E.N.S.S.I.B.
ECOLE NATIONALE SUPERIEURE
DES SCIENCES DE L'INFORMATION
ET DES BIBLIOTHEQUES

UNIVERSITE
CLAUDE BERNARD
LYON I

DESS en Informatique Documentaire

Rapport de recherche bibliographique :

OUTILS D'ANALYSE AUTOMATIQUE DE « NEWS » OU DE FORUMS ÉLECTRONIQUES A DES FINS DE VEILLE

Christine MILLOT

Sous la Direction de
Monsieur Luc GRIVEL
Ingénieur informaticien au Service
Programmes de Recherche Infométrie
INstitut de l'Information Scientifique et Technique
(INIST)

1997

E.N.S.S.I.B.
ECOLE NATIONALE SUPERIEURE
DES SCIENCES DE L'INFORMATION
ET DES BIBLIOTHEQUES

UNIVERSITE
CLAUDE BERNARD
LYON I

DESS en Informatique Documentaire

Rapport de recherehe bibliographique :

OUTILS D'ANALYSE AUTOMATIQUE DE « NEWS » OU DE FORUMS ÉLECTRONIQUES A DES FINS DE VEILLE

Christine MILLOT

Sous la Direction de
Monsieur Luc GRIVEL
Ingénieur informaticien au Service
Programmes de Recherche Infométrie
INstitut de l'Information Scientifique et Technique
(INIST)

1997

Remerciements

Je tiens à remercier Luc GRIVEL, ingénieur informaticien au service Programmes de Recherche Infométrie de l'INIST (INstitut de l'Information Scientifique et Technique), pour l'étude qu'il a bien voulu me confier et l'aide qu'il m'a apportée pour la réalisation de ce travail.

Outils d'analyse automatique de « news » ou de forums électroniques à des fins de veille.

Christine MILLOT

Rapport de recherche bibliographique demandé par Luc GRIVEL, ingénieur informaticien au Service Programmes de Recherche Infométrique de l'INIST (INstitut de l'Information Scientifique et Technique) – 2, allée du Parc de Brabois – 54500 VANDOEUVRE.

Résumé :

Avec la prolifération des flux d'information, notamment sur l'Internet, les outils de recherche ne semblent plus adaptés. L'utilisateur désire une réponse ciblée à sa question. Pour palier à ces problèmes, divers entreprises et laboratoires de recherches ont développé des outils pour trier et élaborer l'information, qu'elle soit au préalable structurée ou non. Ces informations deviennent alors des renseignements permettant une veille stratégique.

Mots-clés :

Internet, veille stratégique, analyse de données, traitement de l'information, classification automatique.

Abstract :

With the increase of information, specially with the Internet, research tools don't apear adapted. The users would like exact answres in order to solve their problems. So, several companies and research centers have developped tools to select and elaborate structured and unstructured information. Then, these ones allow an exploitation in competitive intelligence.

Keywords :

Internet, competitive intelligence, data analysis or data mining, information processing, automatic classification or cluster analysis.

Introduction Générale

Le travail présenté dans ce rapport repose sur l'analyse de l'existence "potentielle" de logiciels de veille portant sur les « news » ou les forums électroniques. Cette recherche a été commanditée par Luc GRIVEL, ingénieur informaticien au service Programmes de recherche infométrique de l'INIST (INstitut de l'Information Scientifique et Technique).

Ce service a, au préalable, développé des outils d'aide à la veille stratégique. Ce sont des programmes infométriques de classification automatique avec une représentation de l'information sous forme de cartographie permettant la navigation hypertextuelle et ce, partant d'une information secondaire structurée (les références bibliographiques). Aujourd'hui, ce service désire s'orienter vers le développement de logiciels de veille portant sur les « news » autrement du texte libre.

C'est dans cette perspective qu'a été orienté ce travail. Ce rapport s'articule autour de trois points. Dans un premier temps, je m'attacherai à vous décrire la méthodologie de recherche ainsi que la stratégie adoptée. Puis, je présenterai succinctement le résultat de la recherche afin de définir les caractéristiques des références obtenues. En dernier lieu, vous trouverez une synthèse élaborée à partir de huit références choisies dans la bibliographie pour leur publication récente et leur analyse portant sur de l'information non structurée.

1. *Méthodologie*

1.1 Stratégie de Recherche

Rappelons que l'objectif de cette bibliographie est de recenser les logiciels de veille sur un corpus de type « news », que ce travail s'ancre dans un projet : celui de concevoir et de développer un outil permettant l'analyse des « news » ayant pour objectif une exploitation en veille.

1.1.1 Présentation du sujet

Afin d'appréhender correctement cette étude, mes connaissances préalables ainsi que le travail effectué au sein du Service Programmes de Recherche Infométrie m'ont été d'un grand recours. En effet, une connaissance des concepts de veille, de scientométrie, infométrie et bibliométrie permettent de posséder les notions de base pour l'approche de tels outils.

Il s'agit, ici, de présenter succinctement ces concepts et de montrer les multiples facettes qu'ils recouvrent. Cette approche a été réalisée d'après les lectures suivantes :

- COURTIAL (Jean-Pierre) .- *Introduction à la scientométrie : de la bibliométrie à la veille technologique.*- Paris : Economica, 1990.- 137 p. (Anthropos)
- DESVALS (Hélène), DOU (Henri).- *La veille technologique : l'information scientifique, technique et industrielle.*- Paris : Dunod, 1992.- 429 p.
- JAKOBIAK (François).- *Exemples commentés de veille technologique.*- Paris : Ed. d'Organisation, 1992. 198 p.
- MARTINET (Bruno), RIBAULT (Jean-Michel).- *La veille technologique, concurrentielle et commerciale.*- Paris : Ed. d'Organisation, 1989.- 300 p. (Collection Hommes et Techniques).
- POLANCO (Xavier).- « Analyse stratégique de l'information scientifique et technique : construction de clusters de mots-clés ».- *Sciences de la société*, n°29, 1993, p.111-126.

Notion de veille :

- ***Veille scientifique et technique :***

La veille en information scientifique et technique permet de mesurer quantitativement et qualitativement l'information stockée (actualité, origine, ...) et de structurer l'information thématiquement sans passer par un code de classement préétabli et figé afin de suivre le développement de la recherche tel qu'il se présente au niveau de la littérature scientifique. Les sources formelles de cette veille concernent principalement les données scientifiques concernant la recherche et le développement.

- ***Veille technologique :***

La notion de veille technologique vient d'une double volonté, sociologique et technique. Sociologiquement, elle provient d'une triple nécessité : nécessité pour les scientifiques de publier pour survivre ; nécessité pour une entreprise de protéger ses acquis (propriété industrielle) ; et enfin, nécessité de conserver le patrimoine scientifique et technique. Ainsi, tous ces documents produits serviront à cette veille. Techniquement, grâce à l'impact des nouvelles technologies, elle facilite l'accès aux informations par l'intermédiaire de bases de données textuelles et iconographiques ; la liaison entre différentes sociétés ou personnes par l'intermédiaire de réseaux télématiques de plus en plus performants ; et enfin, elle favorise le traitement des informations par la micro-informatique. L'observation et l'analyse de l'environnement permettent ainsi une diffusion bien ciblée des informations sélectionnées et traitées, utiles à la prise de décision stratégique.

- ***Veille concurrentielle et commerciale :***

Cette veille consiste en la surveillance des concurrents actuels et potentiels. L'intérêt que l'on porte à un concurrent peut être très divers. Si l'on s'intéresse à son équipement industriel, les techniques de veille concurrentielle s'apparenteront à de la veille technologique. Par contre, si l'on s'intéresse à ses clients, la veille concurrentielle s'apparentera à de la veille commerciale, voire même au marketing. La veille concurrentielle montre l'interaction qu'il peut y avoir entre tous les types de veille.

- ***Veille stratégique :***

Cette expression de veille stratégique permet de faire intervenir d'autres notions que celle de l'information scientifique et technique comme l'aspect concurrentiel et commercial. Il faut reconnaître que ces trois aspects technologique, concurrentiel et commercial sont plus au moins liés. On a du mal à délimiter les frontières de chacun de ces trois types de veille dans une application. La plupart du temps, les trois interviennent mais à des degrés différents.

Notions de scientométrie, de bibliométrie et d'infométrie :

Il s'agit, ici, de définir les concepts de scientométrie, de bibliométrie et d'infométrie. La scientométrie est l'étude de la production d'articles scientifiques et elle sert à mettre en évidence des lois bibliométriques sur la production scientifique et sur la citation. La bibliométrie est à la fois le comptage de tout ce qui peut entrer dans une bibliothèque scientifique et une approche quantitative des techniques de gestion d'une bibliothèque. Enfin, l'infométrie englobe les deux concepts précédents, elle concerne toutes les mesures faites en sciences de l'information.

1.1.2 Etude préalable

Après une première rencontre avec Luc GRIVEL, la démarche de la recherche se dirigea vers la constitution d'une bibliographie recensant les logiciels de veille sur des informations de type « news » ou forums électroniques. Dès le début, mon commanditaire m'a fourni plusieurs pistes (centres de recherche étant susceptibles de travailler sur le sujet) :

- Pôle Universitaire Léonard de Vinci à Paris : Patrick CONSTANT, Olivier JOUVE et Claude VOGEL ;
- CERESI-CNRS à Meudon : Equipe TURNER, BORZIC, Mathilde de SAINT-LEGER avec l'outil DYNATOOLS ;
- CRRM à Marseille : Equipe d'Henri DOU.

Ici, il faut remarquer qu'il ne s'agit pas de recenser tous les logiciels de veille existant mais il s'agit de trouver ceux s'appliquant à un corpus de type « news » ou forums. En effet, ce type d'information est beaucoup plus difficile à exploiter, informatiquement parlant, que les références bibliographiques qui est une information secondaire structurée. Ici, se pose le problème d'une information non structurée, il n'est donc pas évident de définir des balises permettant d'identifier le titre, l'auteur, l'année, etc....

J'ai cherché à identifier les équipes de recherche travaillant sur des outils d'analyse de corpus de « news » ou de forums à des fins de veille. Or, on s'est aperçu que les recherches étaient assez limitées et ce, suivant plusieurs hypothèses :

- soit les travaux ne sont pas encore assez avancés dans ce domaine en ce qui concerne la France ;
- soit les équipes de recherche souhaitent encore tenir au secret l'évolution de leur recherche.

1.1.3 Problèmes soulevés et Nouvelles orientations de la recherche

Lors des premières recherches, j'ai été confrontée à plusieurs problèmes :

- problème cité précédemment : peu de publications sur ce sujet ;
- problème de bruit avec le mot veille et surtout avec le mot « news ». Il est donc nécessaire d'affiner le système de requêtes à l'aide de termes supplémentaires qui serviront de filtres.

Pour le premier point, la recherche s'est donc orientée vers des méthodes à la fois statistiques et linguistiques d'analyse automatique de ce type d'information. La question à se poser est donc :

Quels sont les outils de statistique, d'indexation,
de classification et de cartographie
pour traiter ces données ?

Par conséquent, il a fallu déterminer les étapes d'analyse et d'élaboration d'un tel logiciel c'est-à-dire les méthodes pouvant être utilisées. A ce niveau, mon travail préalable pour le Service "Programmes de Recherche Infométrie" de l'INIST (INstitut de l'Information Scientifique et Technique) m'a permis de déterminer les différentes étapes de conception :

1) Analyse scientométrique, bibliométrique, infométrique (ou statistique) ;
2) Indexation automatique : analyse syntaxique, sémiotique, sémantique ;
3) Classification automatique : analyse des mots associés ou autres méthodes ;
4) Cartographie et navigation hypertexte : représentation finale des données.

