



HAL
open science

Contribution à l'analyse et à la recherche d'information en texte intégral : application de la transformée en ondelettes pour la recherche et l'analyse de textes

Nabila Smail

► To cite this version:

Nabila Smail. Contribution à l'analyse et à la recherche d'information en texte intégral : application de la transformée en ondelettes pour la recherche et l'analyse de textes. Sciences de l'information et de la communication. Université Paris-Est, 2009. Français. NNT : 2009PEST1016 . tel-00504368

HAL Id: tel-00504368

<https://theses.hal.science/tel-00504368>

Submitted on 20 Jul 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THÈSE

pour obtenir le grade de

Docteur de l'Université Paris-Est

Spécialité :

Information Scientifique et Technique

présentée et soutenue publiquement par

Nabila SMAIL

le 27 Janvier 2009

Titre : Contribution à l'analyse et à la recherche d'information en
texte intégral.

Application de la Transformée en Ondelettes pour la
recherche et l'analyse de textes.

Jury :

Monsieur le Professeur S. CACALY, Directeur de thèse
Monsieur le Professeur J.KISTER, Rapporteur,
Monsieur le Professeur H. DOU, Rapporteur
Monsieur le Professeur L. QUONIAM, Examineur
Monsieur le Professeur R. EPPSTEIN, Examineur
Monsieur le Professeur C. LONGEVIALLE, Examineur

À MON PÈRE,

REMERCIEMENTS

Je tiens à remercier M. Serge CACALY d'avoir accepté de diriger mes travaux de recherche.

Je tiens à exprimer ma reconnaissance à M. Renaud Eppstein, pour avoir Co-encadré mes travaux. Je le remercie pour sa disponibilité, son écoute et ses conseils, qui m'ont été toujours précieux, sa confiance, son investissement scientifique et humain qui ont été essentiels à la réalisation de ce travail.

Je voudrais également exprimer toute ma gratitude aux professeurs L. KISTER et H.DOUCHE qui, en leur qualité de rapporteurs, m'ont fait l'honneur d'examiner minutieusement ce travail.

Je remercie J.QUONIAM et C.LONGEVIALLE, je leur en suis reconnaissante et les remercie d'avoir accepté de faire partie du jury de ma thèse.

Je tiens aussi à remercier pour son accueil toute l'équipe du laboratoire Sciences et Ingénierie de l'Information et de l'Intelligence Stratégique (S3IS) de l'Université Paris-Est où j'ai effectué cette thèse.

Je remercie également Christian LONGEVIALLE et Christel PORTE de l'équipe CESD localisé à l'IUT de champs sur Marne pour leur accueil, leur encouragement et leur aide professionnel et personnel.

Je remercie tous les membres du département Services et Réseaux de Communication de l'IUT de Champs sur Marne en particulier : Martine THIREAU, Agnès GILLET, Nicolas CLASSEAU, ainsi que tous le corps enseignants.

Enfin, je remercie toute ma famille et tout particulièrement ma mère, de m'avoir soutenue et encouragée, ma sœur Linda pour son aide dans les moments difficiles.

RESUME

L'objet des systèmes de recherche d'informations est de faciliter l'accès à un ensemble de documents, afin de permettre à l'utilisateur de retrouver ceux qui sont pertinents, c'est-à-dire ceux dont le contenu correspond le mieux à son besoin en information. La qualité des résultats de la recherche se mesure en comparant les réponses du système avec les réponses idéales que l'utilisateur espère recevoir. Plus les réponses du système correspondent à celles que l'utilisateur espère, plus le système est jugé performant.

Les premiers systèmes permettaient d'effectuer des recherches booléennes, c'est à dire, des recherches où seule la présence ou l'absence d'un terme de la requête dans un texte permet de le sélectionner. Il a fallu attendre la fin des années 60, pour que l'on applique le modèle vectoriel aux problématiques de la recherche d'information. Dans ces deux modèles, seule la présence, l'absence, ou la fréquence des mots dans le texte est porteuse d'information.

