Mabrouka EL HACHANI

DEA SCIENCES DE L'INFORMATION ET COMMUNICATION
Option 3 : Système d'Information documentaire

L'INDEXATION AUTOMATIQUE

Note de Synthèse

sous la direction de M. M. HASSOUN
Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB)

Mars 1997

SOMMAIRE

INTRODUCTION

Le travail d'une documentaliste ne consiste pas uniquement à classer des documents ou à rechercher l'information, elle doit également indexer un nombre plus qu'important de documents afin de les retrouver lors d'une recherche documentaire. La tâche est encore plus importante lorsqu'il s'agit d'une grande unité documentaire, mais la technologie a permis bon nombre d'avancée notamment depuis l'entrée de l'informatique dans ce domaine.

L'explosion de l'informatique est un phénomène plus qu'important depuis ces dix dernières années, et le traitement automatique de l'information documentaire est une application en développement constant. Au fur et à mesure, l'implantation des équipements informatiques dans les centres de documentation et bibliothèques ne cesse de s'accroître et de se poursuivre.

Ce traitement se traduit dans la réalité par l'automatisation de diverses tâches ou fonctions documentaires : gestion de fichiers, recherche, diffusion. Il arrive dans certains cas que toute la chaîne documentaire soit automatisée [CHAUMIER, 94]. Les bases de données (BDD) ont permis de rassembler les fichiers manuels qui décrivaient le document et qui occupaient beaucoup d'espace, facilitant ainsi les recherches sur un document ou un auteur.

Nous verrons dans un premier temps comment nous avons procédé aux recherches de références qui nous ont permis de nous documenter sur ce sujet. Dans une seconde partie nous aborderons tout ce qui relève de l'indexation automatique, l'opposition entre indexation manuelle et indexation automatique, les différentes méthodes utilisées pour l'indexation automatique. Puis, nous terminerons notre étude sur l'évaluation de ces systèmes d'indexation et les attentes des professionnels par rapport à ces systèmes.

1-Methodologie de la recherche

1.1.Délimitation du sujet

Nous avons choisi de traiter de l'indexation automatique en abordant les principes et le fonctionnement des principaux systèmes d'indexation. Notre recherche a donc été centrée sur les méthodes d'indexation automatique.

1.2.Recherche de références

***choix des mots clés**

Le choix des mots clés n'a pas véritablement posé de problème, une fois le sujet délimité. Les mots clés choisis sont :

- indexation/indexing
- indexation et documentation (pour éviter l'indexation des prix en économie)/indexing and documentation
- indexation automatique/automatic indexing
- indexation et automatisation/indexing and automatic process
- traitement du langage naturel/natural language processing

Nous avons indiqué pour chaque descripteur français un descripteur équivalent en anglais, qui servira lors d'interrogation de BDD anglo-saxone.

1.2.1.Bibliothèque et Centre de documentation de l'ENSSIB

➢Catalogue informatisé de l'ENSSIB

Nous avons interrogé le catalogue informatisé de l'ENSSIB sur le site Web[1] .

Nous avons procédé à une recherche simple sur le champ *titre de l'ouvrage*, qui comprend les monographies, les thèses, les mémoires de DEA, etc. :

- d'abord à partir du terme *indexation* : 181 références

Pour affiner les recherches, nous avons choisi un descripteur plus spécifique, toujours à partir d'une recherche sur le titre :

[1] http://www.enssib.fr

- *indexation automatique* : 46 références.

Sur ces 46 références quelques-unes étaient identiques à ceux de la première recherche. De nombreuses références nous renvoyaient vers des documents du Centre de documentation de l'ENSSIB. Nous sommes allées consulté les listes de thèses, mémoires, notes de synthèse, ainsi que le Dossier *indexation automatique* qui contenait de nombreux articles sur le sujet.

➢**Dépouillement de certains périodiques**

Nous avons procédé au dépouillement de certains périodiques spécialisés :

- en documentation et en sciences de l'information tels que :

  - Documentaliste-Sciences de l'information,

  - Archimag

  - Journal of Librarianship and information science

  - Journal of American Society for Information Science (JASIS)

- d'autres revues spécialisées dans le traitement automatique des langues telles que :

  - La Tribune des Industries de la Langue,

  - TAL (Traitement Automatique des Langues), celle-ci nous a posé quelques problèmes : elle ne se trouve qu'à la Bibliothèque municipale de Lyon à la Part Dieu, qui ne possède que quelques numéros de cette revue.

1.2.2.Bases de données sur CD-ROM

Nous avons également interrogé des BDD sur CDROM :

➢**BN Opale :** Base bibliographique de la Bibliothèque Nationale de France, elle recense les documents imprimés entrés dans la collection depuis 1970.

Nous avons effectué une recherche simple sur le champ mot-sujet (ms) :

- ms=*Indexation documentation* : 16 références

- ms=*indexation automatique* : 3 références

Devant le peu de références trouvées nous avons décider d'élargir la recherche en choisissant deux termes plus génériques, soit *indexation* et *automatisation*.

ms=*indexation* : 36 références

ms=*automatisation* : 244 références

Recherche combinée (CS) sur mot-sujet : *indexation et automatisation*

cs=0

Le résultat a été beaucoup trop important et non pertinent, car pour le descripteur *automatisation*, nous avions des références d'automatisation en dehors de l'indexation. Aussi avions-nous décidé de combiner les deux mots sujets « indexation et automatisation », le résultat a été nulle.

➢**DOC-THESE** : CD-ROM recensant les thèses de doctorat soutenues en France depuis 1972 pour les lettres, sciences humaines et sociales et les sciences et 1983 pour les disciplines de santé. Il est édité par l'ABES (Association Bibliographique de l'Enseignement Supérieur).

Nous avons procédé à une recherche simple sur le champ mot-clé :

mot-clé : *indexation automatique* : 23 références

➢ **PASCAL & FRANCIS :** les deux bases bibliographiques de l'Institut National de l'Information Scientifique et Technique (**INIST**). Nous avons interrogé les deux bases en français et en anglais sur deux types de champ de recherches :

DEF : mot clé en français DXF : expression en français

DEA : mot clé en anglais DXA : expression en anglais

-Pascal : base de données multilingue (français-espagnol-anglais) recensant l'ensemble de la littérature internationale dans le domaine scientifque, médicale et technique.

Les mots clés choisis sont :

DEF=indexation➔ résultat : 34 réponses

DEF=indexation automatique ➔ résultat : 101 réponses

DXF=indexation documentaire➔ résultat : 19 réponses

Les mots clés choisis en anglais sont :

DXF=automatic indexing➔ résultat : 53 réponses

-Francis : base de données multilingue (français-espagnol-anglais) recensant l'ensemble de la littérature internationale dans le domaine des sciences humaines.

Les mots clés choisis en français sont :

DEF=indexation➔ résultat : 34 réponses

DXF=indexation automatique➔ résultat : 16 réponses

Les mots clés choisis en anglais sont :

DEA=indexing➔ résultat : 9 réponses

DXA=automatic indexing➔ résultat : 226 réponses

Nous avons remarqué une certaine redondance au niveau des résultats : certaines références étaient présentes dans les réponses lors de l'interrogation des deux bases.

1.2.3.Bases de données en ligne

Nous avons interrogé trois BDD sur le serveur DIALOG : Library Information Science Abstract (LISA), PASCAL et Information Science Abstract.

Des perturbations sur le serveur nous ont obligées à formuler des requêtes simples:

1/Automatic indexing	1738 références
2/Language processing natural	0 référence
3/language	76531 références
4/Processing natural	912 références
5/Language and processing natural	95 références
6/(automatic index ?) and (language and process ? natural)	6 références

Des problèmes de connexion sur le serveur DIALOG ont perturbé notre interrogation et se sont traduits par l'apparition de lettres parasites lors de la saisie ce qui donnaient des requêtes incompréhensibles pour le système. Ces perturbations nous ont empêché de continuer

la recherche en affinant les requêtes, le serveur ayant directement mis fin à l'interrogationNous avions néanmoins pu enregistrer les 6 références trouvés.

