cambridge MA, (pp.343-360).

[DEFU86] Defude, B. (1986). Etude et réalisation d'un système intelligent de recherche d'informations : le système IOTA , Thèse de l'institut national polytechnique de Grenoble : spté informatique.

[DUB85] Dubois, D. & Prade, H. (1985). Théorie des possibilités. Applications à la représentation des connaissances en informatique, Paris : Masson.

[ELHA87] El-Hamdouchi, A. & Willett, P. (1987). Techniques for the measurement of clustering tendency in document retrieval systems. *Journal of information science,* 13, 361-365.

[FALOU87] Faloutsos, C. & Christodoulakis, S. (1987). Description and performance analysis of signature file methods for office filing. *ACM transactions on office information systems* 5(3), 235-257.

[FLUH80] Fluhr, C. (1980). SPIRIT : a Linguistic and probabilistic information storage and retrieval system. First international workshop on natural communication with computers, Warsaw.

[FLUH85] Fluhr, C. & Debili, F. (1985). Interrogation en langue naturelle de données textuelles et factuelles, RIAO 85, Grenoble.

[FOUR84] Fournier, J.P. (1984). 'Sauvetage' de raisonnement en langage naturel. Thèse de 3ème cycle, Université Paris-Sud.

[FOX87] Fox, E.A. (1987). Development of the CODER system : a testbed for artificial intelligence methods in information retrieval. *Information processing and management* 23(4), 341-366.

[FOX80] Fox, E.A. (1980). Lexical relations : enhancing effectiveness of information retrieval systems. *ACM SIGIR forum*,15(3), 5-36.

[FOX89] Fox, E.A. (1989). Research and development of information retrieval models and their application. *Information processing and management* 25(1), 1-5.

[HALIN88] Halin, G., Mouaddib, N., Foucaut, O. & Créhange, M. (1988). Semantics of user interface for image retrieval : possibility theory and learning techniques applied on

two prototypes. In RIAO 88, User-oriented content-based text and image handling. March 21-24, MIT, Cambridge MA, (pp. 676-688).

[HOBB82] Hobbs, J.R., Walker, D.E. & Amsler, R.A. (1982). Natural language access to structured text. In J. Horecky, (Ed.), COLING 82 : Proceedings of the 9th International conference on computational linguistics July 5-10, Prague, Tchécoslovaquie. Amsterdam : North Holland (pp. 127-132).

[JACO88] Jacobs, P.S. & Rau, L.F. (1988). Natural language techniques for intelligent information retrieval. In Y. Chiaramella, (Ed.), Proceedings of the 11th international conference on research and development in information retrieval, June 13-15, Grenoble, France. (pp. 85-100). Grenoble : Presses universitaires.

[JARD71] Jardine, N. & Van Rijsbergen, C.J. (1971). The use of hierarchic clustering in information retrieval. *Information storage and retrieval*, 7(5), 217-240.

[JIME88] Jimenez Guarin, C. (1988). Access by content of documents in an office information system. In Y. Chiaramella, (Ed.), Proceedings of the 11th international conference on research and development in information retrieval, June 13-15, Grenoble, France. (pp. 629-650). Grenoble : Presses universitaires.

[KAYS84] Kayser, D. (1984). Examen de diverses méthodes utilisées en représentation des connaissances Actes du coll. AFCET-RFIA, Paris, (pp. 115-144).

[KITT82] Kittredge, R. & Lehrberger, J. (1982). Sublanguages : studies of language in restricted domains. New York : Walter De Gruyter.

[LANC88] Lancel, J.M. & Simonin, N. (1988). Tex-nat : a tool for indexing and information retrieval, In RIAO 88, User-oriented content-based text and image handling, March 21-24, MIT, Cambridge MA, (pp. 369-378).

[LECO89] Le Coadic, Y. (1989). Une politique scientifique pour l'information. *Le Documentaliste*, 26(2), 59-64.

[MICH85] Michelet, B. & Turner, W.A. (1985). 'Co-word search : a system for information retrieval. *Journal of information science* 11, 173-181.

[ODDY77] Oddy, R.N. (1977). Information retrieval through man-machine dialogue. *Journal of documentation.*, 33(1), 1-14.

[PANY87] Panyr, J. (1977). Conceptual clustering and relevance feedback. *International classification*, 14(3), 133-137.

[PEDE87] Pedersen, G.H. & Larsen, H.L. (1987). The design of an information retrieval assistant system. ESPRIT'87, Achievements and impact Proceedings of the 4th Annual ESPRIT Conference Brussels, Sept. 28-29, (pp. 688-700).