D'autres systèmes de recherche d'information adoptent cette approche dans la modélisation des données textuelles et dans le calcul de la similarité entre documents ou par rapport à une requête. SMART (System for the Mechanical Analysis and Retrieval of Text) [4] est l'un des premiers systèmes de recherche à avoir adopté cette approche. Plusieurs améliorations des systèmes de recherche d'information utilisent les relations sémantiques qui existent entre les termes dans un document. LSI (Latent Semantic Indexing) [5], par exemple réalise ceci à travers des méthodes d'analyse qui mesurent la cooccurrence entre deux termes dans un même contexte, tandis que Hearst et Morris [6] utilisent des thésaurus en ligne pour créer des liens sémantiques entre les termes dans un processus de chaînes lexicales.

Dans ces travaux nous développons un nouveau système de recherche qui permet de représenter les données textuelles par des signaux. Cette nouvelle forme de représentation nous permettra par la suite d'appliquer de nombreux outils mathématiques de la théorie du signal, tel que les Transformées en ondelettes et jusqu'à aujourd'hui inconnue dans le domaine de la recherche d'information textuelle.

MOTS CLES

Systemes de Recherche d'Information, Transformées en ondelettes, Analyse multi résolution, Modélisation de l'information, Analyse documentaire.

ABSTRACT

The object of information retrieval systems is to make easy the access to documents and to allow a user to find those that are appropriate. The quality of the results of research is measured by comparing the answers of the system with the ideal answers that the user hopes to find. The system is competitive when its answers correspond to those that the user hopes.

The first retrieval systems performing Boolean researches, in other words, researches in which only the presence or the absence of a term of a request in a text allow choosing it. It was necessary to wait for the end of the sixties to apply the vector model in information retrieval. In these two models, alone presence, absence, or frequency of words in the text is holder of information.

Several Information Retrieval Systems adopt a flat approach in the modeling of data and in the counting of similarity between documents or in comparison with a request. We call this approach 'bag of words '. These systems consider only presence, absence or frequency of appearance of terms in a document for the counting of its pertinence, while Hearst and Morris [6] uses online thesaurus to create semantic links between terms in a process of lexical chains.

In this thesis we develop a new retrieval system which allows representing textual data by signals. This new form of presentation will allow us, later, to apply numerous mathematical tools from the theory of the signal such as Wavelets Transforms, well-unknown nowadays in the field of the textual information retrieval.

KEYWORDS

Information Retrieval Systems, Wavelets Transforms, Multi resolution Analysis, Information modeling, Documentary analysis.

TABLE DES MATIERES

1 Sommaire

REMERCIEMENTS	3
RESUME.....	4
MOTS CLES	5
ABSTRACT	6
KEYWORDS	6
TABLE DES MATIERES.....	7
LISTE DES TABLEAUX	11
LISTE DES FIGURES	12
INTRODUCTION.....	13
CHAPITRE 1 : Cadre de la recherche d'information	16
1 Un survol de l'histoire de la Recherche d'Information	17
Introduction	17
La naissance de la recherche d'information	20
Expérimentations.....	21
Systèmes de Recherche d'Informations	22
Améliorations techniques	23
Ère Internet.....	23
La francophonie de la recherche d'informations.....	24
2 La recherche documentaire	25
3 Qu'est-ce que l'information ?.....	25
3.1. L'information documentaire	25
3.2. L'information spécialisée.....	26
4 Formes de l'information.....	26
5 Propriétés de l'information.....	27
5.1. Information structurée	27
5.2. Information non structurée	27
5.3. Information semi-structurée	28
6 Notions et définitions	29
6.1. La notion de 'besoin' dans la recherche d'information.....	29