1.2.4.Internet

➢**Moteurs de recherche**

Nous avons interrogé à partir des moteurs de recherche tels qu'AltaVista, AltaVista Europe et Infoseek, à partir des mots clés simples :

- *indexation :* le résultat a été de 5 000 références

- *traitement électronique du document* : 40 000 références

Le résultat sur Internet n'est pas pertinent, pour deux raisons : tout d'abord, il n'est pas possible de faire une recherche fine ou pertinente à cause de la faiblesse des outils. La deuxième raison est l'ambiguïté du terme de la requête : elle semble avoir été comprise comme *indexation automatique=moteur de recherche* sur internet, et non les ouvrages traitant de l'indexation automatique. Le résultat a été de 70 000 à 110 000 documents, un bruit impressionnant !

➢**Listes de diffusion**

Ce sont des forums de discussions spécialisés dans un domaine et sont plus sérieux que les forums habituels. Nous nous sommes inscrites à des listes de professionnels en sciences de l'information, en documentation et bibliothèque : Biblio.fr, ADBS.

Nous avons envoyé un message à la liste précisant notre sujet de recherche (l'indexation automatique), nous avons reçu des réponses intéressantes. Ces réponses n'étaient pas forcément des titres d'ouvrages mais des noms de personnes ou d'organismes qui ont travaillé sur le sujet ou qui ont eu à réaliser une indexation automatique comme l'INIST, ou des équipes de recherche au Canada plus précisément l'équipe de l'EBSI de l'Université de Montréal.

➢**Serveurs d'organismes spécialisés en sciences de l'information**

ADBS : Association des Documentalistes et Bibliothécaires Spécialisés

Nous avons interrogé le site Web de l'ADBS[2], nous avons cherché à partir d'un champ *recherche à partir de mot clé* :

- indexation ➔résultat : 42 références

- indexation automatique ➔résultat : 6 références

Les six références de l'ADBS sont celles de la revue Documentaliste-Science de l'information éditée par l'association.

Il semblerait que la recherche aléatoire et le butinage aient été plus efficace que la recherche sur internet ou même sur un catalogue informatisé. En effet, à la bibliothèque de l'ENSSIB, nous avons pris la cote des ouvrages qui traitent de l'indexation automatique en 025.04, nous avons trouvé quelques ouvrages qui n'étaient pas apparus dans les résultats lors de l'interrogation du catalogue informatisé.

Tableau récapitulatif des recherches documentaires

Intitulé	Domaine(s) couvert(s)	nombre de références pertinente
Catalogue informatisé de l'ENSSIB	Sciences de l'information	46
Bibliothèque Nationale de France (BNF)	Bibliographie de la BNF	3
PASCAL	Sciences et techniques, medecine	19
FRANCIS	Sciences humaines	16
DOCTHESE	Thèses de doctorat en lettres et sciences humaines	10
BDD en ligne **LISA/PASCAL/INFORMATION SCIENCE ABSTRACT**	Sciences de l'information	6

[2] http://www.adbs.fr

INTERNET **Moteurs de recherche** : ALTAVISTA/INFOSEEK	Domaines couverts très larges. Nos recherches ont surtout porté sur les sciences de l'information et documentation.	0
Liste de diffusion		
ADBS	Sciences de l'information	0
Biblio.FR	Sciences de l'information et bibliothèques	0
Serveurs d'organismes spécialisés		
ADBS	Sciences de l'information	6
	TOTAL	**106**

Parmi ces 106 références, nous avions trouvé quelques références redondantes.

2.L'indexation automatique

L'indexation automatique est l'opération qui consiste à *faire reconnaître par l'ordinateur des termes figurant dans le titre, le résumé, le texte complet (s'il est enregistré avec la notice documentaire) et parfois même l'indexation humaine, et à employer ces termes, soit tels quels soit après conversion en d'autres termes équivalents ou conceptuellement voisins, pour en faire des critères incorporés dans le fichier de recherche et utilisables pour retrouver le document*[3].

En France, l'âge d'or de la recherche en indexation automatique, dans les années 1965-70, a été marqué par les travaux du laboratoire d'automatique documentaire et linguistique du CNRS, dirigé par J.CL. GARDIN. C'était une époque marquée par les recherches sur le traitement automatique des langues, les études sur la traduction automatique le démontrent. On a en effet rêvé pouvoir traduire les poèmes de Shakespeare, les romans de Balzac, ou encore pour une activité un peu plus documentaire, produire des résumés automatiquement. Paradoxalement, c'est avec l'avancée des recherches en intelligence artificielle et le développement de l'informatique que les chercheurs sont devenus plus modestes, car on est passé de la traduction automatique à la traduction assistée par ordinateur (TAO), et de l'indexation automatique à l'indexation assistée par ordinateur (IAO) [CHAUMIER, 90].

[3] G. VAN SLYPE, *Les langages d'indexation : conception, construction et utilisation dans les systèmes documentaires*, Paris :Les Editions d'organisation, 1987, pp.168-169

Il existe deux types d'indexation automatique ou semi-automatique : le premier consiste à enrichir automatiquement l'indexation humaine par autopostage générique[4] ou encore une indexation automatique non sélective (prise en compte de tous les mots non vides du document). Ce type d'indexation est utilisé de façon généralisée. Le deuxième type d'indexation automatique est l'indexation automatique sélective, c'est-à-dire une prise en compte de certains termes seulement jugés par le système comme les plus représentatifs du contenu du document soit en langage naturel soit en langage contrôlé [VAN SLYPE, 87]. Une grande majorité des systèmes bâtis sur ce type d'indexation était encore en expérimentation jusqu'à l'intégration de module linguistique depuis une dizaine d'années environ.

Pour mieux se rendre compte de l'importance de l'indexation automatique nous allons tout d'abord voir la comparaison entre l'indexation humaine et l'indexation automatisée.

2.1.L'opposition entre l'indexation manuelle et l'indexation automatique

Marie-Gaëlle MONTEIL définit l'indexation comme étant *l'analyse documentaire qui a pour objet de produire une représentation réduite et formalisée des documents en y retenant l'ensemble des éléments essentiels* [5].

L'indexation est donc la réduction du volume des données d'un document par le biais d'une représentation de ce document par des mots clés. G. VAN SLYPE décompose l'indexation humaine en quatre étapes : la prise de connaissance du contenu du document, le choix des concepts, la traduction des concepts en descripteurs, l'établissement de liaisons syntaxiques entre les descripteurs. L'indexeur choisit ces mots clés dans une liste de vocabulaire contrôlé formé par le lexique ou le thesaurus, ce qui permet de garantir l'uniformité de la représentation du document [CHAUMIER, 90]. Mais le choix de ces outils de vocabulaire contrôlé dépendra de divers facteurs dont *le type d'information recherchée, (la) finesse demandée à l'interrogation, (la) disponibilité et (la) compétence des indexeurs* [6]. Cela dépendra également du domaine s'il s'agit d'un domaine généraliste, le travail de l'indexeur sera plus allégé, une indexation de type général permet une recherche plus large. S'il s'agit en revanche d'un domaine beaucoup plus « pointu », la pharmacologie par exemple, il faudra tenir compte

[4] L'autopostage générique : (...) *lors de l'initialisation d'une banque de données tous les descripteurs attribués par l'indexeur humain se verront automatiquement complétés par la machine, à l'aide de descripteurs qui leur sont attachés dans le thésaurus par une relation hiérarchique ascendante (ibidem p.170).*

[5] Marie-Gaëlle MONTEIL, *Indexation manuelle et indexation automatique : comparaison et perspectives*, IDT 95 , 12e Congrès, Paris, 12-15 juin 1995, p.214

de la classe des médicaments lors de l'indexation, en fait tout dépend du niveau d'exigence du public auquel on a à faire.

Le processus d'indexation humaine se déroule de la façon suivante : l'indexeur prend connaissance du document qui peut être une monographie, un article de périodique, des actes de communication d'un congrès, etc. [LE LOADER, 94]. Il lit rapidement le titre, la table des matières, le résumé, l'introduction générale, les introductions et conclusions des principaux chapitres, l'intitulé des légendes et figures (s'il y en a) et la conclusion. Il parcourt en diagonale le contenu du document [VAN SLYPE, 87]. Cette lecture rapide permet à l'indexeur de savoir de quoi traite le texte, de connaître « l'aboutness » pour reprendre l'expression de Hutchins (1971), « l'aboutness » désigne le sujet dont parle le document [HODGE, 92].