Et ce, appliqué aux « news » ou plus largement au texte libre.

En ce qui concerne le second point, pour le terme « news », certaines recherches ont dû s'effectuer avec des substituts de ce terme tel que : « newsgroup », texte libre c'est-à-dire information non structurée ou information en ligne. D'autre part, le terme « veille », utilisé seul, était porteur d'énormément d'informations non pertinentes :

Ex. : la veille du 12 décembre

Afin de palier ce problème, ma recherche s'est orientée vers les termes : veille technologique ou veille stratégique.

Enfin, un dernier handicap linguistique et non des moindres est celui de la traduction du concept de classification automatique en anglais et en américain. En effet, en anglais cette expression se dit « automatic classification » alors qu'aux Etats-Unis, on dit « cluster analysis ». Le terme « classification » référant dans la majorité des cas aux termes français « classement » : en interrogeant sur ce mot-clé, les textes trouvés concernaient donc la plupart du temps le plan de classement CDU ou DEWEY. « *Attention aux faux-amis : beaucoup d'auteurs américains se sont mis à employer le terme classification dans un sens différent du mot français, sens qui correspondrait plutôt à notre terme classement.* » [68]

1.1.4 Etapes de la recherche

Mes recherches se sont déroulées en plusieurs étapes :

1) Tout d'abord, elles se sont orientées vers les thèses afin de connaître l'état des travaux actuels. Ceci s'est fait grâce au CD-ROM Docthèses. Cette première interrogation m'a également permis de mieux cerner les mots-clés potentiels.
2) En second lieu, les recherches se sont orientées vers Internet pour localiser des sites se référant aux pistes données par le commanditaire, et également pour élargir la recherche vers d'autres laboratoires.
3) Ainsi, après identification de plusieurs laboratoires de recherche, de mots-clés ainsi que d'auteurs, cette troisième étape s'est effectuée sur DIALOG et sur les CD-ROM LISA Plus, PASCAL ainsi que celui de la BN-Opale.
4) Suite aux résultats obtenus j'ai désiré peaufiner ma recherche sur Internet afin d'obtenir plus de publications récentes et plus d'information sur les recherches américaines.

1.2 Les Outils utilisés

1.2.1 Recherche sur CD-ROM

1.2.1.1 CD-ROM utilisés :

Mes recherches se sont effectuées sur trois CD-ROM :

- Celui de la BN-Opale : cette base réunit 800 000 notices bibliographiques d'ouvrages entrés par dépôt légal à la Bibliothèque Nationale de France depuis 1970 à nos jours. En plus des ouvrages imprimés, elle comporte également des notices des publications officielles des grands organismes internationaux. Enfin, elle recense les notices de publications en série :
  - notices de périodiques édités en France reçus à la bibliothèque Nationale de France (titres vivants et titres nés à partir de 1960 et ayant cessé de paraître).
  - notices de collections éditées en France reçues à la Bibliothèque Nationale de France.

  Remarque par rapport au sujet : cette base reste trop générale pour une recherche telle que la nôtre. La politique d'acquisition diffère de nos préoccupations sur l'aspect de l'état de l'art en ce qui concerne les recherches informatiques. Il faut également souligner la prédominance des publications françaises ; or, notre sujet s'intéresse également aux recherches étrangères telle que la position américaine sur ce domaine.

- Docthéses : cette base est le catalogue des thèses soutenues dans les universités françaises. Ce CD-ROM recense les thèses de doctorat soutenues en France :
  - depuis 1972 pour les Lettres, Sciences Humaines et Sociales et les Sciences ;
  - depuis 1983 pour les disciplines de la santé (à l'exception des thèses de Médecine vétérinaire, à partir de 1990).

  La localisation permet de connaître l'adresse de la bibliothèque de dépôt et celles qui possèdent une version sur microfiche.

  Remarque par rapport au sujet : Très bonne couverture de la littérature de langue française pour les thèses. Cette base donne une idée de l'état de la recherche concernant notamment les logiciels de veille. Elle permet également de déceler quelles sont les équipes (laboratoires) qui travaillent sur ce sujet : Qui supervise la thèse ?

- LISA Plus : spécialisée en sciences de l'information et en bibliothéconomie, cette base a été établie par la Library Association et par l'ASLIB (deux associations professionnelles anglaises) à partir de tous types de documents.

  Remarque par rapport au sujet : cette base permet d'élargir la recherche aux publications étrangères. De plus, comme cette base est spécialisée en sciences de l'information, elle recense la majorité des publications représentatives de ce thème : elle est donc une source d'information importante pour nous.

1.2.1.2 Méthodes utilisées :

La première recherche sur CD-ROM s'est effectuée sur Docthèses ce qui a permis de déterminer si certaines équipes de recherche travaillaient actuellement en France à l'élaboration de logiciels de « news » à des fins de veille. Cette interrogation a également permis de définir plus précisément les mots-clés pour une orientation de la recherche vers les méthodes.

Après interrogation sur Internet, j'ai travaillé sur le CD-ROM LISA Plus et celui de la BN-Opale. Ce dernier est très général, j'ai donc récolté peu d'information pertinente. Par contre, l'interrogation de LISA Plus s'est avérée intéressante, cette base étant spécialisée en sciences de l'information. De plus, l'interrogation de LISA sur CD-ROM a été un choix plutôt que l'utilisation du serveur DIALOG pour cette base.

Enfin, j'ai tenté à plusieurs reprises d'interroger le CD-ROM PASCAL, normalement accessible en réseau. Ma recherche s'est avérée infructueuse car il m'a été impossible d'accéder au CD-ROM.

1.2.1.3 Mots-clés retenus :

Au début de ma recherche, pour l'interrogation du CD-ROM Docthèses, je me suis servie des pistes données par mon commanditaire et de mots-clés très généraux.

Tableau 1 : Pistes préalables

Mots-clés	Auteurs	Localisation
- Veille stratégique technologique - Information scientifique et technique - Logiciel	- CONSTANT Patrick et JOUVE Olivier et VOGEL Claude	PARIS : Université Léonard de Vinci :
	- TURNER et BORZIC et SAINT-LEGER	MEUDON : CERESI, CNRS Dynatools (produit)
	- DOU	MARSEILLE : CRRM

Partant de là, le système de requêtes a été le suivant :

Tableau 2 : Système de requêtes pour Docthèses

N° de la requête	Formulation de la requête	Nb. de réponses
Question 1 :	AUTEUR : DKAKI	2 références
Question 2 :	MOT-CLE : veille and (stratégique or technologique)	14 références

Pour la question 1 : suite à une première recherche sur Internet et à la visite d'un site correspondant au curriculum vitae de Taoufiq DKAKI (http://atlas.irit.fr/cvtao.html), cette

requête m'a permis d'avoir des précisions sur sa thèse grâce aux résumés conséquents présents dans Docthèses.

Ce début de recherche a été volontairement très général afin de connaître tout ce qui se faisait en logiciel de veille (cf. Question 2). Cette requête m'a également permis de cerner les mots-clés utilisés pour indexer les différentes thèses sur le sujet.

Par la suite, après avoir exploré plus avant les ressources qu'offraient Internet, l'interrogation d'autres CD-ROM semblait souhaitable afin d'estimer la valeur de nouveaux mots-clés et surtout de nouveaux auteurs :

↳ Interrogation du CD-ROM de la BN-Opale :

Tableau 3 : Système de requêtes pour la BN-Opale

N° de la requête	Formulation de la requête	Nb. de réponses
Question 1 : **Question 2 :** **Question 3 :** **Question 4 :** **Question 5 :**	mc[1] = veille mc = technologique cs[2] = 1 and cs =2 mc = Etats-Unis cs = 3 and cs = 4	1 réponse
Question 6 : **Question 7 :** **Question 8 :**	mc = indexation mc = automatique cs = 6 and cs = 7	3 réponses
Question 9 : **Question 10 :** **Question 11 :**	mc = classification mc = automatique cs = 9 and cs = 10	11 réponses

Le CD-ROM de la BN-Opale est assez fastidieux dans son mode d'interrogation, comparé à Docthèses. De plus cet outil est trop général, ainsi peu de documents se sont avérés réellement pertinents (cf. Partie 1.3.2 Pertinence des documents).

[1] mc = recherche sur un mot-clé.
[2] cs = cumuler plusieurs questions.

↳ Interrogation du CD-ROM Lisa Plus :

Tableau 4 : Système de requêtes dans Lisa Plus

N° de la requête	Formulation de la requête	Nb. de réponses
Question 1 :	au[3] = BALLARD	
Question 2 :	kw[4] = indexing	
Question 3 :	cs = 1 and cs =2	3 réponses
Question 4 :	au = FRAKES	2 réponses
Question 5 :	ft[5] = SCISOR	2 réponses
Question 6 :	ft = software	
Question 7 :	ft = news	
Question 8 :	ft = intelligence	
Question 9 :	cs = 6 and cs = 7 and cs = 8	2 réponses
Question 10 :	ft = retrieval	
Question 11 :	ft = information	
Question 12 :	cs = 10 and cs = 11	
Question 13 :	ft = hypertext	
Question 14 :	cs = 12 and cs = 13	
Question 15 :	ft = classification	
Question 16 :	cs = 14 and cs =15	1 réponse
Question 17 :	kw = automatic	
Question 18	kw = indexing	
Question 19 :	cs = 17 and cs = 18	
Question 20 :	kw = news	
Question 21 :	cs = 19 and cs = 20	3 réponses
Question 22 :	cs = 15 and cs = 17 and cs = 20	1 réponse

Le CD-ROM Lisa Plus adopte le même principe hermétique de recherche que celui de la BN-Opale ; cependant, le contenu de son fonds correspond bien à nos attentes. Les thèmes abordés sont les sciences de l'information, l'informatique mais aussi les problèmes linguistiques d'indexation automatique et de classification automatique.

1.2.2 Stratégie de recherche sur serveur DIALOG

1.2.2.1 Identification des banques de données :

Afin d'identifier les bases de données à interroger, j'aurais pu utiliser DIALINDEX qui répertorie les banques de données présentes sur DIALOG. Cependant, j'ai préféré me référer aux sources papier plus facile d'accès.

[3] au = recherche sur un auteur.
[4] kw = recherche sur un mot-clé en anglais.
[5] ft = recherche en texte intégral.

- ADBS, ANRT.- *Répertoire des banques de données professionnelles : Banques et services d'information en ligne*.- Paris : ADBS, 14e ed., 1993.
- KNIGHT-RIDER INFORMATION.- *Database Catalogue*.- [S. l.] : Knight Rider Information, 1996.- 110 p.

Ces deux ouvrages m'ont permis de déterminer les banques de données à interroger grâce à leur index par sujet. Les sujets sélectionnés furent dans le thème Sciences, Technologies et Ingénierie :

- les sciences en générale et la technologie,
- les sciences humaines et sociales.

1.2.2.2 Banques de données interrogées :

Mes recherches se sont effectuées sur sept banques de données :

- INSPEC (base n°2) : depuis 1979, cette banque de données recense l'information sur la littérature générale concernant la physique, l'électronique, l'informatique et les technologies de l'information.

  Remarque par rapport au sujet : son fonds cumule les données sur les publications du domaine informatique et celui des sciences de l'information ; INSPEC est, par ces aspects, doublement intéressante pour notre sujet.

- NTIS : National Technical Information Service (base n°6) : depuis 1964, les résultats de la recherche, du développement et de l'ingénierie américaine financée par le gouvernement y sont recensés.

  Remarque par rapport au sujet : cette banque de données est très pertinente ne serait ce que pour connaître l'avancée des recherches américaines.