6.2.	La notion de pertinence	29
6.3.	Structures de stockage de l'information	30
6.4.	L'utilisation d'une 'stop list'	31
7	Différentes approches d'indexation.....	32
7.1.	Définition de l'indexation	32
7.2.	Les débuts de l'indexation dans la recherche d'information.....	33
7.2.1.	Indexation manuelle avec vocabulaire contrôlé	33
7.2.2.	Le texte intégral.....	34
7.3.	Les approches actuelles	35
8	Processus et architecture d'un SRI.....	36
9	Les Modèles de Recherche d'Information	37
1.	Le modèle Booléen ou ensembliste.....	37
i.	Formulation de la requête.....	38
ii.	Les limites du modèle booléen.....	39
iii.	Recherche booléenne pondérée	40
2.	Le modèle vectoriel.....	41
i.	Vecteurs documents et vecteurs requêtes	41
ii.	Les mesures de similarité	42
iii.	La sélection des termes d'indexation	43
iv.	Les schémas de pondération.....	43
v.	Prise en compte des dépendances dans modèle vectoriel	46
3.	Le modèle LSI.....	47
4.	Le modèle DSIR.....	47
5.	Modèle probabiliste.....	48
i.	Représentation des documents et des requêtes.....	49
ii.	Fonction de correspondance.....	49
iii.	Prise en compte des dépendances dans le modèle probabiliste.....	49
6.	Le modèle logique	50
i.	Représentation des documents et requêtes	50
ii.	Fonction de correspondance	50
7.	L'évaluation des Systèmes de Recherche d'Information	50
7.1.	Le rappel : calculer l'exhaustivité de la recherche	51
7.2.	La précision : combien de non pertinent ?.....	52
7.3.	Combiner précision et rappel.....	53

CHAPITRE 2 : Modélisation et visualisation des données textuelles.....	54
Introduction	55
1. Modèles de représentation des données textuelles	55
1.1 Approche ‘sac de mots’	55
1.1.1 Identification des termes d’indexation	56
1.1.2 Méthodes d’analyse de l’information.....	57
1.1.3 Modèles de visualisation : la cartographie des données textuelles	60
1.2 Approche de document structuré.....	66
1.3 Le contexte local d’un mot dans un texte	68
1.4 Les thèmes dans un document.....	68
1.5 Visualisation multidimensionnelle spectrale.....	70
CHAPITRE 3 : Les Transformées en ondelettes et leurs utilisation actuelle.....	74
1. Pourquoi a-t-on besoin de Transformées?.....	75
1.1 Naissance de la Transformée de Fourier	76
1.1.1 Transformée de Fourier des fonctions périodiques	76
1.1.2 Transformée de Fourier des fonctions non périodiques	78
1.2 Signification physique de la Transformée de Fourier	79
1.3 Quelques applications de la Transformée de Fourier	80
1. Applications aux signaux monodimensionnels	80
2. Applications aux signaux bidimensionnels	80
3. Applications fondées sur la propagation des ondes électromagnétiques	80
1.4 Limites de la Transformée de Fourier	80
1.4.1 Analyse temps- fréquence	81
1.4.2 Principe d’incertitude d’Heisenberg.....	81
1.5 Transformée de Fourier Fenêtrée	82
1.6 La Transformée en Ondelettes.....	83
1.6.1 Définition	84
1.6.2 Les propriétés des Ondelettes.....	84
1.6.3 L’Ondelette de Haar	86
1.6.4 Exemple de calcul	87
1.6.5 L’utilisation actuelle des Ondelettes	89
Conclusion.....	95
CHAPITRE 4 : Modélisation Spectrale des données textuelles : vers un Système de Recherche d’Information Spectral	96