Les différents modèles d'indexation manuelle, c'est-à-dire le type de représentation qui sera donnée au texte indexé, sont les suivants : *l'indexation dite « à plat »* où tous les descripteurs sont placés au même niveau d'importance par rapport au texte indexé, *l'indexation pondérée* où on distingue les descripteurs principaux et secondaires, c'est la manière d'indiquer qu'un document traite prioritairement d'un sujet. Et enfin, *l'indexation à rôles* où l'on souligne les relations qu'entretiennent entre eux les descripteurs retenus pour mettre en évidence de manière plus fine le sujet du document.

Le processus d'indexation demande beaucoup de qualité de la part de l'indexeur, notamment la compréhension du domaine, de la langue, ainsi que le processus de transfert de l'information et évidemment l'utilisation des langages documentaires [CHAUMIER, 90]. De fait, l'indexation humaine est lourde à gérer et exige des analystes compétents. Elle peut être performante à condition que le nombre de documents à indexer ne soit pas trop élevé, que les documents ne soient pas sur un support informatique et que l'essentiel ne réside ni dans la rapidité ni dans le repérage de l'information [CLAVEL, 93]. Tout ceci représente donc un obstacle, la surcharge des documents à indexer, les connaissances plus ou moins importante de l'indexeur rend son indexation aléatoire sans compter la subjectivité de ses choix.

En entreprise, le temps de la recherche d'information doit être court, en effet les questions souvent fines et pointues exigent un délai de réponse assez rapide. Les enjeux économiques étant considérables dans les systèmes d'information, ils vont faire croître et généraliser l'archivage électronique, et donc l'automatisation de certaines fonctions documentaires dont

[6] Sophie RANJARD, *Indexer et résumer, pourquoi et comment ?* Archimag n°80 décembre/janvier 1995, p.41

l'indexation automatique. En effet, le coût moyen d'un document à indexer manuellement est de 50F, ainsi pour une banque de données moyenne d'un volume de 100 000 documents, le coût d'indexation de la base est de 5 millions de francs, et pour un accroissement de 20 000 documents par an, l'indexation annuelle revient à 1 million de francs, cela représente un budget colossale [CHAUMIER, 90].

Autre exemple, l'analyse et l'indexation de 500 000 documents par an à l'Institut Nationale de l'Information Scientifique et Technique (INIST) nécessite l'intervention de 140 ingénieurs documentalistes. Les coûts financiers de cette activité a contraint l'INIST à s'engager dans un processus d'automatisation de cette fonction documentaire en 1989 et 1990 [CORET , 91]. Il faut surtout ne pas perdre de vue que le coût de travail de l'homme augmente toujours alors que celui de la machine ne cesse de baisser.

Mais on peut se demander si cette activité essentiellement intellectuelle et humaine peut faire l'objet d'une automatisation ? En effet l'indexation automatique exige de la machine une double compétence : celle de la langue et celle de la pensée scientifique pour traiter indifféremment un article de physique en même temps qu'un article de sociologie comme le ferait un être humain. J.C. GARDIN définit l'indexation automatique comme *un ensemble de règles assurant le passage automatique d'un texte écrit dans une langue naturelle à une représentation de ce texte qui soit censée en exprimer le sens du point de vue largement intuitif où se placent habituellement le documentaliste*[7].

Il est vrai qu'aujourd'hui, l'indexation automatique est pertinente à 60% contre 95% pour l'indexation humaine. Cette différence au niveau du résultat trouve son explication dans l'ambiguïté de la langue, facile à traiter pour un être humain mais complexe pour un outil automatique disposant de compétences linguistiques réduites. Mais cette limite n'est que provisoire du fait de l'évolution permanente et sensible des outils et des ressources linguistiques dans le cadre du traitement automatique du document.

De plus, la qualité des documents à traiter est très variable car l'indexation automatique a besoin d'un minimum d'information textuelle à traiter estimé, approximativement, à dix lignes de texte « informatif » pas toujours présent dans un résumé d'auteurs. Mais l'intérêt de l'indexation automatique est évidemment économique, on a besoin

[7] J. CHAUMIER, M. DEJEAN*, L'indexation documentaire, de l'analyse conceptuelle humaine à l'analyse morphosyntaxique*, Documentaliste, vol.27, n°6, novembre-décembre 1990, p.276

de traiter rapidement une grande masse d'information, en simplifiant et optimisant les circuits documentaires. Cette simplification et optimisation doit se faire avec une cohérence et une homogénéité au niveau de l'indexation. Mais ceci n'est pas le cas pour l'indexation manuelle qui manque d'uniformité, en effet, les consignes d'indexation sont peu formalisables et donc la « profondeur » de l'analyse est généralement dépendante de l'indexeur [MONTEIL, 95].

En fait, l'indexation automatique pourrait être le complément de l'indexation humaine ou inversement, car l'indexation automatique comporte deux phases : la première est une phase de préindexation automatique au cours de laquelle l'ordinateur analyse le texte qui lui est soumis et lui associe un certain nombre de descripteurs généralement extraits d'une liste d'autorité qu'il propose au documentaliste. La seconde phase est une phase de dialogue entre le documentaliste et l'ordinateur au cours de laquelle la liste proposée au cours de la phase précédente est affinée par voie humaine. Il semblerait que l'indexation automatique et l'indexation humaine ne soient pas si opposées [VAN SLYPE, 87].

Indexation assistée par ordinateur ?

L'indexation automatique n'existe pas véritablement, aucun système pour le moment n'indexe de façon totalement autonome des textes numérisés, c'est pour cela que l'on parle d'indexation assistée par ordinateur. En effet, on peut dire que les systèmes actuels s'ils remplacent l'homme pour une importante part de son expertise, ils ne le remplaceront pas complètement, car l'expression « indexation automatique » suppose une intervention totale du système, ce qui est loin d'être le cas et l'intervention humaine est toujours nécessaire [BLANQUET, 94]. On peut citer en exemple SINTEX et ALEXDOC comme logiciels d'indexation semi-automatique d'indexation assistée par ordinateur.

2.2.Les différentes méthodes d'indexation automatique

De nombreuses méthodes (linguistiques, statistiques, par assignation, etc.) ont été développées pour concevoir, ou améliorer dans certains cas, les systèmes ou les logiciels d'indexation automatique.

2.2.1.La méthode linguistique

On constate en indexation que l'objet analysé c'est-à-dire le texte et les descripteurs, utilisés pour la représentation de celui-ci, font tous deux appel à la linguistique. De plus, le fait

que certains systèmes d'indexation utilisent les techniques du traitement automatique des langues, démontre la pertinence d'une approche linguistique.

On a remarqué que l'absence de syntaxe lors des interrogations peut provoquer un bruit important au niveau du résultat. C'est pourquoi des méthodes basées sur une désambiguïsation syntaxique ont été élaborées. *Ces méthodes vont d'une simple marque distinctive affecté à chaque descripteur (un rôle) à une véritable structuration syntaxique de l'énoncé documentaire* [8]. On peut citer deux systèmes qui fonctionnent sur ce modèle : PRECIS (Preserved Context Indexing System, c'est-à-dire « système d'indexation respectant le contexte ») de D. AUSTIN, adopté par la Bibliographie Nationale Britannique (BNB) et SYNTOL (Syntagmatic Organization Language), réalisé en 1960 par J.C. GARDIN et ses collègues du CNRS et de la Maison des sciences de l'homme avec le soutien de l'Euratom.

Ces travaux permirent d'ouvrir la voie à divers autres travaux qui aboutirent à la conception de logiciels, expérimentés pour la plupart sur des banques de données de différents organismes importants. On peut citer PIAF-DOC (Programme Interactif d'Analyse du Français) expérimenté par la Documentation Française ou encore le système SINTEX (Système d'INdexation de TEXtes), de la Société d'Information Européenne (SIE), expérimenté sur les banques de données des Communautés Européennes, ECODOC et CELEX. Les travaux de recherche réalisés à l'université ne sont pas négligeables, l'exemple du logiciel PASSAT (Programm zur Automatischen Selektion von Stichwörtern Aus Texten) de Siemmens, c'est en effet un programme élaboré à partir de travaux de chercheurs de l'Université de Heidelberg à la fin des années 1970. Ce programme, après modification et amélioration, est toujours opérationnel et utilisé par Volkswagen pour l'élaboration de sa base de données [CHAUMIER, 90].