[POGU87] Pogue, Ch. A. & Willett, P, (1987). Use of text signatures for document retrieval in a highly parallel environment. *Parallel computing* 4, 259-268.

[PRAD85] Prade, H. & Testemale, C. (1985). Système de base de données autorisant des requêtes vagues et des données imprécises. Actes AFCET, Dijon.

[ROUAU87] Rouault, J. (1987). Linguistique automatique, Applications documentaires, Berne : Peter Lang.

[RICH79] Rich, E. (1979). User modeling via stereotypes. *Cognitive Science* 3, 329-354.

[RUM86] Rumelhart, D.E., McClelland, J.L. and the PDP research group. (1986). Parallel distributed processing , MIT Press, Cambridge, Mass.

[SAG78] Sager, N. (1978). Natural language information formatting : the automatic conversion of texts to a structured data base. In M. C. Yovits, (Ed.), Advances in computers, 17, New York : Academic Press,

[SALT83] Salton, G., Fox, E.A. & Wu, H. (1978). Extended boolean information retrieval. *Communications of the ACM* 26(12), 1022-1036.

[SALT85] Salton, G. & Vorhees, E. (1985). Automatic assignment of soft boolean operators, Proceedings of the eighth annual international ACM SIGIR conference on research and development in information retrieval, (pp. 54-69). New York : Association for computing machinery.

[SALT86a] Salton, G. (1986). Recent trends in automatic information retrieval. In F. Rabitti, (Ed.), Proceedings of the ACM Conference on research and development in information retrieval, September 8-10, Pisa, Italy.

[SALT86b] Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM* 29(7), 648-656.

[SALT89] Salton, G. (1989). Automatic text processing : the transformation, analysis and retrieval of information by computer, New York : Addison Wesley.

[SCHAN81] Schank, R.C., Kolodner, J.L. & De Jong, G.D. (1981). Conceptual information retrieval, In R.N. Oddy, S.E. Robertson, C.J. Van Rijsbergen, & P. W. Williams, (Eds.), Information retrieval research. (pp.94-116). London : Butterworths.

[SHOV85] Shoval, P. (1985). Principles, procedures and rules in an expert system for information retrieval. *Information Processing and Management*, 21(6), 475-487.

[SFBA87] Société française de bibliométrie appliquée. (1987). Les systèmes d'informations élaborées. Congrès S.F.B.A, Ile Rousse, 23-25 Septembre. SFBA, BP 1507 - 75327 Paris Cedex 07.

[SFBA89] Société française de bibliométrie appliquée. (1989). Les systèmes d'informations élaborées. Congrès S.F.B.A, Ile Rousse, 31 Mai - 2 Juin. SFBA, BP 1507 - 75327 Paris Cedex 07.

[TSIC83] Tsichritzis, D., Christodoulakis, S., Economopoulos, P., Faloutsos, C., Lee, A., Lee, D., Vandenbroek, J. & Woo, C. (1983). A multimedia office filing system. Proceedings of 9th International Conference on VLDB, Oct.31-Nov. 2, Florence, Italie, (pp. 2 -7).

[TURN85] Turner, W., Chartron, G. & Michelet, B. (1985). Describing scientific and technological problem networks using manually and automatically indexed full text data bases : some co-word analysis techniques, Colloque OCDE, June 10-13, Paris.

[VANR86] Van Rijsbergen, C.J. (1986). A new theoretical framework for information retrieval. In F. Rabitti, (Ed.), Proceedings of the ACM Conference on research and development in information retrieval, September 8-10, Pisa, Italy.

[VOOR86] Voorhees, E.M. (1986). The efficiency of inverted index and cluster searches. In F. Rabitti, (Ed.), Proceedings of the ACM Conference on research and development in information retrieval, September 8-10, Pisa, Italy, (pp. 164-174).

[WILL87] Willett, P. (1987). Effectiveness of retrieval in clustered document files. In 11th Online information, 8-10 December. Oxford : Learned information (pp. 137-146).

[ZARR88a] Zarri, G.P. (1988). Conceptual representation for knowledge bases and 'intelligent' information retrieval systems. In Y. Chiaramella, (Ed.), Proceedings of the 11th international conference on research and development in information retrieval, June 13-15, Grenoble, France. (pp. 551-566). Grenoble : Presses universitaires.

[ZARR88b] Zarri, G.P. (1988). Etat de l'art - Les nouvelles tendances de l'informatique documentaire. *Bulletin du CID*, 32, 11-40.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique

ISSN 0249-6399