Introduction :	97
Exemple.....	99
1. Pourquoi une modélisation spectrale.....	101
2. Notions et fonctions.....	101
3. La mise en œuvre du Système de Recherche d'Information Spectrale	108
3.1 Modélisation thématique spectrale des documents	108
3.1.1. Algorithme de construction des signaux thématiques.....	110
3.1.2 Expérimentation : la modélisation spectrale	111
3.1.3 Résultats de l'analyse multi résolution	126
3.2 Représentation spectrale des requêtes	129
3.2.1 Introduction.....	129
3.2.2 Modélisation Spectrale des requêtes	130
3.2.3 Processus de comparaison spectrale document /requête.....	131
3.2.4 Expérimentation.....	134
3.2.5 Comparaison des résultats.....	137
3.2.6 Discussion	138
CONCLUSION	140
ANNEXE 1	143
ANNEXE 2	151
BIBLIOGRAPHIE	153

LISTE DES TABLEAUX

Tableau 1. Exemple d'un index simple	30
Tableau 2. Tableau explicatif des opérateurs logiques.....	38
Tableau 3. Différentes organisations des thématiques dans un texte	70
Tableau 4. Transformée de Haar du signal $S = [2\ 4\ 8\ 12\ 14\ 0\ 2\ 1]$	88
Tableau 5. Pertinence des mots de la phrase exemple 1. Les mots vides sont éliminés.....	100
Tableau 6. Exemple des thématiques de la base LISA.....	111
Tableau 7. Extrait de l'ensemble des thèmes représentatifs du Corpus LISA	116
Tableau 8. Pertinence des termes du texte 1 par rapport à la thématique « Library network ».....	118
Tableau 9. Pertinence des mots du texte 2 par rapport à la thématique « Library network ».....	120
Tableau 10. Transformée de Haar d'un signal de dimension 1.	123
Tableau 11. Tailles de segments à différents niveaux de résolution pour un texte de 9 mots.....	124
Tableau 12.valeurs de corrélation entre les deux signaux S1 et S2 à différents niveaux de détails.	129
Tableau 13. Corpus des données pour la recherche d'information spectrale	135
Tableau 14. Un ensemble de termes d'indexation avec leurs signaux dans les deux documents.	136

LISTE DES FIGURES

Figure 1. Extrait d'un texte en langage XML.....	28
Figure 2. Architecture d'un Système de Recherche d'Information.....	37
Figure 3. Exemple d'une représentation du modèle vectoriel avec deux documents et une requête.	42
Figure 4. Principe du rappel dans la recherche d'information	52
Figure 5. Principe de la précision dans la recherche d'information	53
Figure 6. Tour de Salton.....	61
Figure 7. Les « TilesBars » de Hearst	62
Figure 8. Un exemple de Cone Tree.....	63
Figure 9. La cartographie d'un corpus de dépêches d'agences de presses selon MOKRANE [30].....	64
Figure 10. Un cluster extrait d'un corpus « armement nucléaire » réalisé sous SAPMLER	65
Figure 11. Analyse multidimensionnelle selon Tétralogie	66
Figure 12. Prototype de visualisation Topic-O-Graphy: forme 'Wave':.....	71
Figure 13. Visualisation thématique multidimensionnelle Topic- O- Graphy	71
Figure 14. Visualisation de texte par coloriage	72
Figure 15. La structure de TSA-tree : X représente les données originales, AX_i et DX_i représente la tendance et la surprise au niveau i	91
Figure 16. Hachage aléatoire de l'image de Lena (décomposition en 3 niveaux avec l'Ondelette de Haar)	92
Figure 17. Signal de pertinence correspondant à l'exemple 1.....	99
Figure 18. Processus des différents traitements du corpus en vue d'une modélisation spectrale des documents.....	109
Figure 19. Processus de construction des signaux thématiques	110
Figure 20. Extrait de l'index réalisé sous SAPMLER.....	113
Figure 21. Cluster « Computer Network » avec les cooccurrences relatives aux termes du texte	114
Figure 22. Signal thématique 'Library network' dans le texte 1	119
Figure 23. Signal thématique 'Library network' dans le texte 1	121
Figure 24. Propriété de multi-résolution de l'Ondelette de Haar	124
Figure 25. Transformée de Haar du signal 'Library network' correspondant au texte 1 (coefficients d'approximations et de détails)	126
Figure 26. Transformée de Haar du signal 'Library network' correspondant au texte 2.....	127
Figure 27. Faible similarité entre les deux signaux thématique au niveau de similarité 0.	128
Figure 28. Similarité modérée entre les deux signaux thématique au niveau de similarité 1.....	128
Figure 29. Processus de comparaison spectrale document /requête	131