Ces systèmes basés sur une analyse syntaxique ne sont pas très performants lors de l'interrogation. Le taux de bruit et de silence ont certes été réduits par rapport aux systèmes de départ et l'analyse syntaxique a permis de nombreuses avancées. Des travaux, axés sur l'analyse de contenu et sur l'analyse de textes ont permis de concevoir d'autres systèmes dont notamment les analyseurs syntaxiques et morphologiques qui sont apparus dans la décennie 1980. On peut citer les logiciels AlexDoc[9] de la société GSI-Erli et SPIRIT de la société SYSTEX qui permettent une indexation sur texte intégral et sur résumé ainsi que l'interrogation

[8] J. MANIEZ, *Les langages documentaires : conception, construction et utilistion dans les systèmes documentaires*, Paris :Les Editons d'organisation, 1987, p.251

des bases de données en texte intégral par le biais d'une indexation des questions. Les systèmes ALETH et DARWIN fonctionnent aussi sur le même schéma : ils cherchent la réponse par essais et erreurs successifs qui ne peuvent influencer, en cas d'impasse, les sous systèmes déjà résolus.

Ces systèmes, combinant différentes analyses linguistiques pour le traitement en langage naturel, sont formés de plusieurs modules de traitement linguistique ayant chacun une analyse spécifique :

- niveau morphologique : on isole chaque terme par le biais d'un dictionnaire qui permet le contrôle des chaînes de caractères et le repérage des mots,

- niveau lexical : il s'agit ici de polymorphisme de mot appartenant à un même concept, le traitement se traduit par la suppression des variantes combinatoires (flexion, dérivation, conjugaison) pour obtenir une forme canonique par réduction ou lemmatisation. Les outils nécessaires à ce procédé de réduction sont des dictionnaires de correspondances entre formes fléchies ou dérivées et formes canoniques (par exemple produira, produisent, ont produit etc., auront la même forme canonique *produire*) ainsi que des règles d'établissement par correspondance.

Les traitements morphologique ou lexical engendrent des ambiguïtés sémantiques qui peuvent être levées par d'autres analyses linguistiques :

- la syntaxe permet dans certains cas de résoudre des ambiguïtés mais elle est impuissante face à certaines ambiguïtés de la langue naturelle telle que « le pilote ferme la porte » ou encore « la belle porte le voile » !

- le niveau pragmatique, l'être humain fait appel à ses connaissances du monde et du contexte pour résoudre les cas de polysémie, dans le cas du traitement automatique, on fait appel aux réseaux sémantiques. Mais ces réseaux ne concernent que des domaines très spécialisés et sont donc d'utilisation restreinte [CLAVEL, 93].

D'autres recherches en traitement linguistique de l'information ont été menées pour améliorer la qualité de l'indexation automatique. On s'est intéressé à la sélection des mots isolés qui représente la faiblesse majeure de l'indexation automatique. En effet la reconnaissance des

[9] AlexDoc est l'une des couches d'applications du logiciels d'enregistrement et de recherche documentaire ALEXIS. AlexDoc comporte des programmes d'aide à l'indexation des documents et des questions.

syntagmes nominaux [10] et de leur fonction syntaxique permettraient une recherche plus fine. Des analyseurs morpho-syntaxiques ont été élaborés dans ce sens. Plusieurs équipes de recherche ont travaillé sur cet aspect du syntagme nominal on peut citer le groupe SYDO à Lyon, ou d'autres équipes du monde entier, toutes les langues ou presque ont en un (portugais, espagnol, japonais, arabe, etc.)[11]. Le problème de ce type d'analyse est comme on l'a précédemment cité l'ambiguïté de la langue naturelle, comment opérer un choix dans ce cas ? Michel LE GUERN du groupe SYDO préconise de prendre en compte les deux solutions pour éviter que, lors de l'interrogation, l'utilisateur ne soit en face d'un silence du système.

2.2.2.La méthode statistique

L'initiateur des méthodes d'indexation automatique reste sans aucun doute H.P. Luhn avec son célèbre article *The automatic creation of literature abstracts* paru en 1958 dans le Journal of Research and Development d'IBM. Il déclare : « (...)*au lieu de tirer l'information au hasard comme le fait normalement le lecteur, la nouvelle méthode automatique choisit les phrases d'un article qui représentent le mieux l'information pertinente »*[12] [CHAUMIER, 90]. H. P. Luhn ouvrit la voie aux travaux sur l'indexation automatique par voisinage appelée aussi méthode statistique, en quoi consiste-t-elle ? « *la fréquence d'un mot dans un article fournit la mesure utile de la sigifiance d'un mot (...), la cooccurrence relative dans une phrase de mots auxquels ont été affectés des poids de signifiance est une mesure utile de la signifiance de la phrase (...)* [13]. Plus certains mots sont souvent rencontrés en compagnies les uns des autres dans une phrase plus on peut dire que ces mots sont lourds de sens [CHAUMIER, 90].

Les travaux de Luhn ont ouvert la voie à d'autres analyses statistiques, on peut citer en France, Francis LEVERY qui, en 1963, fait remarquer que *l'indexage (terme utilisé à l'époque pour indexation) des textes techniques à l'aide de mots du vocabulaire naturel appelés « mot-clé » permet d'obtenir une sélection pertinente* [14]. Donc dans l'analyse statistique, la fréquence des données brutes sera prise en compte ainsi que la mesure de la distance entre les mots.

---

[10] Le syntagme nominal est la plus petite unité porteuse de sens dans une phrase.
[11] Se reporter à la revue La tribune des industries de la Langue n°9, 1992
[12] J. CHAUMIER, M. DEJEAN, *L'indexation documentaire, de l'analyse conceptuelle humaine à l'analyse morphosyntaxique*, Documentaliste, vol.27, n°6, novembre-décembre 1990, p.276
[13] ibidem, p.276
[13]ibidem, p. 275

Les méthodes statistiques, ou encore méthodes par extraction, sont basés sur deux types de traitement : le premier est basé sur le calcul de fréquence statistique (avec prise en compte des synonymes (avec dictionnaire ou sur racines)), de fréquence selon une table (table spécialisée par corpus homogène ou table selon la loi de Zipft[15]). Le deuxième type de traitement est construit sur une recherche de voisinage (que l'on appelle également méthodes par co-occurrence) : avec ou sans élimination de polysémies ou avec le calcul de la distance moyenne [CHAUMIER, 90].

La méthode statistique est, en fait basée, sur le mot plein (le lexique). En effet, une fois que tous les mots vides, ceux qui ne portent pas de sens en soi (mots grammaticaux, articles, etc.), sont éliminés, il ne reste que les mots pleins. On tient compte du fait que plus un mot plein est présent dans un texte et plus il est signifiant et servira ainsi de descripteur et pourra apparaître lors d'une interrogation [BLANQUET, 94].

2.2.3.La méthode mixte

Aujourd'hui de nombreux outils, logiciels ou systèmes, selon la dénomination de leurs concepteurs, sont basés non pas sur une méthode d'indexation mais combinent plusieurs analyses en même temps. Il s'agit pour la plupart d'une combinaison linguistique (morphologie, syntaxe, lexique voire sémantique) et statistique. On peut citer quelques logiciels qui combinent ces systèmes : SPIRIT de T-GID, STAND d'IBM France. Ces systèmes permettent une indexation en langage naturel et intègrent des outils d'analyse linguistique multilingue [GACHOT, 95]. Ces outils sont évidemment plus performants que les autres.