- Information Science Abstracts (base n°202) : depuis 1966, sa couverture bibliographique concerne les sciences de l'information.

  Remarque par rapport au sujet : cette base est performante pour notre sujet car elle concerne également les thèmes connexes aux sciences de l'information.

- Microcomputer Abstract (base n°233) : depuis 1981, cette banque produit des résumés sur des extraits de publications traditionnelles concernant le marché du micro-ordinateur en industrie, éducation, librairies et centres de documentation.

  Remarque par rapport au sujet : cette banque de données est un bon outil pour le sujet qui nous préoccupe. Elle permet de déceler si un outil tel qu'un logiciel de veille portant sur les « news » est déjà commercialisé et par quelle société.

- Softbase : Reviews, Companies and Products (base n°256) : cette banque de données est un répertoire des producteurs de logiciels de mini, micro-ordinateurs et de grands systèmes.

  Remarque par rapport au sujet : sa couverture documentaire est récente et très dense, cette banque de données est une source d'information intéressante.

- *IAC* Computer Database (base n°275) : depuis, 1983, cette banque recense les informations sur les producteurs et les sociétés de logiciels et de matériels informatiques, de télécommunications et d'électronique.

  Remarque par rapport au sujet : la seule remarque concerne son fonds très orienté état du marché et non-état des recherches.

- Computer News Fulltext (base n°674) : cette banque de données recense le contenu de deux journaux de communication IDG traitant de technologies relatives au domaine informatique et celui des réseaux. Ces deux journaux sont :

  - Computerworld : depuis 1989,
  - Network World : également depuis 1989.

  Remarque par rapport au sujet : son fonds permet d'identifier la situation des développements informatiques et des sociétés concernées.

1.2.2.3 Méthodes utilisées :

Ces banques de données ont été interrogées les unes après les autres puisque beaucoup de requêtes ont été formulées avec des noms d'auteur. Or, l'intitulé des noms d'auteur est fortement éclectique selon les bases : certaines bases mettent le NOM puis le Prénom séparés par une virgule, d'autres ne mettent que la première lettre du prénom (séparée ou non d'une virgule).

Pour faciliter l'interrogation par les auteurs, la commande « Expand » m'a été d'un grand recours. Elle m'a permis de visualiser, pour chaque base, la façon dont un auteur est cité afin d'y adapter la syntaxe de ma requête.

1.2.2.4 Mots-clés retenus :

Les mots-clés présentés ci-dessous ont été structurés sous la forme d'un tableau afin de pouvoir comparer les réponses entre les banques de données interrogées. Ce tableau permet également d'identifier les groupes de mots utilisés pour ces requêtes.

D'autre part, le tableau suivant est scindé en deux parties :

- une pour les mots-clés,
- l'autre concernant les auteurs et leur affiliation.

Tableau 5 : Système de requêtes des banques de données

	B. 2	B. 6	B. 202	B. 233	B.256	B. 275	B. 674
Data Analysis Extraction	22 (+ competitive intelligence = 1)	2147 (+ newsgroup = 0)	2	12 (+ newsgroup = 1)			12
Data Analysis Classification		456 avec le terme automatic	2	13	1		0
Data Modeling Computer Software	4 avec le terme cartography	9		57	0	10	
Economic Intelligence Computer	1 avec le terme newsgroup	25 avec le terme software	100 sans le terme computer	14 avec economic intelligence	7	0 sans le terme computer	14 sans le terme computer
DynaTools	0	0	0	0	0	0	0
Method Analysis	22 avec economic intelligence		73 (+ newsgroup = 0)	2	5		2
Data Analysis	7 avec semiotic						1 avec newsgroup
CS = CERESI ou CNRS AU = "Turner WA."	Uniquement l'auteur = 43 Le tout = 7	1	2	0	0	0	0
AU = Dou	0	0	11	0		0	0
Au = "Vogel C."	25		42	0	0	0	0
AU = "Constant P."	7	0	1	0	0	0	0
Au = Jouve		0	2	0	0	0	0
CS = "CERESI"	4		2	0	0	0	0

1.2.3 Recherche sur Internet

1.2.3.1 Identification des moteurs de recherche :

Face à la multitude des moteurs de recherche disponibles sur Internet, j'ai préféré identifier ceux étant les plus pertinents pour le sujet et mon système de requêtes. Pour cela, je me suis référée à l'ouvrage :

☞ LARDY, Jean-Pierre.- *Recherche d'information dans l'Internet : Outils et Méthodes*.- Paris : ADBS, 2e ed., 1996.-100p.

Cet ouvrage m'a permis de définir les moteurs de recherche les plus adéquats afin de trouver au mieux une information de nature très diverse (journaux, informations institutionnelles, catalogues de librairies et de bibliothèques, banques de données, contributions à des forums, etc.).

1.2.3.2 Moteurs de recherche utilisés :

Deux types d'outils ont été utilisés, lors de la recherche sur Internet :

- Les répertoires raisonnés dont l'indexation est humaine, ont l'avantage de classer les ressources "manuellement" mais ceci au détriment de l'exhaustivité.

Yahoo : (Yet Another Hierarchically Organized Oracle). Outil le plus ancien, début 1994, il visite chaque jour environ 60 sites.
Remarque : ce moteur manque d'exhaustivité.

- Les moteurs de recherche font intervenir des logiciels-robots pour l'indexation ; ils garantissent donc une meilleure exhaustivité et mise à jour mais au détriment d'un classement raisonné.

Lycos : apparu en juin 1995, ce moteur explore l'Internet chaque jour et effectue une indexation du début des pages et non du texte intégral.
Remarque : ce moteur effectue une troncature à droite implicite, ce qui amène énormément de bruit lors de l'interrogation sur notre sujet.

Infoseek : ce moteur recherche l'information dans les serveurs W3, Gopher, FTP et les pages de « news ». Sa mise à jour s'effectue toutes les une à deux semaines et l'indexation se fait sur le texte intégral.
Remarque : comme Lycos, ce moteur effectue une troncature à droite implicite. De plus, l'interrogation est réalisée sans opérateur booléen. Ce sont donc deux inconvénients majeurs pour notre recherche.

AltaVista : datant de décembre 1995, AltaVista est le service le plus récent. Cet outil indexe en texte intégral et permet deux types de recherche : simple ou avancée.
Remarque : outil le plus complet pour notre sujet avec la possibilité d'une recherche avancée permettant de combiner des termes à l'aide des opérateurs AND, OR, NOT et NEAR et d'interroger sur les mots composés grâce aux guillemets.

J'ai utilisé Yahoo lors d'une première interrogation pour une recherche globale. Par la suite, AltaVista m'a permis de spécifier ma recherche.

1.2.3.3 Mots-clés retenus :

Le système de requêtes présenté ci-dessous est élaboré chronologiquement, c'est-à-dire qu'il suit l'évolution de la recherche : recherche générale au départ pour devenir de plus en

plus spécifique. On part donc des pistes données par mon commanditaire pour arriver à des termes d'indexation et de classification automatique.

D'autre part, ce tableau recense volontairement les requêtes posées avec la syntaxe du moteur de recherche AltaVista car cet outil a été le plus pertinent pour aborder l'Internet. D'autres moteurs ont été utilisés mais ils me semblaient moins efficaces (cf. les remarques ci-dessus).

Enfin, au début de ma recherche, des requêtes ont également été posées français, elles ne se trouvent pas ici car les références trouvées n'ont pas été retenues dans la bibliographie : cette recherche m'a permis d'acquérir une première approche des différents moteurs de recherches.

Tableau 6 : Système de requêtes dans Internet

N° Question	Intitulé de la requête	Nb. Réponses
1	pôle near universitaire near Leonard near Vinci	70
2	free near text and (competitive or economic) near intelligence	67
3	*Idem* + and software and research	23
4	((free near text) or (online near information) and (economic or competitive) near intelligence) and automatic near indexing	1
5	(automatic near indexing) and (free near text) or (online near information)	80
6	"competitive intelligence" and software and (newsgroup or (information near online))	11
7	Callon and (economic or competitive) near intelligence	1
8	"boite noire" and Callon	2
9	automatic near indexing and (news or newsgroup) and USA	150
10	research and tool and USA and ((economic or competitive) near intelligence) and (free near text)	23
11	(automatic near indexing) and (free near text) and tool	50
12	(syntactic near scanning) and linguistic	2
13	(word near frequency) and (linguistic near research)	38
14	(cluster near analysis and (news or newsgroup) and software and (economic or competitive) near intelligence)	150
15	*Idem* + and software *Avec en plus un intervalle pour les dates pour obtenir uniquement l'information la plus récente :* Start date : 01/Jun/96 End date : 15/Feb/97	80

1.3 Evaluation de ces outils

1.3.1 Evaluation du temps passé et du coût de l'information

CD-ROM :

L'accès à l'information sur CD-ROM étant gratuit, j'ai uniquement comptabilisé le temps passé à la recherche. Au total, j'ai passé 2 h ½ à la consultation de ce type de support, cette durée se décompose comme suit :

- pour le CD-ROM Docthèses : ½ heure de recherche ;
- pour le CD-ROM de la BN-Opale : 1 heure de recherche ;
- pour le CD-ROM Lisa : 1 heure de recherche.

DIALOG :

Etant donné que les tarifs des banques de données sont en dollars (prix de connexion par heure et prix d'une référence), pour obtenir l'équivalent en franc, j'ai effectué la conversion à un taux de 5,7 (taux indiqué dans le quotidien *Le Monde* du 28 février 1997).

Tableau 7 : Tarif et durée d'interrogation

Bases de données	Temps		Références		Total par base
	Durée	**Tarif**	**Nb.**	**Tarif**	
COMPUTER DATABASE	11 mn	62,60 f	5	84,05 f	146,65 f
COMPUTER NEWS FULLTEXT	28 mn	159,35 f	15	149,60 f	308,95 f
INSPEC	1 h 36 mn	815,10 f	111	917,40 f	1732,50 f
MICROCOMPUTER ABSTRACTS	14 mn	79,35 f	30	213,70 f	293,05 f
NTIS	32 mn	176,45 f	42	335,10 f	511,6 f
SOFTBASE	20 mn	170,80 f	15	149,60 f	320,4 f
INFORMATION SCIENCE ABSTRACTS	34 mn	290,35 f	21	113,70 f	404,05 f
			Total Final		3717,20 f

INTERNET :

L'information sur Internet est gratuite, tout comme pour les CD-ROM, mais le temps passé à l'exploration via les moteurs de recherche et les liens hypertextuels est considérable. En effet, 22 h 40 de travail m'ont été demandées afin de déceler des informations pertinentes par rapport à notre sujet. Ceci s'explique par le manque d'homogénéité des termes d'indexation et surtout par l'utilisation de programmes d'indexation automatique sur du texte intégral. Ainsi, certains termes utilisés comme "veille" ou "news" engendrent énormément de bruit lors d'une interrogation.

D'autre part, la facilité de navigation grâce à l'hypertexte pose un problème majeur : celui de vouloir tout consulter et par conséquent de s'orienter par ces liens vers de l'information s'éloignant de la recherche de départ.

1.3.2 Pertinence des documents

Le tableau suivant se propose de comparer les différents outils utilisés par rapport notamment au tarif, au temps de recherche nécessité, aux nombres de références trouvées et aux nombres de références finalement gardées afin de donner une idée sur le mode d'accès le plus facile et le plus pertinent à l'information.