INTRODUCTION

L'objet d'un système de recherche d'information est de faciliter l'accès à un ensemble de documents, afin de permettre à l'utilisateur de retrouver ceux qui sont pertinents, c'est-à-dire ceux dont le contenu correspond le mieux à son besoin en information. La qualité des résultats de la recherche se mesure en comparant les réponses du système avec les réponses idéales que l'utilisateur espère recevoir. Plus les réponses du système correspondent à celles que l'utilisateur espère, plus le système est jugé performant.

Dans le domaine de la recherche textuelle, un système de recherche d'information doit fournir une représentation précise du texte. Une représentation similaire doit être créée pour une requête. Le modèle doit ensuite déterminer la correspondance entre un texte et une requête à partir de leurs représentations. Le système s'appuie sur un calcul de similarité pour identifier et sélectionner les documents les plus pertinents pour l'utilisateur.

Les premiers systèmes permettaient d'effectuer des recherches booléennes, c'est à dire, des recherches où seule la présence ou l'absence d'un terme de la requête dans un texte permet de le sélectionner. Il a fallu attendre la fin des années 60, pour que le modèle vectoriel soit appliqué aux problématiques de la recherche d'information. Le modèle vectoriel représente les documents et les requêtes par des vecteurs dans un espace constitué par les termes du vocabulaire d'indexation. Ce modèle est encore utilisé aujourd'hui pour sa robustesse et sa capacité d'ordonner les documents par ordre de pertinence décroissante.

Dans les deux modèles cités avant, seule la présence, l'absence, ou la fréquence des mots dans le texte est porteuse d'information. D'autres systèmes de recherche d'information utilisent cette approche dans la modélisation des données textuelles et dans le calcul de la similarité entre documents ou par rapport à une requête. SMART (System for the Mechanical Analysis and Retrieval of Text) [4] est l'un des premiers systèmes de recherche à avoir adopté cette approche. Plusieurs améliorations des systèmes de recherche d'information utilisent les relations sémantiques qui existent entre les termes dans un document. LSI (Latent Semantic Indexing) [5], par exemple réalise ceci à travers des méthodes d'analyse qui mesurent la cooccurrence entre

deux termes dans un même contexte, tandis que Hearst et Morris [6] utilisent des thésaurus en ligne pour créer des liens sémantiques entre les termes dans un processus de chaînes lexicales.

Ces méthodes considèrent les termes des documents indépendamment de leurs positions dans les textes. Ainsi, elles perdent toute information relative à l'ordre d'apparition de ces termes.

Nos travaux présentent une nouvelle méthode de la recherche d'information dans des collections de textes. Comme les méthodes de recherche classiques, l'interrogation sera basée sur la localisation des termes de la requête dans les documents, mais à la différence des méthodes classiques, nous représentons les données (documents et requêtes) dans le domaine spectral.

L'originalité de cette modélisation des informations repose sur la façon de représenter un document, non plus simplement comme un ensemble de vecteurs mais comme un ensemble de signaux décrivant le contenu du document. Cette nouvelle forme de représentation nous permettra par la suite d'appliquer de nombreux outils mathématiques connus en théorie du signal, tel que les Transformées en ondelettes et jusqu'à aujourd'hui inutilisés dans le domaine de la recherche d'information textuelle.

Nous traitons le sujet de la manière suivante : le chapitre 1 présente la problématique générale de la recherche d'information et il introduit le vocabulaire lié à ce champ de recherche. Ce chapitre présente également les différentes approches d'indexation ainsi que les principaux Systèmes de Recherche d'Information actuels, leurs principes de fonctionnement ainsi que leurs limites.