Ainsi SPIRIT (Système Syntaxique et Probabiliste d'Interrogation et de Recherche de l'Information Textuelle), et MICRO-MIND, développé par la suite par la même société, représentent un système de documentation automatique dont toutes les procédures de l'indexation à l'interrogation sont entièrement automatisées, l'interrogation se fait en langage naturel [ANDREEWSKI, 87]. Il a connu un grand succès et a été adopté par de nombreux organismes importants (CEA, EDF-GDF, AUPELF, etc.). Il est composé de quatre modules :

[15] La loi de Zipf est la suivante : si l'on dresse une table de l'ensemble des mots d'un texte quelconque, classés par ordre de fréquence décroissante on constate que la fréquence d'un mot est inversement proportionnelle à son rang dans la liste. La loi est représentée de la façon suivante : f*r=c (f=fréquence, r=rang, et c= constante). Cette égalité vraie en approximation est indépendante du locuteur, du type de document et de la langue.

- un module linguistiques (dictionnaire morphologique, analyseur syntaxique),

- un module fichier inverse qui permet d'accéder aux documents à partir de concepts pondérés et ce dans un temps optimisé,

- un module statistique (ou probabiliste) qui permet d'affecter à chaque concept une fonction de poids informationnel et d'ordonner les réponses en fonction de leur pertinence,

- un quatrième et dernier module d'interrogation en langage naturel et de proximité sémantique.

En plus de cette méthode mixte SPIRIT[16] intègre des dictionnaires généraux (français, anglais, allemand), pour l'interrogation des BDD multilingue.

Il est à noter qu'aujourd'hui la plupart des sytèmes tendent vers une combinaison de différentes méthodes et non pas sur une seule, ils gagnent d'ailleurs à être plus performants ainsi.

2.2.4.La méthode par assignation

Le thesaurus est l'outil habituel de l'indexation, le plus souvent on y trouve trois types de relations : synonymie, hiérarchie et voisinage sémantique. Toutes ces relations sont polysémiques, elles peuvent être représentées sous forme de graphes mais sont différentes des réseaux sémantiques qui eux représentent les différents sens d'un même mot. Les réseaux sémantiques semblent être, sur le plan informationnel, de meilleure qualité car les relations sont définies et non polysémiques. Ils permettent une organisation multiples et multidimensionnelles, le grand avantage c'est que l'utilisateur final peut utiliser ces réseaux pour ses interrogations. Mais les thesaurus comme les réseaux sémantiques doivent être remis à jour fréquemment [CATTENAT, 93].

Les thesaurus peuvent être considérés comme une représentation des connaissances, ils sont utilisables si l'on connaît bien le domaine à partir duquel ils sont établis. De ce fait, ils ne

[16] La société SYSTEX, qui gère SPIRIT avec la société T-GID, a participé à différents programmes européens pour le développement et l'amélioration d'outils linguistiques et autres. A partir de 1990, et ce pendant trois ans, SYSTEX a participé au projet EMIR (European Multilingual Information Retrieval) pour l'élaboration d'un logiciel de recherche documentaire multilingue dans le cadre du programme ESPRIT2 (European Strategic Program for Research in Information and Technology), avec la collaboration de l'Université de Liège et la société allemande Transmodul.

peuvent pas toujours aider l'utilisateur final à exprimer ou préciser sa question [CATTENAT, 93].

Le thesaurus est néanmoins un outil indispensable qui permet non seulement de contrôler le vocabulaire d'indexation mais aussi de gérer les problèmes de relations sémantiques entre les termes. Chaque domaine a un thesaurus spécifiques par exemple un thesaurus destiné à l'indexation et la recherche d'image doit, en plus des relations classiques d'un thesaurus (c'est-à-dire hiérarchique), multiplier les liens d'association, d'implication, et d'équivalence entre les descripteurs. On peut concevoir des liens entre les termes abstraits et les indices visuels qui les évoquent [GUILBAUD, 95]. Pour créer des relations il faudra partir des questions des utilisateurs donc de leur vocabulaire [CATTENNAT, 93].

En effet, depuis une dizaine d'années on voit apparaître des modules de gestion de thésaurus associés à des logiciels de stockage et de recherche documentaire. Ces modules permettent l'enregistrement des termes et celui de leurs descripteurs (par exemple non-descripteur vers descripteur, descripteur spécifique vers descripteur générique). Ils permettent également de produire automatiquement les relations inverses (générique à spécifique). En plus de la gestion du thésaurus, ces modules permettent de faire un tri alphabétique des termes et d'en éditer une liste complète structurée sur un niveau hiérarchique comme GOLEM ou BASIS ou sur tous les niveaux hiérarchiques MISTRAL [VAN SLYPE, 87].

On trouve également des logiciels de gestion de thésaurus multilingue comme BASIS, ou encore des logiciels qui permettent une assistance interactive pour la construction de thésaurus. C'est le cas d'ALEXIS de ERLI qui permet d'enregistrer les termes et les relations, de valider immédiatement chaque entrée et d'émettre un message quand cela est nécessaire, de mettre à jour le fichier, d'enregistrer les modifications.

Ainsi dans le cas d'une gestion automatisée d'un thésaurus, il s'agit de comparer les mots clés avec les descripteurs d'un thesaurus, l'indexation automatique simule ici l'indexation manuelle.

2.3.Les autres méthodes

2.3.1.L'indexation automatique des titres

Au départ les ordinateurs ont été utilisés en documentation pour préparer et éditer automatiquement différents types d'index, ces index servaient par la suite d'instruments manuels de recherche rétrospectives ou de support de diffusion de l'information. Les index KWIC (Key Word In Context) sont une solution intéressante à la recherche et à la diffusion de l'information. Ces index sont, en fait, produits directement par l'ordinateur à partir de données mémorisées et sans l'intervention d'un indexeur. Comment s'effectue dans ce cas l'indexation ? Elle se fait sur les mots du langage naturel contenus dans le texte en l'occurrence le titre, qui a été retenu et qui doit servir de base à l'édition de l'index. Le cas le plus fréquemment utilisé, pour ce type d'index, est celui des titres d'articles de périodique car plus explicite et plus précis en général que les autres types de documents, dans ce cas, on parle de Key Word In Title (KWIT) c'est-à-dire de mot clé du titre. Dans ce type d'index, on exclut les mots vides de la même manière que pour la méthode statistique.

Beaucoup d'index sont préparés selon la méthode KWIC, ils permettent, en effet, la diffusion d'un volume important de référence, de nombreuses publications célèbres sont réalisées à partir de cette méthode, on peut citer Chemical titles ou encore Biological abstracts [CHAUMIER, 94].

Les index KWIC et KWIT sont identiques, mais il existe une autre variante de cette méthode celle du Key Word Out of Context (KWOC) c'est-à-dire *le mot clé hors contexte*. Le principe est le même : élimination des mots vides et permutation des termes significatifs, la seule différence réside dans la présentation du mot au sein de l'index. En effet, le mot sur lequel porte la permutation n'est plus mis en évidence en colonne centrale mais placé en en-tête des titres comme des vedettes matières. La méthode KWOC est plus avantageuse dans le sens où il n'y a pas de limitation de caractères et permet une meilleure lisibilité, en revanche l'inconvénient de ces méthodes est qu'elles ne sont pas d'utilisation pratique.

2.3.2.Les systèmes experts et l'indexation automatique

Le développement de l'intelligence artificielle a permis la conception de systèmes informatiques différents des systèmes classiques, ce sont les systèmes experts. Ces nouveaux systèmes sont basés sur l'exploitation dans un domaine particulier des connaissances explicites

et organisées, et peuvent se substituer à un expert humain. Ces systèmes ont fait leur apparition dans les années 1970.

On peut citer COALSORT qui permet une assistance à l'indexeur tout comme JAKS qui en plus traite le texte intégral. Mais aucun de ses systèmes ne traite le langage naturel sauf RIME[17], conçue à partir d'une méthode déterministe basée sur le traitement du langage naturel. [VANDEUR, 90]

On peut dire qu'à l'instar des systèmes à bases de connaissances comme les systèmes experts, les thesaurus peuvent également servir de base de connaissances dans un domaine [MONTEIL, 95].

Il est à noter que l'indexation automatique basée sur un système expert, ne peut se passer de l'intervention humaine, et on parlera plus volontiers dans ce cas d'indexation assistée par ordinateur.