Tableau 8 : Pertinence des outils de recherche

Critères	CD-ROM	DIALOG	INTERNET
Accès à l'information	Système de requêtes sur : auteur, titre, mots-clés et texte intégral.	Système de requêtes sur : titre, auteur, affiliation, date, mots-clés et texte intégral.	Requête généralement sur du texte intégral et navigation hypertextuelle
Tps de consultation	Total = 2 h 30	Total = 3 h 59	Total = 22 h 40
Nb. de documents consultés	Total = 45	Total = 239	Sites obtenus = 847 Sites consultés = 240
Nb. de documents pertinents	Total = 23	Total = 67	Total = 29•
Rapport de pertinence	**51 %**	**28 %**	Calcul sur les sites consultés = **12 %**

Il est à observer que les thèses sont très bien placées dans ma bibliographie. Il m'a semblé pertinent d'insister sur ce point car beaucoup d'auteurs faisaient la remarque que les thèses rédigées au préalable étaient pour eux une grande source d'inspiration pour leur recherche. « *Grâce à ces thèses nous disposons déjà de nombreux résultats permettant d'envisager sérieusement l'étude des dimensions sociales et organisationnelles de l'intelligence.* » William A. TURNER [73].

• Ceci concerne les sites présentés dans la bibliographie finale et également des articles, des ouvrages et des congrès trouvés dans des bibliothèques virtuelles et des bibliographies sur le Web.

Il est à souligner que dans ce tableau les doublons n'ont pas été retirés c'est-à-dire qu'un document présent dans la bibliographie, située à la fin de ce travail, peut correspondre à la fois à une interrogation issue de Dialog et de CD-ROM ou encore d'Internet.

De plus, le nombre de documents consultés est tel que je les ai extraits des CD-ROM, des banques de données et d'Internet avant lecture attentive. Par contre le nombre de documents pertinents correspond aux documents conservés et présents dans la bibliographie située à la fin de ce rapport.

2. *Analyse du corpus*

Cette partie présente une étude scientométrique du corpus constitué dans la bibliographie. Cette étude a été élaborée pour une première approche de cette bibliographie afin de guider mon commanditaire et les éventuels lecteurs de ce rapport. En effet, la connaissance de quelques données statistiques permet de caractériser les données trouvées c'est-à-dire de cerner le type d'information et ses spécificités. Le tableau synoptique suivant est chargé de donner une vue globale de la démarche adoptée pour cette analyse ainsi que d'expliquer le rôle d'une telle étude :

Statistiques sur :	*Eléments de conclusion concernant :*
- Type de document	- Support de communication
- Date de publication	- Fraîcheur de l'information
- Auteurs principaux	- Acteurs, communauté scientifique

2.1 Type de document

Les références de la bibliographie, comme le montre le tableau 1, représentent plus de 29 % des congrès et colloques. En effet, actuellement les recherches sur ce domaine sont importantes, elles font l'objet d'un enjeu économique non négligeable. Les ressources Internet sont également bien placées avec plus de 28 %. Il faut dire que la recherche sur ce type de support a été privilégié pour l'obtention d'informations récentes : la plupart de sites trouvés dates de fin 1996. Ceci prouve que la recherche dans ce domaine est en émergence et qu'il est nécessaire de continuer cette étude pour voir son évolution.

Tableau 9 : Type de document

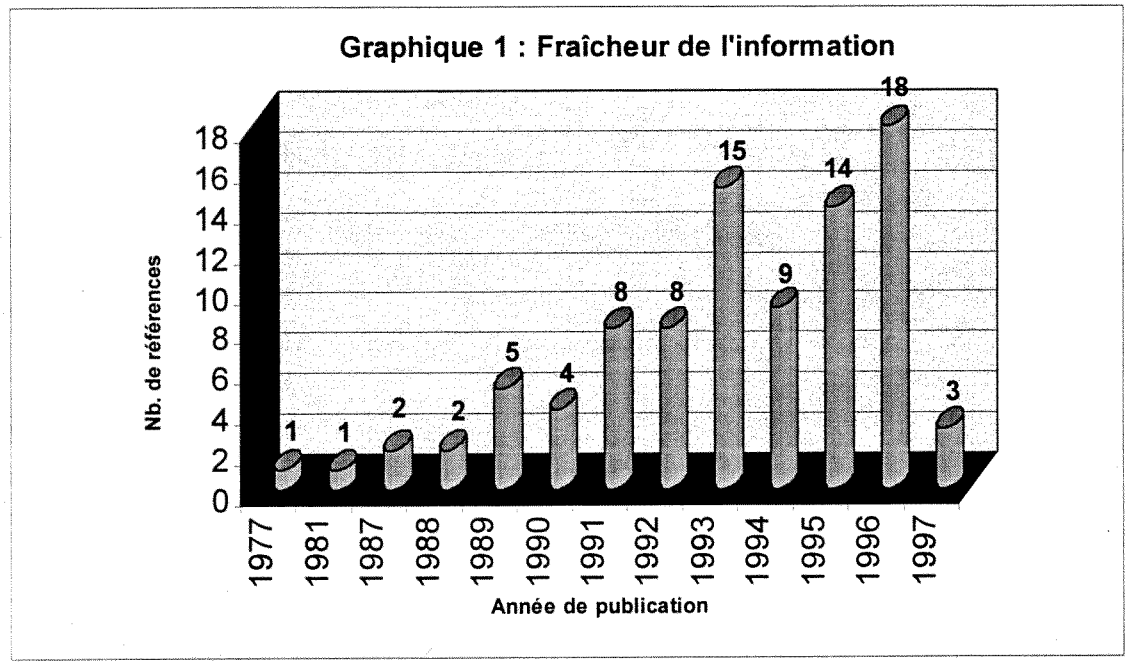
Type du document	Nombre de références	%
1) Congrès	27	29,67 %
2) Ressources Internet	26	28,57 %
3) Articles	18	19,78 %
4) Thèses	11	12,09 %
5) Ouvrages	6	6,59 %
6) Rapports	2	2,20 %
7) Symposium	1	1,10 %
Total	91	100,00 %

On sait que dans la tradition de l'information documentaire, il existe une distinction entre la littérature scientifique à proprement parler, une littérature publique au sens où elle est publiée par des organes scientifiques de communication, et la littérature grise. Dans notre cas, cette dernière est très bien représentée. Elle est constituée par les rapports, les thèses et les actes de congrès ce qui représente plus de 44 % de nos références. L'exploitation de ce type d'information est, de par sa nature plus complexe et cependant plus intéressante au niveau

informationnel. L'analyse des thèses est un moyen de connaître le monde de la recherche universitaire et des grandes écoles (auteurs et sujets de thèse, discipline concernée, directeurs de thèse, membres des commissions d'examens, universités).

2.2 Fraîcheur de l'information

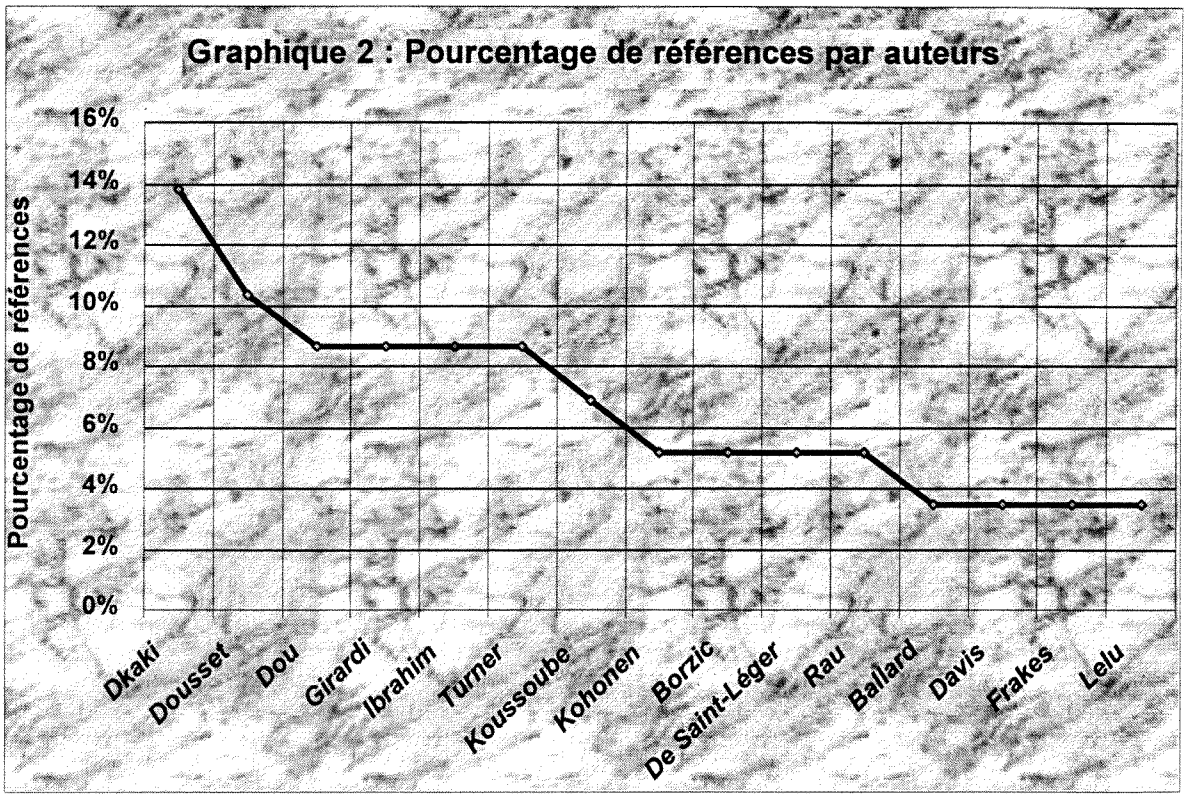
On mesure l'actualité de l'information par la distribution des références selon la date (année) de publication (voir ci-dessous Graphique 1).



A propos de ce graphique, on s'aperçoit qu'en général, l'information y est récente : ces cinq dernières années couvrent environ 65 % de l'information obtenue c'est-à-dire 59 références sur 90 au total puisqu'une référence n'a pas de date de publication. Ainsi, les récentes recherches qui ont trait à notre sujet sont présentes dans la bibliographie. Je tiens à souligner que certains documents peuvent paraître un peu "vieux" (1977, 1981) ; ce sont en fait des ouvrages de références étant souvent la base des recherches actuelles.

2.3 Auteurs les plus souvent cités

Le graphique suivant (graphique 2) est représentatif des principaux auteurs de ma bibliographie qui sont essentiellement des producteurs de littérature grise.



Dans cette figure, on retrouve la communauté française avec, notamment DOUSSET, DOU, TURNER, BORZIC et DE SAINT-LEGER ainsi qu'une partie de la communauté étrangère avec GIRARDI, IBRAHIM, KOHONEN, RAU, BALLARD et FRAKES. Cette dernière est essentiellement représentée par les américains.

3. *Synthèse*

Avant Propos

Pour la rédaction de cette synthèse, j'ai choisi volontairement des textes s'appuyant sur des méthodes de classification et d'indexation automatique puisqu'il n'existe pas réellement de logiciels de veille portant sur les « news » ou les forums électroniques.

Cette partie se propose donc d'être une présentation des méthodes d'analyse de données susceptibles d'être pertinentes pour la conception et le développement de tels outils. Leurs avantages et inconvénients y sont également abordés. Ainsi, mon commanditaire pourra avoir une première approche des méthodes d'analyse et consulter, s'il le désire, plus avant les documents correspondant à ses intérêts.