Nous abordons ensuite la modélisation et visualisation des données textuelles, en décrivant l'approche de représentation de 'sac de mots', les différentes mesures d'identification des termes d'indexation qu'elle utilise et les méthodes d'analyse de l'information utilisées dans le traitement et la cartographie traditionnelle des données textuelles. Nous présentons également dans ce second chapitre Ce chapitre décrit également une approche structurée par rapport à l'approche de 'sac de mots', qui permet de décrire et détecter la présence des différentes thématiques dans un document. Cette approche structurée constitue la base d'un nouveau mode de visualisation multidimensionnelle spectrale qui permet de détecter la présence mais également, les ruptures et les changements thématiques dans les documents longs.

Dans le chapitre 3, nous présentons la « boîte à outils » que nous proposons de mobiliser dans le cadre de l'analyse et de la comparaison documentaire. Nous présentons tout d'abord une vue d'ensemble sur la Transformée de Fourier, son principe de fonctionnement, ses applications ainsi

que ses limites. Nous présentons également la Transformée en ondelettes et ses avantages par rapport à la Transformée de Fourier et Transformée de Fourier Fenêtrée ainsi que ses utilisations récentes dans différents domaines.

Le chapitre 4 constitue le cœur de notre proposition, Nous formulons une description du modèle de Recherche d'Information Spectral. Ce modèle est décrit selon son mode de représentation des documents et des requêtes et son modèle de correspondance entre document et requête. La première partie du système décrit concerne la modélisation des textes par un ensemble de signaux thématiques. Dans cette représentation, les termes du texte seront remplacés par les pertinences relatives à un thème donné, dont le flot peut être tracé et représenté comme un signal à travers le texte. La seconde partie du système concerne la recherche d'information et le calcul de la similarité entre documents ou entre document et requête dans le domaine spectral. Les deux parties font l'objet d'évaluations sur un ensemble de tests et de comparaison avec la représentation vectorielle classique.

Enfin, nous concluons ces travaux en mettant en perspective notre contribution et en présentant les différentes pistes de recherche qui en découlent.

CHAPITRE 1 : Cadre de la recherche d'information

La Recherche d'Information abrégée en (RI) ou (IR pour Information Retrieval en anglais), est la science qui consiste à rechercher l'information dans le contenu des documents, dans les métadonnées qui décrivent les documents, ou dans les différentes bases de données relationnelles ou sur des réseaux comme le Web, Internet, ou les Intranets.

Le Vocabulaire de la documentation (ADBSⁱ, 2004) distingue la recherche **d'**information de la recherche **de l'**information :

- La recherche *d'*information est « l'ensemble des méthodes, procédures et techniques permettant, en fonction de critères de recherche propres à l'utilisateur, de sélectionner l'information dans un ou plusieurs fonds de documents plus ou moins structurés ».
- La recherche *de l'*information est « l'ensemble des méthodes, procédures et techniques ayant pour objet d'extraire d'un document ou d'un ensemble de documents les informations pertinentes ».

Dans ce premier chapitre nous allons présenter les différentes notions utilisées dans le domaine de la recherche d'information. Tels que la notion de la pertinence, le besoin en information, les différentes méthodes d'indexation actuels, ainsi que les différents systèmes de recherche d'informations actuels.

1 Un survol de l'histoire de la Recherche d'Information

Introduction

Les sociétés et les entreprises ont toujours essayé de mieux préparer leur avenir en se dotant d'outils et de méthodes afin de se rendre le plus compétitifs vis-à-vis de leurs voisins et concurrents, en utilisant les techniques de renseignement, d'espionnage et des stratégies prévisionnelles, c'est-à-dire différentes formes de veille.