3.L'avenir de l'indexation automatique

3.1 Evaluation de ces systèmes

L'évaluation de l'indexation automatique et de l'efficacité des systèmes en texte intégral soulève de nombreuses questions théoriques et pratiques. Comme dans tout système, l'indexation automatique n'est jamais parfaite, cependant elle permet beaucoup de facilité. Cependant l'indexation automatique a tendance à générer beaucoup d'index qui ne sont pas toujours faciles à gérer. De plus, le traitement de la polysémie est difficile et passe par des analyses pragmatiques qui obligent à rassembler des connaissances sémantiques importantes.

Pour définir les méthodes d'évaluation de l'indexation, il est important de rappeler que l'indexation documentaire n'a de sens que par rapport à la recherche documentaire où à des services plus ou moins dérivés telle que la diffusion sélective d'information. C'est une seule et même chose car il s'agit de faire correspondre les termes de la requête avec ceux utilisés à l'indexation. R. FIDEL indique que l'indexation et la recherche documentaire sont les deux faces d'une même pièce : (...) *indexing and searching are two sides of the same coin*[18].

[17] RIME est spécialisé dans la gestion et l'archivage d'imagerie médicale, il est distribué par EURODIM
[18] Raya FIDEL, *User-centered indexing*, JASIS september 1994, p.575

Ainsi comme on ne peut pas évaluer directement un système d'indexation, on est obligé de passer par la recherche documentaire : on procède donc à une recherche et on évalue ensuite la pertinence des références obtenues en réponse par rapport à la requête posée. Cette évaluation s'articule autour de deux phases : la première concerne une macro-évaluation qui permet de mesurer globalement le rappel (documents signalés mais ne répondant pas à la question) et la précision (proportion de documents répondant à la question) d'un système documentaire. La seconde est une micro-évaluation qui consiste à analyser les résultats obtenus d'un échantillon de quelques centaines de recherches documentaires, c'est donc une évaluation limitée à un ensemble de requêtes sur un échantillon de références et non comme au niveau de la macro-évaluation, une évaluation du système documentaire par le biais des requêtes posées directement au systèmes. On peut ainsi identifier les causes de dysfonctionnement.

Cette évaluation permet d'apprécier la qualité d'indexation d'un système par le biais des descripteurs utilisés d'une part et d'autre part de voir quels sont les documents qui ont été mal indexés [VAN SLYPE, 87].

Une autre méthode d'évaluation a été testé par le Service d'Information et de Documentation (SID) de l'EDF qui a mis en place, en 1992, un circuit d'indexation automatique des rapports internes pour pouvoir analyser les dysfonctionnements de l'indexation automatique et améliorer la qualité de celle-ci. Les documentalistes ont ainsi pu révisé l'indexation automatique de ces rapports, l'objectif était de passer à l'automatisation totale de l'indexation des documents à la fin de l'année 1994 [MONTEIL, 95].

Le SID de l'EDF a en effet mis en place un circuit expérimental en parallèle du circuit d'exploitation normal du fonds documentaire. Ceci a permis aux documentalistes d'évaluer la qualité du résultat de l'indexationn automatique du titre, du résumé et des mots clés libres (proposés par les auteurs). Ils ont pu identifié plusieurs améliorations à mettre en œuvre dont l'enrichissement linguistique du système d'indexation. Ils ont réalisé, selon le même principe, une nouvelle évaluation de la version améliorée de l'outil, qui a mis en évidence l'utilité des filtrages des résultats et le travail nécessaire à apporter au thésaurus, à partir duquel s'effectue l'indexation. Ces filtrages statistiques et ces améliorations apportées au thésaurus ont abouti à l'exploitation de l'indexation automatique en 1995.

Il est a remarquer que l'évaluation des systèmes d'indexation automatique est difficile à mettre en place car le manque d'homogéneïsation de ces systèmes ne permet pas de concevoir une évaluation unique.

3.2.Les attentes des professionnels par rapport à ces systèmes d'indexation automatique

Les documentalistes spécialisés en indexation n'ont certes pas vu d'un très bon œil l'arrivée de ces systèmes qui menaçaient leur emploi. C'est peut-être aussi pour cela qu'on parle plus souvent d'indexation assistée par ordinateur que d'indexation automatique, tout simplement pour conforter une profession.

Les méthodes d'analyse linguistique progressent, permettant ainsi une analyse de plus en plus fine des textes. De même, les méthodes et les outils documentaires évoluent offrant ainsi des applications diverses. Concernant les applications, l'enrichissement de la documentation interne et l'accès à la documentation externe constituent des champs d'expérimentation nouveaux : constitution d'un patrimoine interne, contrôle de la qualité de la documentation, veille technologique. On peut dire qu'à la fonction d'analyste du documentaliste s'ajoute désormais celle de gestionnaire des terminologies de l'entreprise [MONTEIL, 95].

CONCLUSION

Nous avons vu un panorama des méthodes d'indexation automatique ainsi que l'évolution qu'a impliquée l'entrée de l'informatique dans le domaine de la documentation :

- un gain de productivité que peut permettre l'indexation par rapport au volume d'information à traiter souvent croissant,

- une facilité et une rapidité à réindexer une base entière après mise à jour du langage documentaire sur lequel l'indexation automatique s'appuie,

- une meilleure qualité (même si la qualité est très variable) à l'inverse de l'indexation humaine, qui est moins stable au sens où un même document n'est pas indexé de la même manière par deux personnes à deux moments différents. De plus, l'intérêt pour l'indexation automatique grandit lorsque le renouvellement des indexeurs humains est important dans un service de documentation [LE LOADER, 94]. Elle permet ainsi de réduire les coûts de l'indexation humaine.

Aujourd'hui, l'arrivée d'internet, la multiplication des bases et banques de données en texte intégral et les enjeux économiques que cette activité documentaire représente, ont poussé certaines entreprises à investir dans des systèmes d'information automatisés : indexation des courriers, développement de l'intranet, bases de données en langage naturel pour le grand public, etc.

Après une époque centrée sur les méthodes d'organisation des connaissances, nous sommes rentrés dans celle de la recherche d'information pertinente, et le développement d'outils capable d'indexer correctement et de trouver une information utile est un engagement non négligeable dans une société d'information.

GLOSSAIRE

Analyse morphologique : étude de la forme et des variations des mots.

Analyse pragmatique : étude de la signification en fonction du contexte.

Analyse sémantique : étude du sens au plan du lexique, cette analyse traite aussi de la synonymie et de l'hyperonymie (relation d'inclusion qui associent des termes spécifiques à des termes génériques).

Analyse syntaxique : étude des relations que les mots entretiennent entre eux au niveau de la phrase.

Base de données : ensemble de données organisées en vue de son utilisation par des programmes correspondant à des applications distinctes et de manière à faciliter l'évolution indépendante des données et des programmes.

Base de connaissances : partie d'un système expert contenant l'ensemble des informations, en particuliers des règles et des faits qui constituent le domaine de compétence du système.

BNB : British National Bibliography (Bibliographie Nationale Britannique)

Bruit : proportion des documents proposés lors de l'interrogation bien que ne répondant pas à la question.

IAO : Indexation Assistée par ordinateur

IBM : International Business Machine

Intelligence artificielle : discipline relative au traitement par informatique des connaissances et du raisonnement.

KWIC / Key Word in Context : mot clé dans le contexte

KWIT : Key Word in Title : mot clé dans le titre

KWOC : Key Word out of Context : mot clé hors contexte

Langage naturel : en informatique, on parle de langage naturel dès lors que l'on considère un langage non seulement du point de vue informatique mais aussi linguistique. Ce langage

naturel subit toujours une normalisation ou une modélisation et n'est jamais la reproduction d'un langage parlé ou écrit, tel que l'utilise normalement un locuteur.

Lemmatisation : processus de réduction terminologique ou l'on ramène le mot à la forme canonique (infinitif pour les verbes, masculin singulier ou féminin singulier pour les adjectifs et les substantifs).

Mot clé : descripteur extrait du texte qu'il caractérise ou d'un thésaurus.

Moteur d'inférence : partie d'un système expert qui effectue la sélection et l'application des règles en vue d'un problème donné.

PASSAT : Programm zur Automatischen Selektion von Stichwötern Aus Texten

PIAF-DOC : Programme Interactif d'Analyse du Français

PRECIS : Preserved Context Indexing System (système d'indexation respectant le contexte)

Réseau sémantique : cartographie représentant les différents sens d'un mot.