D'autre part, les documents cités dans cette partie ont été choisis parce qu'ils font tous référence à de l'information non structurée tels que les « news » ou les forums électroniques. Par conséquent, ils correspondent au type d'information qui nous intéressent pour notre sujet. Les textes choisis pour élaborer cette synthèse sont :

- Réf. [72] : CADIS, Inc.- *Welcome to CADIS : the innovators in internet and intranet information classification, publishing, storage and retrieval*, [En ligne], (Page créée en 1996).- Adresse URL : http://www.cadis.com
- Réf. [75] : DE SAINT LEGER, Mathilde ; CERESI/CNRS Centre de recherche en Sciences Infométriques.- *DynaTools : un outil de gestion dynamique des flux d'informations pour une veille scientifique*, [En ligne], (Page créée en 1996).
  Adresse URL : http://atlas.irit.fr/vsst/toulouseM2.html
- Réf. [77] : GIRARDI, M.R. ; IBRAHIM, B.- *Automatic Indexing of software artifacts*, [En ligne], (Page créée le 7 Juin 1996).
  Adresse URL : http://cuiwww.unige.ch/eao/www/ROSA.papers/SR94/paper.html
- Réf. [78] : HUMPHREY, Pete.- *Natural Language Processing at EDS*, [En ligne], (Page créée le 9 Avril 1992).
  Adresse URL : http://www.edsr.eds.com/edsr/papers/natlang.html
- Réf. [79] : KOHONEN, T.- *WEBSON – Self-organizing map for internet exploration*, [En ligne], (Page créée le 23 Février 1997).- Adresse URL : http://nucleus.hut.fi/nnrc/new_book.html
- Réf. [81] : LELU, Alain.- *De l'émergence des concepts : réflexions à partir du traitement « neuronal » des bases de données documentaires*, [En ligne], (Page créée le 10 Septembre 1996).
  Adresse URL : http://www.info.unicaen.fr/bnum/jelec/Solaris/d02/2lelu.html
- Réf. [88] : TURNER, William A.- *Penser l'entrelacement de l'Humain et du Technique : les réseaux hybrides d'intelligence*, [En ligne], (Page créée en 1994).
  Adresse URL : http://www.info.unicaen.fr/bnum/jelec/Solaris/d01/1turner.html
- Réf. [89] : VOGEL, Claude.- *SEMIO Corporation – SemioMap Search*, [En ligne], (Page créée en 1996).- Adresse URL : http://www.semio.com

Introduction

Face à l'expansion des autoroutes d'information et la multiplication des sources, la tendance actuelle est de développer des programmes permettant de gérer ces énormes flux. Ce type d'information est riche de conséquences pour celui qui sait la maîtriser et en tirer les conclusions qui s'imposent. Il ne s'agit plus uniquement de récolter de l'information mais il faut la structurer et voir son évolution dans le temps afin de refléter au maximum la réalité de la recherche et du développement dans tout domaine.

C'est dans cette optique que de nombreux logiciels de veille ont vu le jour : le suivi permanent des flux d'information permet de déceler les prémisses d'un changement et de repérer les thèmes naissant. Par conséquent, ces logiciels permettent aux décideurs de dégager les menaces et les opportunités pour leur entreprise en un temps t(x).

Le texte suivant est composé de deux grands axes. Le premier est une présentation de nouveaux types d'outils pour le traitement de données textuelles non structurées. La seconde partie développe les méthodes utilisées pour l'indexation et la classification automatique.

3.1 De nouveaux outils pour le traitement de l'information

L'explosion de l'information virtuelle a créé un vaste embouteillage de la productivité, en particulier sur les réseaux. Un gros corpus d'information en ligne n'est pas exploitable sans un outil capable de le structurer et de l'organiser sous une forme élaborée et de manière interactive. C'est pourquoi, nous allons déjà voir l'émergence de ces outils pour ensuite en étudier quelques-uns.

3.1.1 Les outils préalables

Face à la prolifération de l'information, les utilisateurs d'Internet se noient dans l'information. Les outils actuels de recherches retournent des listes interminables de documents. Généralement, ils réalisent une recherche exhaustive qui engendre énormément de bruit et une perte de temps considérable. Ces méthodes de recherche qui ne sont pas adaptées pour explorer les idées et les événements. Ainsi, pour déceler l'information relative à sa requête, l'utilisateur se trouve face à un casse-tête.

A force de reformuler leurs questions et de naviguer grâce aux liens hypertextuels, les utilisateurs arrivent à trouver une information quelque peu en rapport avec le thème recherché. Claude VOGEL soulève ce problème d'accès à l'information [89] : *« To find answers to open-ended queries, users switch from one tool to the next, until they happen to sumble upon something relevant. »* En effet, aucun détail n'est omis avec ces systèmes. Ils rapportent, en plus, environ une centaine de documents qui n'ont qu'une relative cohérence avec le sujet. Ces documents ont néanmoins besoin d'être lus et analysés. Aussi, à l'heure actuelle, la majorité des outils proposés sont des applications qui effectuent des recherches exhaustives sur du texte intégral à l'aide des opérateurs booléens. Ceci explique leur mauvais

résultat face à un désir de précision sur une requête : *« Often the search results show high recal and low precision, or vice versa. »* [79]. Depuis peu de temps, afin de palier à ce problème majeur, de nouveaux outils apparaissent pour l'exploration "intelligente" d'un gros corpus de textes non structurés.

Actuellement, de nouvelles applications commencent timidement à surgir sur le marché. Leurs méthodes sont du reste révélées, pour la majorité, sur l'Internet depuis peu (novembre 1996). Ces outils travaillent sur de l'information électronique non structurée où il n'y a pas de champ pour les mots-clés et ou les expressions tiennent du vocabulaire courant. Ils permettent d'explorer l'information sans un balayage ou une lecture d'une grande quantité de documents.

3.1.2 De nouveaux besoins

Lors d'une recherche, l'utilisateur s'intéresse à une question spécifique et désire trouver au plus vite l'information correspondant à sa requête pour valider ses problèmes et non pas toutes les informations en rapport avec le sujet. Aussi, Claude VOGEL [89] montre la nécessité de nouveaux outils pour l'exploration de l'information, *« Users need a tool that uncovers relationships between words and ideas to end their aimless searching »*, ainsi que le groupe CADIS : *« Products that enable content to be published in an easily searchable way that overcomes the frustration of existing keyword search systems. »*

Aujourd'hui, il s'agit de mettre en valeur les relations entre les documents d'Internet (en information externe) et aussi d'Intranet (en information interne). Le besoin d'outils "intelligents" se fait ressentir, ils permettent de couvrir et d'identifier les relations entre les mots et les idées. *« Uncovering and discavering valuable relationships beetween ideas and words contained in text-based documents located on corporate networks, online databases and, of course, the Internet. »* [89]. Il est important de développer des méthodes qui permettent d'explorer minutieusement des collections de documents. Pour cela, les outils de recherche intègrent des index créés par des personnes physiques qui assignent chaque document à une catégorie.

Mais de nombreuses autres méthodes intégrant un traitement automatique existent. Ainsi, je vous propose de découvrir ces outils à la pointe de la technologie dont l'apport pour la stratégie d'un centre de recherche ou d'une entreprise est loin d'être négligeable.

3.1.3 Présentation de quelques outils de traitement de l'information appliqué à du texte libre

Diverses entreprises et laboratoires de recherche ont développé des outils répondant aux besoins présentés précédemment. Ces outils seront décrit un à un par organisme de développement, dans un souci de clarté.

Le CERESI (Centre d'Etudes et de Recherche En Sciences Infométriques) le CNRS (Centre National des Recherches Scientifiques) ont développé l'outil DynaTools [75] :

- DynaTools est un progiciel de recherche documentaire qui prend en compte la science disponible sur l'Internet et ce en temps réel. Pour cela, il effectue un

sondage périodique des flux sur les réseaux et cible les sources d'information représentatives du domaine à étudier. Une fois l'information sélectionnée, cet outil, composé de divers modules, automatise toutes les procédures de modélisation, de pondération et de filtrage des "mots remarquables". « *Un descripteur est dit remarquable si son poids au cours des périodes étudiées, dépasse un seuil défini statistiquement. Seuil obtenu à partir de la valeur moyenne du poids des descripteurs et de leur écart-type.* » [75]. En effet, l'information pour devenir un renseignement doit certes être pertinente mais elle doit être aussi synonyme de changement ou au contraire de stabilité. Il s'agit donc de pondérer l'information en fonction de son importance dans un domaine et également en fonction de la période étudiée.

En second lieu, DynaTools permet d'obtenir des cartes thématiques chronologiques du domaine à partir des flux d'information. Ici, deux méthodes sont utilisées en complémentarité :

- Vue Macroscopique : la cartographie tient compte à la fois des différentes associations qui existent entre descripteurs et des variations de ces associations. Le calcul des associations ne consiste pas en un simple calcul de fréquence mais utilise la méthode des mots associés [61] permettant de définir la cooccurrence entre deux descripteurs. Ici, le modèle de base a été le programme Leximappe développé au Centre de Sociologie et de l'Innovation de l'Ecole des Mines de Paris en collaboration avec le CNRS.
- Vue Microscopique ou individuelle de chaque information du flux : tous les thèmes sont positionnés sur une carte thématique, leur situation sur la carte détermine sils sont des thèmes importants ou des thèmes isolés

Pour avoir des compléments d'information sur le calcul du poids d'un mot-clé, une lecture plus approfondie du document [75] (Partie 4) permettra d'obtenir des précisions conséquentes. Enfin, il faut souligner que DynaTools travaille bien sur de l'information en ligne mais cette information comporte déjà un traitement préalable : ces documents possèdent un champ descripteur ce qui n'est pas le cas des «news » et des forums électroniques.

L'entreprise SEMIO Corporation a développé une nouvelle méthode de recherche appelée la "recherche de la découverte" (Discovery Search) ainsi que deux outils complémentaires [89] :

- La méthode "Dicovery Search" répond à des questions ouvertes ou fermées en décelant les relations entre termes. « *A discovery search tool surfaces content hidden deep inside documents and displays it in an organized, navigable way* » [89]. Comme dans DynaTools, le contenu des documents est organisé d'une manière structurée et navigable, l'utilisateur peut découvrir un thème, explorer l'information et déceler les changements à deux périodes données.
- SemioMap : outil qui permet d'analyser et de catégoriser des documents en montrant les connexions logiques entre les mots et les idées sous forme de cartes graphiques qui sont le départ de la navigation hypertextuelle. Il travaille sur de l'information non structurée et a la possibilité d'analyser le récit. Tout comme DynaTools, il est basé sur la méthode des mots associés [61].
- SemioLex : il permet le traitement lexical et l'analyse sémiotique. Au départ, cet

outil était voué à améliorer la prise de décision dans un environnement de veille stratégique.

L'organisme CADIS a conçu deux outils [75] :

- CADIS-PMX : outil qui a été créé pour permettre une solution de management de l'information. Il adopte une organisation des données pour faciliter leur lecture et leur accès aux utilisateurs finaux.
- CADIS-Krakatoa : il permet de naviguer dans l'information à l'aide de classes et de sous-classes par navigation hypertextuelle. La manière de structurer les informations est donc interactive comme les autres solutions précédemment présentées. Le texte [75] décrit les technologies utilisées et l'environnement nécessaire (ex. : bibliothèque API, langage de programmation JAVA).