La stratégie de ces organismes consiste à recueillir l'information, la synthétiser et tirer les conclusions pouvant orienter leur développement. Mais toute information ne peut contribuer à l'amélioration de la productivité et à la compétitivité d'une organisation que lorsqu'elle répond aux vrais besoins des responsables, à savoir progresser, moderniser, innover et diversifier.

Toutefois, la recherche de cette information plus qu'indispensable pour toutes les fonctions d'une organisation se heurte généralement à des obstacles de nature à réduire son efficacité, notamment :

- L'abondance des supports d'information réels et potentiels sur le marché de la communication,
- le flot de l'information pouvant entraîner l'inopportunité et la non pertinence des données lorsqu'elles ne répondent pas aux besoins précis des décideurs, alors que ces derniers ont besoin d'une information précise, analysée, filtrée et condensée.

Il s'agit ainsi, d'une information sur mesure, personnalisée et gérée pour répondre aux besoins spécifiques et de plus en plus exigeants des décideurs, or « sans gestion d'information, pas d'organisation viable ».

Des outils d'observation et de mesure ont été créés tout au long de l'histoire pour aider les sociétés à mieux mesurer leur environnement. Les Grecs ont développé des mécanismes d'observation très complexes capables de prédire les cycles de la terre. Ces mécanismes seront transmis aux horlogers européens via les arabes. Ils donneront naissance à différentes machines

de calculs (machine à calcul de Pascal, les cartes perforées de Jacquard) pour arriver à la création des premiers ordinateurs.

Depuis l'avènement d'Internet qui facilite l'accès à une grande masse de données et le développement des nouvelles technologies, la veille est à la mode, elle s'élargit, de l'entreprise privée, elle devient une affaire d'état nommée Intelligence économique. Avec la chute de l'URSS, les agences des services secrets et les militaires se sont converti au civil utilisant les moyens légaux pour la cueillette d'informations.

Les nouvelles technologies de l'information et de la communication ont ainsi conduit :

- 1- A une transformation des pratiques de gestion de l'information : les fichiers manuels se transforment en fichiers informatisés, en banques de données,
- 2- à la législation et la gestion électronique des documents qui amènent à une véritable ingénierie informationnelle,
- 3- au besoin d'être tenus correctement informés qui devient une nécessité vitale de toutes les catégories d'utilisateurs, notamment des entrepreneurs, des chercheurs, etc.

Ceci a engendré de nouvelles pratiques, entre autres :

- La veille stratégique qui consiste à surveiller l'environnement externe de l'entreprise par le service d'information afin de recueillir l'information nécessaire à la prise des décisions stratégiques et aux actions au sein de toute organisation,
- la veille technologique qui consiste à observer l'environnement technologique et suivre les évolutions qu'il subit afin de dégager les opportunités et les menaces qu'il offre et que le service d'information doit prendre en considération,
- la veille concurrentielle qui consiste à suivre de près et de manière systématique les concurrents réels et potentiels du service d'information, leur expansion dans le temps et dans l'espace, leurs produits, leurs services, leurs innovations,

- la veille commerciale qui, pour rationaliser les achats et ventes, consiste à suivre les marchés de matières premières, la situation des circuits commerciaux, etc.

Bref, l'intelligence économique, qui est une démarche globale qui vise à inclure tous les types de veille en une approche globale permettant non seulement de surveiller mais aussi de prévoir toutes les menaces et opportunités relatives au contexte concurrentiel, juridique, technologique, commercial, sociétal, etc. de l'organisation.

Etant donné ces nouvelles pratiques, le professionnel de l'information est tenu de :

- Savoir et pouvoir maîtriser l'information de veille, de découverte, d'innovation et d'ouverture sur le monde,
- savoir et pouvoir développer et exploiter l'information utile qui rend possible l'activité quotidienne des individus, des centres et des laboratoires de recherche, des entreprises,...
- valoriser l'information auto produite, tenir compte de l'information vivante, de l'information de communication, etc....
- raisonner en terme de différenciation fonctionnelle multidimensionnelle et donc de richesse d'intervention potentielle avec autant de compétences spécifiques à développer.