SIE : Société d'Information Européenne

Silence : proportion des documents non proposés lors de l'interrogation bien que répondant à la question.

SINTEX : Système d'INdexation de TEXtes

SPIRIT : Système Syntaxique et Probabiliste d'Interrogation et de Recherche de l'Information Textuelle

SYNTOL : Syntagmatic Organization Language

Système expert : ensemble de logiciels exploitant dans un domaine particulier des connaissances explicites ou organisées, pouvant se substituer à un expert humain.

Thésaurus : dictionnaire de mots ou expressions du langage naturel, termes normalisés et préférentiels, organisés d'une manière conceptuelle présentant le terme groupé par affinité sémantique et complété d'indications de relations.

TAO : Traduction assistée par ordinateur

T-GID : Technologies Gestion Informatique Documentaire, société filiale du Groupe Technologies et gérante de SPIRIT avec la société SYSTEX.

Traitement automatiques du langage : discipline qui regroupe toutes les recherches de développement dans le domaine linguistique automatique, appliquée aux langues écrites et plus récemment aux langues parlées.

BIBLIOGRAPHIE

1-Monographie

[BLANQUET, 94] BLANQUET Marie-France, *Intelligence artificielle et système d'information*, Collection Systèmes d'information et nouvelles technologies, Paris : ESF Editeurs, 1994, 269p.

[BORKO, 78] BORKO Harold, L. BERNIER Charles, *Indexing concepts and methods*, Library and Information Science Series, Academic Press INC. New York, 1978, 261p.

[CHAUMIER, 77] CHAUMIER, Jacques, *L'analyse documentaire : le traitement linguistique de l'information documentaire*, Paris : Entreprise Moderne d'Edition, 1977, 126p.

[CHAUMIER, 94] CHAUMIER, Jacques, *Les techniques documentaires*, Collection Que Sais-je n°1419, Paris : PUF, 1994, 125p.

[CFCE, 94] Centre Français du Commerce Extérieur, Direction des industries et Services, *Industries de la langue*, Document arrêté en mai 1994, Paris :CFCE, 1994, 76p.

[COYAUD, 67] COYAUD Maurice, SIOT-DECAUVILLE Nelly, *L'analyse automatique des documents*, Paris : Mouton et Compagnie, 1967, 149p.

[HODGE, 92] HODGE, Gail M., *Automated support to indexing*, 1992 NFAIS Report Series n°3, USA, Philadelphia, PA : National Federation of Abstracting and Information Services,1992, 176p.

[LOIS, 85] LOIS Mai Chan, RICHMOND A. Phyllis, SVENONIUS Elaine, *Theory of subject analysis, a source book*, Libraries unlimited Inc. Littleton Colorado, 1985, 415p.

[MANIEZ, 94] MANIEZ Jacques, *Les langages documentaires et classificatoires : conception, construction et utilisation dans les systèmes documentaires*, Paris : Les Editions d'Organisation, 1987, 291p.

[VAN SLYPE, 87] VAN SLYPE George, *Les langages d'indexation : conception, construction et utilisation dans les systèmes documentaires*, Paris : Les Editions d'Organisation, 1987, 277p.

2-Publications en série

[ABDOUN, 91] ABDOUN, Abdelkrim, *La lexicométrie documentaire : contribution à l'utilisation des techniques documentaires comme méthodologie d'études en sciences sociales*, RIST, vol.1, n°2, 1991, pp.87-93

[ALDOUS, 96] ALDOUS K.J., *A system for the automatic retrievial of information from a specialist database*, Information processing & management GBR, 1996,vol. 32 n°2, pp.139-154

[ALEXANDER, 96] ALEXANDER M., *Automatic indexing of document images using EXCALIBUR EFS*, Library technology news, GBR, 1995, n°16, pp. 4-6

[AL-KHARASHI, 94] AL-KHARASHI I. A., EVENS M.W., *Comparing Words, stems and roots as index terms in an Arabic information retrieval system* JASIS N°8, vol.45, september 1994, pp. 548-560

[ANDREEWSKI, 87] ANDREEWSKI Alexandre, FLUHR Christian, *L'intelligence artificielle : une nécessité pour le futur*, Temps futur CSEN, 1987, pp.7-8

[ANDREEWSKI, 96] ANDREEWSKI, Alexandre, *Les systèmes documentaires SPIRIT et MCRO-MIND*, Bulletin du Centre de Hautes Etudes Internationales D'Informatique documentaire, N°61-Mars 1996, CID Paris, pp.29-41

[BOUCHE, 89] BOUCHE, Richard, *Le syntagme nominal, une nouvelle approche des bases de données textuelles* in : Actes du Colloque terminologie et industries de la langue, META, Journal des traducteurs, Montréal, vol. 34, n°3 septembre 1989, pp. 429-434

[CHAUMIER, 90] CHAUMIER Jacques, DEJEAN Martine, *L'indexation documentaire, de l'analyse conceptuelle à l'analyse morphosyntaxique*, Documentaliste, vol.27, n°6, novembre-décembre 1990, pp.275-279

[CHAUMIER, 92] CHAUMIER Jacques, DEJEAN Martine, *L'indexation assistée par ordinateur, principes et méthodes*, Documentaliste, vol.29, n°1, 1992, pp.3-6

[CISSE, 92] CISSE, Sofiane, *Le génie linguistique : une réalité*, Processeurs, 15 mai 1992, pp.29-30

[CLAVEL, 93] CLAVEL Geneviève, Walther Frédéric, WALTHER Joëlle, *Indexation automatique de fonds bibliothéconomie*, ARBIDO-R8, 1993, pp.14-19

[CATTENAT, 93] CATTENAT Annette, PAUL Gérard, *Thésaurus ou réseaux sémantiques pour l'aide à l'interrogation de base de données textuelles par l'utilisateur final*, IDT 93, 6e congrès, 1993, pp. 138-142

[CAULIER, 96] CAULIER Sophie, *La gestion et le suivi des courriers* Archimag, n°95, juin 1996, pp.46-47

[CHARTRON, 89] CHARTRON Ghislaine, DALBIN Sylvie, MONTEIL M.G., VERILLON M., *Indexation manuelle et indexation automatique : dépasser les oppositions*, Documentaliste, vol. 26, n°4-5, 1989, pp.181-187

[CHECROUN, 92] CHECROUN Alain, ELGHOUL M., *L'approche SIAD en Gestion Documentaire*, Bulletin du Centre de Hautes Etudes Internationales, 1992, pp.5-13

[COHEN, 95] COHEN J. D., *Highlights : language and domain-indépendant automatic indexing terms for abstracting*, JASIS N°3, vol.46, april 1995, pp.162-173

[CORET, 91] CORET Annie DUCLOY J., MENILLET D., *Les stations de travail des ingénieurs à l'INIST*, IDT, 1991, pp.189-194

[DAILLE, 95] DAILLE Béatrice, *Repérage et extraction de terminologie par une approche mixte statistique et linguistique*, TAL, vol. 36, n°1-2, 1995, pp. 101-118

Dans la lunette de l'observatoire, La Tribune des Industries de la Langue, juillet/octobre 1990, p.14

[DEBILI, 93] DEBILI Fathi, SAMMOUDA Elyès, ZRIBI Adnane, *Indexation interactive et interrogation multilingues français-anglais-arabe : outils pour la confection d'ouvrages électroniques plurilingues*, IDT 93, 1993, pp.96-99

[DEMAILLY, 92] DEMAILLY André, *Robert Pagès et l'analyse codée*, Documentaliste, vol. 29, n°2, 1992, pp.59-72

[FAVRE, 96] FAVRE G. CHAUVET P., PAVIOT B., *Le traitement de l'identification automatique : l'exemple du courrier*, Archimag, n°95, juin 1996, pp.42-44

[FIDEL, 94] FIDEL R., *User-centered indexing*, JASIS N°8, vol.45, september 1994, pp.572-576

[GACHOT, 95] GACHOT, Isabelle, *Linguistique + statistique + informatique = indexation automatique ?,* Archimag, n°84, mai 1995, pp.34-37