Le Centre de Recherche sur les Réseaux Neuronaux de l'université de technologie d'Helsinki a élaboré l'outil WebSom [79] :

- WebSom : comme les outils précédents, il permet d'organiser l'information textuelle non structurée sous forme de cartographie étant le point de navigation. *« The user may view any area of the map in detail by simply pointing to the map image with the mouse »* [79]. Ici, les documents sont situés les uns par rapport aux autres d'après la méthode des réseaux neuronaux. Cet outil est basé sur l'algorithme SOM ("Self-Organizing Map"), datant de 1981, qui permet de structurer les documents dans un espace à deux dimensions. Au préalable, cet algorithme a permis des applications pour l'analyse de l'image, les télécommunications et la catégorisation des données économiques. Aujourd'hui, il sert au traitement naturel du langage naturel. Pour un complément d'information sur ce point, consulter l'article [9]. WebSom adopte également à une méthode alternative ainsi qu'une méthode de traitement du texte intégral. *« If no keywords are available and the texts are very colloquial such as the free-form discussions in the Internet newsgroups are, new full-text searching methods have to be developed »* [79]. En dernier lieu, cet outil permet également de structurer l'information tels que les « mail ».

3.2 Méthodes d'indexation et de classification automatique

Il existe toutes sortes de texte, d'une information très structurée telle que la référence bibliographique à une information libre tel que le texte intégral. De même, les supports sont très variés : papier, CD-ROM, information en ligne par Minitel, serveur ou Internet. Ainsi, tant de types d'information engendrent autant de traitements différents et de programmes au niveau de complexité diverse. Les problèmes posés par les fichiers textuels sont nombreux. Ils concernent la taille du fichier, son format et la structure du document. Enfin, avec un fichier non structuré on se trouve face à l'ambiguïté et à la complexité du langage naturel.

Concernant les logiciels sur de l'information non structurée : ils doivent permettre la reconnaissance de la structure grammaticale du texte c'est-à-dire les phrases, les paragraphes, etc.

3.2.1 Traitement linguistique et extraction de concept

Les problèmes au niveau linguistique font l'objet, depuis longtemps, d'études et de théories. Cependant, beaucoup de limites restent succinctement résolues et dépendent encore une fois du type et de la structure de l'information à laquelle on fait face. En effet, lorsqu'il traite une information non structurée, l'Homme fait intervenir diverses connaissances implicites lui permettant d'identifier les structures grammaticales, de reconnaître sans effort la variation d'un mot (pluriel/singulier), de traiter les synonymes, d'évaluer la signification d'un terme selon le contexte, etc. « *Il fait intervenir le langage, le raisonnement, mais aussi des phénomènes largement inconscients comme notre perception du monde et notre propre identité.* » [81] Tout cela, il faut l'apprendre à l'ordinateur. Or, il existe tellement de structures différentes qu'il est difficile de palier à tous les problèmes. Certains linguistes ont cependant imaginé des solutions comme celle de calculer l'occurrence d'un terme qu'il soit au singulier ou au pluriel par une troncature à droite : création d'un algorithme de recherche partielle de mot.

Utilisant les dérivés des méthodes issues de l'intelligence artificielle (tableaux, arbres, hiérarchies ainsi que des listes) puisque *« notre intérêt pour les flux nous conduit à poser comme hypothèse de travail l'existence d'un monde ouvert par opposition aux mondes clos de l'intelligence artificielle »* [88], l'indexation automatique consiste à faire émerger des concepts. A cette fin, des recherches ont été établies en croisant l'analyse des données et l'analyse des réseaux neuronaux formels. Aussi, nous allons observer cette méthode d'émergence des concepts appliquée aux documents électroniques.

Les concepts, également appelés micro-mondes ([81] et [88]), permettent de réduire le flot des informations et de produire une sorte de synthèse de l'information. A ce niveau, on désire simplifier le processus de conceptualisation pour être implanté dans un programme d'ordinateur. Alain LELU [81] part donc du postulat qu'il existe des phénomènes observables que l'on peut résoudre à un cas binaire dans le cas de questions fermées : phénomènes observés ou non aboutissant à la création de vecteurs décrivant chaque observation (où chaque vecteur contient toujours le même nombre d'éléments).

De là, un tableau de données est élaboré dont l'ordre des lignes et des colonnes est indifférent. *« Puisque l'ordre des lignes et des colonnes est indifférent, profitons-en : une première approche consiste à ré-ordonner à la main les lignes et les colonnes (c'est la méthode graphique due à Jacques Bertin) »* (Figure 4 [81]). Ce tableau est à deux dimensions répertoriant les documents et les termes d'indexation et permettant le calcul de l'ensemble des coefficients d'association. Une analyse basée sur ce même principe, plus rapide et moins fastidieuse puisque automatique, est appelée méthode d'analyse factorielle. Pour appliquer ces méthodes aux documents, il faut détecter les groupes de lignes et de colonnes car le nombre de mots-clés n'est pas identique pour chaque document (cf. Figure 6 [81]). L'énorme avantage de cette méthode est qu'elle simplifie le regroupement de données issues de grand tableau, elle n'est donc pas limitée par la taille des données à traiter. Par contre, un inconvénient majeur est à lui reprocher, celui de réduire à deux ou trois dimensions seulement les données alors que normalement, il existe autant de dimensions que de colonnes. Ici, l'objectif est de rendre visible la densité du nuage de points. Pour redonner un simulacre de niveau de dimensions, une solution est proposée dans l'article [81] : *« Comment visualiser, rendre sensible, la densité d'un nuage de points ? Réponse : en "épaississant", en "empâtant" chaque point du nuage, c'est-à-dire en déposant successivement des "pâtés" élémentaires de même forme dans le voisinage de chaque point. »* L'avantage de ce système est qu'il permet

d'obtenir des indicateurs d'appartenance plus ou moins forte à un groupe de termes. L'inconvénient est que les axes n'ont pas de repères les unes par rapport aux autres.

Actuellement, des recherches sont en cours afin de gérer des flux d'information quotidiennement [88]. Ces recherches concernent :

1. Dictionnaire d'indexation car tous les termes ne possèdent pas la même valeur informationnelle, ceci dépend de leur profil de distribution. Il s'agit de combiner des techniques statistiques et linguistiques.
2. Liste de mots retenus après calcul : on sonde les flux d'information pour incorporer les documents comportant les mots en question dans la base documentaire.
3. Mise à jour périodique du contenu de ces mémoires locales car le télédéchargement de données est constant ou plutôt quotidien. Il faut donc faire évoluer les tableaux de documents et de mots-clés.

3.2.2 Méthodes pour le traitement du langage naturel

Différentes approches ont été élaborées pour le traitement du langage naturel à partir du système des langages de programmation ([77] et [78]). Ces logiciels incluent l'extraction de l'information et l'indexation automatique de documents. Le tableau ci-dessous propose de décrire chaque logiciel et de présenter leurs avantages et inconvénients.

Tableau 10 : Méthodes de traitement du langage naturel

Méthodes	Descriptif	Avantages	Inconvénients
EDS	Règles de grammaires et analyse grammaticale. Algorithme de traitement et de l'environnement grammatical.		
Approche basée sur la connaissance	Sens du texte est en fonction du sens des mots pris individuellement et de leur combinaison. La signification globale est guidée par un petit ensemble de relations sémantiques primitives.	Structure sémantique. Prise en compte de la syntaxe.	Manque de régularité de la structure d'une phrase → création d'un algorithme.
« One of the most impressive systems to come out of this work, also showed how having a strong goal-driven aproach could enable a system to glean information from text by skimming for key terms » [78]. De plus, cette approche inclut un système d'analyse grammaticale syntaxique. Il existe deux approches différentes : ATN et CFGs.			
ATN	Utilisée en Intelligence Artificielle. Ce système permet de définir la structure pour les expressions syntaxiques et les relations sémantiques.	ATN tient compte de la sémantique et de la syntaxe.	Cette grammaire est non intuitive et elle est difficile à maintenir. De plus, elle ne permet pas de traiter une grosse quantité de textes.

Méthodes (suite)	Descriptif (suite)	Avantages (suite)	Inconvénients (suite)
CFGs (Grammaire hors contexte)	Le but est de formaliser les régularités des expressions du langage naturel.		Difficulté de prendre en compte tous les cas de figure de la langue. Les règles du langage naturel sont beaucoup plus complexes que les règles définies pour un langage de programmation
GPSG	Il a été choisit comme base de l'EDS ; il est lui-même basé sur le CFGs.	Théorie simple mais capable de reconnaître toutes les structures de la plupart des langages naturels.	
EPSG	Il permet de développer des outils basés sur l'analyse grammaticale. Approche syntaxique : "syntax first". Nécessité de développer un algorithme pour l'analyse grammaticale [78].	Langage robuste Système général facilement adaptable.	
NLP	Basé sur l'EPSG, il permet de traiter les textes en langue anglaise.	Identification des substantifs dans les phrases pour une indexation automatique. Traitement des sources d'information en ligne.	
« The NLP aproach presents clear advantages from the point of view of costs over either manual indexing aproaches or knowledge-based systems where knowledge bases are contructed manually' [77].			

3.2.3 Méthodes de classification automatique

De nombreuses méthodes ont été élaborées et leur degré de complexité varie considérablement d'une méthode à l'autre. L'objectif du tableau ci-dessous est de présenter succinctement les méthodes développées dans les articles choisis pour la synthèse. Ensuite, différents types d'algorithmes sont cités car ils sont complémentaires à ce type de traitement.

Tableau 10 : Méthodes de classification

Méthodes	Principe général	Classes recouvrantes	Classement des lignes	Coefficient d'appartenance	Somme pour une ligne
Classification hiérarchique Ex. : Leximappe[6]	Contexte	Avec ou Sans			
Classification non hiérarchique Ex. : K-Means		Avec ou Sans			
Type résolution d'adéquation : Ex. : ACP[7] et Analyse des correspondances	Recherche de valeurs propres et de vecteurs propres.				
- Analyse factorielle dite orthogonale	Répartition équitable autour de 0 ; donc résumés équilibrés.			Positif et négatif	
- Analyse factorielle dite oblique	Répartition non équitable autour de 0 ; résumés dissymétriques			Positifs et négatifs	
Classification directe Ex. : Sériation par blocs ; Analyse relationnelle	Indicateurs de liens entre mots de chaque classe	Oui	Oui et des colonnes		
Extensions : Classification floue	Inverse du cas binaire ; contexte	Oui	Oui	Positif ou nul	1
Classification semi-floue	Contexte	Oui	Oui	Nul	1

Le texte [81] aborde bien les différents types d'algorithmes. L'auteur présente ainsi les algorithmes neuronaux, l'ACP (Analyse en Composantes Principales) et l'ACL (Analyse en Composantes Locales). *« Au lieu de définir la fonction objectif par la somme des carrés des*

[6] Leximappe est un outil à l'origine de nombreux programmes actuels ([69], [73] et [79]).
[7] Analyse en Composantes Principales

projections du nuage de point [comme dans l'ACP], *nous la définissons* [l'ACL] *par la somme des carrés des projections tronquées. La projection tronquée d'un point du nuage sur un axe passant par l'origine est une grandeur positive ou nulle qu'on obtient en retirant à la valeur de cette projection une quantité fixe, quantité qui constitue précisément notre paramètre de finesse d'analyse : plus ce paramètre est proche de 1 et plus le paysage d'inertie locale, partielle, sera accidenté. »*

Conclusion

De nombreuses méthodes pour le traitement du langage naturel ont été développées depuis plusieurs années. Aujourd'hui, il s'agit de les adapter afin de permettre le traitement de données non structurées comme le texte libre. Il faut également intégrer aux programmes une classification automatique des données et enfin, une méthode de représentation des données comme la cartographie.

Par conséquent, depuis peu de temps, divers outils facilitant l'exploitation de données textuelles provenant de l'Internet tels que les « news » ou les forums électroniques commencent à faire leur apparition. Parmi eux, il faut retenir SemioMap, SemoiLex et WebSom. Ils structurent l'information primaire libre en une information dite "intelligente" puisqu'elle permet de dégager les menaces et les opportunités pour une entreprise. Ces outils sont donc exploitables pour toutes sortes de veille, quelle soit économique, scientifique et technique ou technologique.

Conclusion Générale

Ce travail m'a permis d'améliorer mes connaissances sur l'interrogation des CD-ROM, des banques de données ainsi que d'Internet. On peut en retenir un certain nombre de banques de données mais surtout des sites Internet à consulter régulièrement puisque la majorité de ces ressources sont très récentes (novembre 1996).

Selon l'exploitation qui sera faite de ce document, les requêtes pourront être orientées vers l'état des recherches aux Etats-Unis avec les termes :

- "data mining" signifiant littéralement "fouille de donnée" au lieu d'analyse de données en français,
- "text information filtering" pour le filtrage de l'information.

Il est également possible d'affiner la recherche sur les méthodes de traitement de données comme les algorithmes génétiques de classification et la sériation par blocs.

Cette étude n'est évidemment pas exhaustive. Elle est un point de départ pour la surveillance des recherches de ce domaine. Elle a également permis de cerner les termes utilisés aux Etats-Unis qui sont très différents de la traduction littérale du français.

Bibliographie

J'ai choisi de structurer ma bibliographie par type de document afin d'illustrer la prépondérance de la littérature grise. D'autre part, un plan de classement thématique des références ne m'a pas paru adéquat car beaucoup de documents abordent des thèmes différents : l'analyse de données, l'indexation automatique, la classification automatique, la cartographie, ... Ils sont, de ce fait, difficile à cataloguer dans une seule catégorie.

Articles de revue :

[1] BALLARD, R.M.- « Indexing and its relevance to technical processing ».- ***Cataloging & Classification Quarterly***, Vol. 16, n° 3, 1993, p. 79-88.

[2] CISLER, Steve.- « Searching for a Better Verity, Inc's Topic Software (Micro Monitor) ».- ***Online (Weston)***, Vol. 12, n° 6, Nov. 1988, p. 99-102.

[3] DE BRITO, M.- « Information systems in natural languages : the search for an automatic indexing system ».- ***Ciencia da Informacao***, Vol. 21, n° 3, Sept./Dec. 1992, p. 223-232.

[4] DEVANBU, P. ; BRACHMAN, R. ; SELFRIDGE, P. ; BALLARD, B.- « LaSSIE: A Knowledge-based Software Information System ».- ***CACM***, Vol. 34, n° 5, Mai 1991, p. 34-49.

[5] FRAKES, W. B. ; POLE, T. P.- « Proteus: A Software Reuse Library System that Supports Multiple Representation Methods ».- ***SIGIR Forum***, Vol. 24, n° 3, 1990, p. 43-55.

[6] GOODWIN, G.- « Agenda : a PC solution for online database analysis ».- ***Database***, Vol. 14, n° 1, Fev. 1991, p. 29-33.

[7] GRANT, F.- « Software throws a fresh perspective on statistical analysis ».- ***Scientific Computing World***, n° 23, Nov. 1996, p.33, 35, 37.

[8] JACOBS, P. S. ; RAU, L. F.- « SCISOR: Extracting Information from On-line News ».- ***CACM***, Vol. 33, n° 11, Nov. 1990, p. 88-97.

[9] KOHONEN, T.- « Self organizing maps ».- ***Information sciences***, Vol. 30, 1995, [s. p.].

[10] LELU, A. ; TISSEAU-PIROT, A.-G.– « Vers une nouvelle génération de systèmes documentaires évolués : une chaîne expérimentale de génération automatique d'hypertextes ».- ***La Tribune des Industries de la Langue***, n° 15, 1994.

[11] LUNDEEN, Gerald W. ; TENOPIR Carol.- « Text Retrieval Software for Microcomputers and Beyond : An Overview and a Review of Four Packages (Product Comparison) ».- ***Database***, Vol. 15, n° 4, Août 1992, p. 51-63.

[12] MCCARTHY, Michael.- « Nexis News Plus Offers Low-Cost Preparation for On-Line Searches (Impressions) ».- ***InfoWorld***, Vol. 11, n° 43, Oct.1989, p.84.

[13] PRESTON, C.M.- « Report of the first international conference on information and knowledge management ».- ***Library Hi Tech News***, n° 102, Mai 1993, p. 7-8.

[14] PRIETO-DIAZ, R.- « Implementing Faceted Classification for Software Reuse ».- ***CACM***, Vol. 34, n° 5, Mai 1991, p. 89-97.

[15] PRITCHARD-SCHOCH, T.- « Natural language comes of age ».- ***Online***, Vol. 17, n° 3, Mai 1993, p. 33-43.

[16] RAU, L.- « Knowledge organisation and access in a conceptual information system ».- ***Information Processing & Management***, Vol. 23, n° 4, 1987, p. 419-428.

[17] RAU, L.- « Information extraction and text summarization using linguistic knowledge acquisition ».- ***Information Processing & Management***, Vol. 25, n° 4, 1989, p. 419-428.

[19] TURNER, W.A. (Cellule de Recherche en Sciences de l'Information, Paris, France).- « An infometrics program for developing research into information science ».- ***Documentaliste - Sciences de l'Information***, Vol. 27, n° 3, Août 1990, p. 123-125.

Congrès :

[20] ALLEN, C.- « Information challenges in the global marketplace ».- *Proceedings of National Online Meeting*.- New York : Learned Inf., 1994.- p.15-28.

[21] BALLARD, R.M.- « TELI : a powerful easily adaptable natural language question answering system for information retrievial ».- *Online ' 87 Proceedings of Conference*.- Anaheim : Online Inc.,1987.- p.9-13.

[22] BORZIC, B. ; DE SAINT LEGER, M. ; TURNER, W. A..- *Vers un médiateur intelligent : l'impact d'Internet sur la gestion électronique des documents*.- Paris : Communication Présentée à la 3 Conférence Hypertexte Hypermedia, Mai 1995.

[23] BORZIC, B. ; TURNER W. A.- *Vers un médiateur intelligent : l'impact d'Internet sur la gestion d'une base de données documentaire dynamique*.- [S. l.] : Texte présenté à la Journée sur les outils logiciels de la VSST, octobre 1995.

[24] BURKE, R. ; HAMMOND, K. ; KOZLOVSKY, J. (Artifical Intelligence Lab., Chicago Univ., IL, USA).- « Knowledge-based information retrieval from semi-structured text ».- *AI Applications in Knowledge Navigation and Retrieval : . Papers from the 1995 AAAI Fall Symposium*.- Menlo Park : AAAI Press, 1995.- p.15-19. (Tech. Report FS-95-03).

[25] CERRI, S.A.(Dipartimento di Scienze dell'Informazione, Milan Univ., Italy).- « Computational mathetics tool kit: architectures for modelling dialogues ».- *Intelligent Tutoring Systems : Third International Conference, ITS '96. Proceedings*.- Berlin : Springer-Verlag, 1996.- p.343-352.

[26] DE SAINT LEGER, M. ; RENNER, I. ; TURNER, W. A.- *L'évolution de la violence dans la société allemande : étude dynamique à partir de la modélisation des flux d'informations*.- Ile Rousse : Journées d'études du S.F.B.A. sur les systèmes d'information élaborée, 1995.

[27] DKAKI, T. ; DOUSSET, B.- *Génération de règles pour l'explication des résultats des classifications*.- Rabat : Premières journées de Mathématiques appliquées, Volume 1, 1992.- p. 1-7.

[28] DKAKI, T.- *Une méthode de détection et de suivi des collaborations dans le domaine de la recherche scientifique*.- Toulouse : Veille stratégique scientifique et technique, 25-27 octobre 1995.

[29] DOUSSET, B. ; DKAKI, T. ; KOUSSOUBE S.- *Aspect multitâche dans la conduite d'une analyse de données en bibliométrie : la coopération des méthodes au service de la validation des résultats*.- Ile Rousse : Journée d'études sur les systèmes d'informations élaborées : bibliométrie - Informatique stratégique, 1991.- p. 169-174.

[30] DOUSSET, B. ; DKAKI, T. ; KOUSSOUBE S.- *Les apports de la représentation de la quatrième dimension en analyse de données multidimensionnelles : une recalcification de l'analyse de données exploratoire*.- Ile Rousse : Journée d'études sur les systèmes d'informations élaborées : bibliométrie - Informatique stratégique - Veille technologique, 1991.- p. 98-105.

[31] DOUSSET, B. ; DKAKI, T. ; KOUSSOUBE S.- *Qualité de l'information et analyse de données.*- .- Ile Rousse : Journée d'études sur les systèmes d'informations élaborées : Bibliométrie - Informatique stratégique - Veille technologique, 1993.- p. 169-174.

[32] EVANS, D. A. ; GINTHER-WEBSTER, K. ; HART, M. ; LEFERTS, R. G. ; MONARCH, I. A.- « Automatic Indexing Using Selective NLP and First-Order Thesauri ».- *Proceedings of the Conference on Intelligent Text and Image Handling.*- Barcelona : [S. l.], Avril 1991, p. 624-643.

[33] FOUNTAIN, A.M. ; HALL, W. ; HEATH, I. ; DAVIS, H.C.- « MICROCOSM : An open model for hypermedia with dynamic linking, in A. Rizk, N. Streitz and J. Andre eds. Hypertext : Concepts, Systems and applications ».- *The proceedings of the European Conference on Hypertext, INRIA, France.*- Cambridge : University press, 1990. – [x p.].

[34] FRAKES, W.B.- « A study of the impact of representation in information retrieval systems ».- *The information community : an alliance for progress proceedings of the 44th ASIS Annual Meeting 1981.*- NewYork : Knowledge Industry Publications, 1981.

[35] GIRARDI, M. R. ; IBRAHIM, B.- « New Approaches for Reuse Systems ».- *Position Paper Collection of Second International Workshop on Software Reuse*, Mars 1993.

[36] GIRARDI, M.R. ; IBRAHIM, B.- « An approach to improve the effectiveness of software retrieval ».- *Proceedings of the 3rd Irvine Software Symposium (ISS'93).*- Irvine : IRUS, Avril 1993, p. 89-100.

[37] GIRARDI, M.R. ; IBRAHIM, B.- « A software reuse system based on natural language specifications ».- *Proceedings of 5th International Conference on Computing and Information (ICCI'93).*- [S. l.] : IEEE Computer Society, Mai 1993, p. 507-511.

[38] GIRARDI, M.R. ; IBRAHIM, B.- « A Similarity Measure for Retrieving Software Artifacts ».- *Proceeedings of Sixth International Conference on Software Engineering and Knowledge Engineering (SEKE'94).*- [S. l.] : [S. n.], Juin 1994, p. 478-485.

[39] GULIKERS, L. ; WILLEMSE, R.- « A Lexicon for a Text-to-Speech System ».- Proceedings of the ICSLP.- [S. l.] : [S. n.], 1992, [x p.].

[40] HANSEN, M.D. ; GRIGGS, E.B. ; RUNYARD, R.C. ; STEINER, M.A. ; SIGSBEY, R.C.(Ford Aerosp. Corp., Colorado Springs, CO, USA).- « User-system interface (USI) prototyping ».- *Proceedings of the Human Factors Society 33rd Annual Meeting. Perspectives.*- Santa Monica : Human Factors Society, Vol.2, 1989.- p.1158.

[41] HAYES, P.J.- « Intelligent text technologies and their successful use by the information industry ».- *Proceedings of the 14th National Online Meeting.*- [New York] : Learned Information, 1993.- p.189-196.