[GUILBAUD, 95] GUILBAUD Elisabeth, *Comment indexer l'image ?,* Archimag, n°86, juillet-août 1995, pp.41-43

[GUIMIER, 93] GUIMIER-SORBET, Anne-Marie, *Des textes aux images : accès aux informations multimédia par le langage naturel*, Documentaliste, vol. 30, n°3, 1993, pp.127-134

[GIROLLET, 92] GIROLLET Dominique, *Quelques aspects entrevues à la SEPLN*, Tribune des Industries de la Langue n°9, 1992, pp.27-29

[GINOUVES, 95] GINOUVES Véronique, PERENNOU Véronique, *L'indexation du son inédit ou comment constituer une base ethnographique*, Archimag n°85, juin 1995, pp. 57-59

Indexation automatique et interrogation de base de données textuelles en language naturel, Tribune des industries de la langue n°10, Novembre 1992, pp. 46-49

Indexation automatique : reconnaître l'écriture (SIAD), Archimag n°76 juillet-août 1994, p.13

Interroger des banques de textes, Revue Traitement de texte, mai 1984, pp.3-5

[KSIBI, 95] KSIBI, Ahmed, *Les préliminaires à la participation des pays en voie de développement dans les réseaux internationaux d'information*, IDT, 1995, pp.77-79

L'alliance du texte intégral et du langage naturel, Archimag n°88, octobre 1995, p.35

La rencontre de la linguistique et de l'informatique, La Recherche, octobre 1985, p.6

[LAYNE, 94] LAYNE S. S., *Some issues in the indexing of images*, JASIS N°8, vol.45, september 1994, pp. 258-260

[LE GUERN, 91] LE GUERN Michel, *Un analyseur morpho-syntaxique pour l'indexation automatique*, Le Français moderne, 1991, tome LIX, n°1, pp. 22-35

[LE LOARER, 94] LE LOARER Pierre, *Indexation automatique, recherche d'information et évaluation*, IDT 94, Paris, 1994, pp. 266-277

[LE LORAER, 96] LE LOARER Pierre, NORMIER Etienne, *Techniques linguistiques et statistiques pour sélectionner l'information pertinente*, IDT 96, Paris, 1996, pp. 115-120

[LORETTE, 96] LORETTE, Guy, *Le traitement automatique de l'écrit et du document : état de la recherche*, Documentaliste, vol.33, n°4-5, 1996, pp.214-217

[MATEUS, 92] MATEUS, Marie-Hélène, *L'ingénierie linguistique au Portugal*, Tribune des Industries de la Langue n°9, 1992, pp.26-27

[MILSTEAD, 94] MILSTEAD J. L., *Needs for research in indexing*, JASIS N°8, vol.45, september 1994, pp.574-576

[MITRI, 95] MITRI M., *Combining semantic networks with multi-attribute utility models : an evaluate data indexing method* ; Expert systems with applications, GBR, 1995, vol. 9, n°3, pp. 283-294

[MONTEIL, 95] MONTEIL Marie-Gaëlle, *Indexation manuelle et automatique : comparaison et perspectives*, IDT 95, 12e Congrès, Paris, 12-15 juin 1995, pp. 214-215

[NIEDERCORN, 92] NIEDERCORN Frank, *Table ronde autour de SPIRIT*, 01 Informatique, 09/02/1990, pp.12-13

[PARENT, 92] PARENT, Richard, *DELTA, le français à la faveur des nouvelles technologies de communication*, La Tribune des industries de la Langue, n°9, 1992, pp.32-35

[PUGEAUT, 95] PUGEAUT F., MONTEIL M.G., *Une étude pour l'extraction d'index structurés à la DER*, Génie linguistique 95 : Montpellier, 27-30 juin 1995, pp. 165-175

[RANJARD, 95] RANJARD Sophie, *Indexer et résumer, pourquoi et comment ?* Archimag, n°80, décembre-janvier 1995, pp.41-43

[RAPHAEL, 96] RAPHAEL B., KUMAR B., *Indexing and retrieval of cases in a case design system*, Artificial intelligence for engineering design, analysis and manufacturing, GBR 1996, vol. 10, n°1, pp. 47-63

[RIVIER, 90] RIVIER Alexis, *Construction des langages d'indexation*, Documentaliste, n°27, n°6, novembre-décembre 1990, pp.263-274

[SOERGEL, 94] SOERGEL D., *Indexing and retrieval performance*, JASIS N°8, vol.45, september 1994, pp. 589-599

[SVENONIUS, 94] SVENONIUS E., *Acces to nonbook materials*, JASIS N°8, vol.45, september 1994, pp. 600-606

SYSTEX : Un itinéraire exemplaire, SYSTEMES EXPERTS, décembre 1991, p. 21

[TIBBO, 94] TIBBO H. R., *Indexing for the humanities*, JASIS N°8, vol.45, september 1994, pp. 607-619

[VANDEUR, 90] VANDEUR, Marc, *Approche de l'élaboration d'un système expert d'aide à l'indexation*, Cahier de la Documentation-Bladen Voor de Documentatie, n°4, 1990, pp.75-90

[WELLISCH, 94] WELLISCH H. H. *Book and periodical indexing*, JASIS N°8, vol.45, september 1994, pp. 620-627

[WOODRUFF, 94] WOODRUFF A. G., PLAUNT C., GIPSY *: Automated geographic indexing of text documents*, JASIS N°9, vol.45, october 1994, pp. 645-655

[ZADO, 92] ZADOUNAISKY, Stéphane, *L'empire des signes*, Tribune des Industries de la Langue n°9, 1992, pp.17-26

3-Thèses

[AIT HAMLAT, 83] AIT HAMLAT, Akila, *Applications des méthodes statistiques à l'indexation automatique de documents*, thèse de doctorat de 3^{e} cycle, Paris 6 : 1983

[BERRUT, 88] BERRUT Catherine, *Une méthode d'indexation fondée sur l'analyse sémantique de documents spécialisés : le prototype RIME et son application à un corpus médical*, Thèse de doctorat d'Etat : Sciences et Techniques communes : sciences de l'information et documentation, Grenoble 1 : 1988

[EYMARD, 92] EYMARD, Gilbert, *Traitement documentaire des sommaires : des mots clés à l'extraction de connaissances : application à une documentation technique*,

Thèse : Sciences information et communication, documentation et sémiologie, Grenoble 2 : 1992

[GREFENS, 83] GREFENS Grégory, *Traitements linguistiques orientés vers la documentation automatique*, Thèse de doctorat de 3e cycle, Paris 11 : 1983

[LAINE, 82] LAINE Sylvie, *Extraction et sélection de descripteurs complexes dans un ensemble de textes pour leur indexation automatique*, Thèse de doctorat d'ingénieur : Sciences et Techniques communes : sciences de l'information et documentation, Lyon 1 : 1982

[LALLICH, 86] LALLICH BOIDIN Geneviève, *Analyse syntaxique automatique du français écrit : applications à l'indexation automatique*, Thèse de doctorat : Mathématiques appliquées aux sciences sociales : linguistique, information, documentation, Grenoble 2 : 1986

[LAROUK, 93] LAROUK Omar, *Extraction des connaissances à partir de documents textuels : traitement automatique de la coordination (connecteurs et ponctuation*), Thèse de doctorat spécialité informatique, Lyon1 : 1993

[MERLE, 82] MERLE Alain, *Un analyseur pré-syntaxique pour la levée des ambiguïtés dans les documents écrits en langage naturel : application à l'indexation automatique*, Thèse de doctorat d'ingénieur : Sciences et Techniques communes, sciences de l'information et documentation, Grenoble : 1982

[METZGER, 88] METZGER Jean-Paul, *Syntagmes nominaux et information textuelle : reconnaissance automatique et représentation*, Thèse de doctorat d'Etat : Sciences et Techniques communes : Sciences de l'information et documentation, Lyon 1 : 1988

[SEGUIN, 77] SEGUIN Gérard, *Génération automatique du vocabulaire représentatif d'un domaine : essais d'indexation automatique*, Thèse de doctorat de 3e cycle : Sciences et techniques communes : sciences de l'information et documentation, Lyon 1 : 1977