[42] HOLLOWAY, J. (Comput. Services, Indiana Univ., Bloomington, IN, USA).- « Culture, cost, and conscience: a strategy for introducing electronic computing publications at Indiana University ».- *Proceedings. ACM SIGUCCS 1993. Toward New Horizons.*- New York : ACM, 1993.- p. 424-433.

[43] KOUSSOUBE S. ; DOUSSET, B. ; DKAKI, T.- *Outils et méthodologie d'études de la cohérence et de la complétude dans les bases de connaissances.*- Dourdan : Journées francophones de la Validation et de la Vérification des systèmes a base de connaissances, 1992.- p 17-24.

[44] LINDLEY, D.- « Classifying news ware databases using hierachical cluster analysis ».- *Online Information'93 : proceeding of the 17th International Online Information Meeting.*- Londres : 7-9 Decembre 1993.- p. 149-159.

[45] SALTON, G. ; SMITH, M.- « On the Application of Syntactic Methodologies in Automatic Text Analysis ».- *Proceedings of the Twelfth Annual International ACMSIGIR Conference on Research and Development in Information Retrieval (SIGIR'89).*- Cambridge : Special Issue of SIGIR Forum, Juin1989, p. 137-150.

[46] SIMSOVA, S. (Data Help, London, UK).- « Representation of cultures on the Usenet: a survey ».- *Online Information 95 : 19th International Online Information Meeting Proceedings*.- Oxford : Learned Inf., 1995.- p.437-454.

Ouvrages :

[47] BERTIN, J.- *La graphique et le traitement graphique de l'information*.- Paris : Flammarion, 1977.

[48] CELEUX, G. ; DIDAY, E. ; GOVAERT, G. ; LECHEVALIER, Y. ; et al.- *Classification automatique des données*.- Paris : Dunod, 1989.- 285 p. (Dunod Informatique)

[49] CENTRE FRANÇAIS DU COMMERCE EXTERIEUR : Direction des industries et des services.- *l'intelligence économique et concurrentielle aux Etats-Unis*.- Paris : CFCE, 1993.- 308 p.

[50] DUBOIS, D.- *Sémantique et cognition : Catégories, prototypes, typicalité*.- Paris : Ed. du CNRS, 1991.

[51] FAURE, C.- *Descriptions et classifications*.- Paris : Ecole Nationale Supérieure des Télécommunications, 1992.- 7 p.

[52] LEBART, L. ; SALEM, A.- *Statistique textuelle*.- Paris : Dunod, 1994.

Rapport :

[53] BERNSEN, N.O., et al.- *European Strategic Research in Speech and Natural Language*.- [S. l.] : ELSNET Research Coordination Task Group, 1993.- [x p.]
Rapport : ELSNET Research Coordination Task Group : 1993

[54] KOHONENE, T., et al.- *Newsgroup Exploration with WEBSON method and browsing interface*.- Helsinky : University of technology, Laboratory of computer and information science, [s. d.].- [x p.]
Rapport : Helsinky : University of technology, Laboratory of computer and information science : [s. d.]

Symposium :

[55] DKAKI, T. ; DOUSSET, B.- *Competitive intelligence : Data extraction and analysis : International Symposium on Intelligent Data Analysis (IDA'95)*.- Baden-Baden (Germany) : 17-19 August 1995.

Thèses :

[56] BALDIT, Patrick ; DOU, Henry.- *La sériation des similarités spécifiques : un outil pour la recherche de l'information stratégique en veille technologique*.- Aix-Marseille 3 : CRRM, 1994.- [x p.].
Thèse de doctorat : Sciences et Techniques : Aix-Marseille 3 : 1994.

[57] BOUQUET, Valérie ; DOU, Henry.- *Système de veille stratégique au service de la recherche et de l'innovation de l'entreprise : principes-outils-application.*- Aix-Marseille 3 : CRRM, 1995.- [x p.].
Thèse de doctorat : Sciences et Techniques : Aix-Marseille 3 : 1995.

[58] CASTANO, Eric ; DOU, Henry.- *Conception et installation d'un système de veille technologique : application au domaine pétrolier.*- Aix-Marseille 3 : CRRM, 1994.- [x p.].
Thèse de doctorat : Sciences et Techniques : Aix-Marseille 3 : 1994.

[59] DKAKI, T. ; LAUDET, M.- *Outils informatiques et méthodes automatiques pour la veille technologique.*- Toulouse : [S.n.], 1993.- [x p.].
Thèse de doctorat : Sciences et Techniques : Toulouse : 1993.

[60] GEORGEL, A.- *Classification statistique et réseaux de neurones formels pour la représentation des banques de données documentaires.*- Paris VII : CERSI, CNRS, 1992.- [x p.].
Thèse de doctorat : Sciences et Techniques : Paris VII : 1992.

[61] LI, Z.- *Information retrieval for automatic link creation in hypertext systems.*- Southampton : University, 1993.- [x p.].
Thèse : Sciences et Techniques : Southampton : 1993.

[62] MICHELET, B.- *L'analyse des associations.*- Paris VII : SERPIA, INIST, CNRS, 1988.- [x p.].
Thèse de doctorat : Sciences et Techniques : Paris VII : 1988.

[63] PINCHARD, Frédéric ; COIFFET, P.- *Etude et réalisation d'une base de données documentaire pour la robotique.*- Paris 6 : [S.n.], 1994.- [x p.].
Thèse de doctorat : Sciences et Techniques :Paris 6 : 1994.

[64] ROSTAING, Hervé ; DOU, Henry.- *Veille technologique et bibliométrie : concepts, outils et applications.*- Aix-Marseille 3 : CRRM, 1993.- [x p.].
Thèse de doctorat : Sciences et Techniques : Aix-Marseille 3 : 1993.

[65] SAMIER, Henry ; DUCHAMP, R.- *Contribution de la veille technologique à la conception de produits.*- Paris : ENSAM, 1995.- [x p.].
Thèse de doctorat : Sciences et Techniques : Paris : 1994.

[66] TEIL, Geneviève ; CALLON, Michel.- *CANDIDE, un outil de sociologie assistée par ordinateur pour l'analyse quali-quantitative de gros corpus de textes.*- Paris : ENSAM, 1991.- [x p.].
Thèse de doctorat : Sciences et Techniques : Paris : 1991.

Ressources Internet :

[67] *APTEX-The content mining compagny*, [En ligne], (Page créée le 1er décembre 1995).- Adresse URL : http://www.aptex.com

[68] *Cambridge Research Laboratory-USA*, [En ligne], (Page créée en 1995).- Adresse URL : http://www.research.digital.com/CRL/home.html

[69] *George town linguistics : Graduate courses*, [En ligne], (Page créée le 26 Octobre 1996).- Adresse URL : http://www.georgetown.edu/cball/lx_courses.html

[70] *IDT 96-Presse*, [En ligne], (Page créée le 118 Mars 1996).- Adresse URL : http://ww.idt.fr/fran/cp_int-intel.html

[71] *Lexis-Nexis*, [En ligne], (Page créée en 1997).- Adresse URL : http://www.lexis-nexis.com/infopro/caltrain.html

[72] *Sawtooth software's web page*, [En ligne], (Page créée le 22 Novembre 1996).- Adresse URL : http://www.sawtoothsoftware.com

[73] CADIS, Inc.- *Welcome to CADIS : the innovators in internet and intranet information classification, publishing, storage and retrieval*, [En ligne], (Page créée en 1996).- Adresse URL : http://www.cadis.com

[74] CELEX.- *CELEX and (psycho)linguistic research*, [En ligne], (page créée le 29 Nov. 1996).- Adresse URL : http://www.kun.nl/celex/subsecs/section_psy.html

[75] DAVIS, Hugh.- *Creating user defined bundles from digital Library*, [En ligne], (Page créée le 9 Juin 1995).- Adresse URL : http://www.csdl.tamu.edu/DL95/papers/davis/davis.html

[76] DE SAINT LEGER, Mathilde ; CERESI/CNRS Centre d'Etudes et de Recherche En Sciences Infométriques/ Centre National des Recherches Scientifiques.- *DynaTools : un outil de gestion dynamique des flux d'informations pour une veille scientifique*, [En ligne], (Page créée en 1996).- Adresse URL : http://atlas.irit.fr/vsst/toulouseM2.html

[77] FAHMY, Thierry.- *xlSTAT data analysis toolbox*, [En ligne], (Page créée le 23 Novembre 1996).- Adresse URL : http://www.inapg.inra.fr/~fahmy/toolsfr.html

[78] GIRARDI, M.R. ; IBRAHIM, B.- *Automatic Indexing of software artifacts*, [En ligne], (Page créée le 7 Juin 1996).- Adresse URL : http://cuiwww.unige.ch/eao/www/ROSA.papers/SR94/paper.html

[79] HUMPHREY, Pete.- *Natural Language Processing at EDS*, [En ligne], (Page créée le 9 Avril 1992).- Adresse URL : http://www.edsr.eds.com/edsr/papers/natlang.html

[80] KOHONEN, T.- *WEBSON – Self-organizing map for internet exploration*, [En ligne], (Page créée le 23 Février 1997).- Adresse URL : http://nucleus.hut.fi/nnrc/new_book.html

[81] Laboratory for Advanced Information Technology.- *UMBC CSEE Department*, [En ligne], (Page créée en 1995).- Adresse URL : http://www.cs.umbc.edu/lait/

[82] LELU, Alain.- *De l'émergence des concepts : réflexions à partir du traitement « neuronal » des bases de données documentaires*, [En ligne], (Page créée le 10 Septembre 1996).- Adresse URL : http://www.info.unicaen.fr/bnum/jelec/Solaris/d02/2lelu.html

[83] LEXIS NEXIS COMMUNICATION CENTER.- *How competitive intelligence professionals use the LEXIS-NEXIS services*, [En ligne], (Page créée en 1997).- Adresse URL : http://www.lexis-nexis.com/ci

[84] McGraw-Hill College Division.- *Chapter 6 : Marketing research and information technology*, [En ligne]. (Page créée le 1 Novembre 1996).- Adresse URL : http://www.tmhe.com/ced/ch6.html

[85] MCJONES, Paul.- *SRC Research : access to information*, [En ligne], (Page créée le 9 Août 1995).- Adresse URL : http://www.research.digital.com/SRC/org/information.html

[86] NOYER, Jean-Max.- *Utilisation d'outil infométrique « Candide » dans le contexte d'une réflexion stratégique*, [En ligne], (Page créée en 1995).- Adresse URL : http://www.info.unicaen.fr/bnum/jelec/Solaris/d01/1turner.html

[87] PIRA.- *Oasis Green Paper*, [En ligne], (Page créée en 1996).- Adresse URL : http://www.pira.co.uk/people/david/public/Everything/4_3_6.html

[88] The Queen's University of Belfast.- *Data mining notes data mining techniques*, [En ligne], (Page créée le 26 Mars 1996).
Adresse URL : http://www-pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_4.html

[89] TURNER, William A.- *Penser l'entrelacement de l'Humain et du Technique : les réseaux hybrides d'intelligence*, [En ligne], (Page créée en 1994).- Adresse URL : http://www.info.unicaen.fr/bnum/jelec/Solaris/d01/1turner.html

[90] VOGEL, Claude.- *SEMIO Corporation – SemioMap Search*, [En ligne], (Page créée en 1996).- Adresse URL : http://www.semio.com

News :

[91] *INFORMATION TODAY via NewsNet*, [En ligne], (Page créée le 24 Oct.1996).- Adresse URL : http://www.newsnet.com/libiss/pb07.html

[92] SCHWARTZ, Ray.- *SIG/CR News April / 1995*, [En ligne], (Page créée le 17 Oct. 1996).- Adresse URL : http://cpmcnet.columbia.edu/www/asis/news951.html

Table des matières