La veille suppose une maîtrise de l'information nécessaire à la surveillance des environnements précis (sociopolitiques et économiques). C'est un processus continu et systématique de gestion de l'information stratégique.

Un processus de veille comporte en général trois étapes essentielles :

- 1- La cueillette : il s'agit de bien rassembler les données pour dresser un bilan sur le contexte donné, ses principaux défis sont :
 - Le traitement d'un très grand volume de données dans un temps assez court,
 - la classification des données.

Dans cette étape, la recherche s'effectue dans les bases de données, les sites Web et l'échange entre veilleurs, à l'aide des répertoires, des annuaires, des bases de connaissances commerciales sur le Web, des outils linguistiques, ... etc.

- 2- l'analyse et la synthèse : cette étape sert à synthétiser les données rassemblées afin de faire afin de découvrir les principales tendances qui serviront à convertir certaines stratégies en scénarios,
- 3- la diffusion : il s'agit de présenter aux décideurs divers scénarios qui faciliteront leur prise de décision. Les principaux défis se résument en :
 - La pertinence des choix en fonction du long terme,
 - le développement de stratégies conduisant aux innovations.

Pour exécuter un tel processus, les outils de veille se divisent le plus souvent en 2 catégories :

- Les outils de recherche d'information,
- les outils de surveillance.

La *recherche d'information* concerne les mécanismes qui facilitent l'accès à une base d'informations. Il existe un grand nombre de *modèles* de recherche d'information. Ces modèles diffèrent principalement sur la façon dont les informations disponibles sont représentées et sur la façon d'interroger la base. Notre travail de thèse présenté ci-après porte sur le point particulier des outils de la recherche d'information.

La naissance de la recherche d'information

Le domaine de la recherche d'information remonte au début des années 1950, peu après l'invention des ordinateurs, les chercheurs voulaient les utiliser pour automatiser la recherche des informations, qui dépassaient les capacités humaines à cause de l'explosion de la quantité d'information après la deuxième guerre mondiale.

Le terme de recherche d'information '*Information Retrieval*' fut donné par Calvin N. Mooers en 1948 pour la première fois dans son mémoire de maîtrise [7] et la première conférence dédiée à ce thème – International Conference on Scientific Information - s'est tenue en 1958 à Washington.

Les premiers problèmes qui intéressaient les chercheurs portaient sur l'indexation des documents. Déjà à la 'International Conference on Scientific Information', Luhn avait fait une démonstration de son système d'indexation 'KWIC' qui sélectionnait les index selon la fréquence des mots dans les documents et filtrait des mots vides. C'est à cette période que le domaine de la recherche d'information est né.

Expérimentations

Dès les premiers travaux, l'aspect d'expérimentation occupait une place particulière, les chercheurs testaient toutes les méthodes. Cette tradition est restée bien ancrée dans la communauté de la recherche d'information.

Voici quelques grands projets d'expérimentations dans l'histoire de la recherche d'information.

Projet Cranfield (dirigé par Cyril Cleverdon, 1957-1967) [8] Dans la première phase de ce projet, on testait l'efficacité de différentes méthodes d'indexation et de recherche des documents. Une collection de test est constituée d'un ensemble d'articles (18 000 dans Cranfield I) et un ensemble (1 200) de requêtes. Ces requêtes sont évaluées par des experts afin de déterminer les réponses souhaitées et les résultats d'une recherche automatique sont comparés avec les réponses souhaitées pour mesurer la performance en termes de précision et rappel.

Le projet Cranfield a une influence marquante sur toute l'histoire de la recherche d'information, on utilise encore aujourd'hui les mêmes principes d'évaluation pour la plupart des systèmes.

SMART (Gérard Salton, 1^{ière} version 1961-1965) [4]

Dans ce projet, une série d'expérimentations a été menée, portant sur divers sujets comme :

- La comparaison entre l'indexation manuelle et l'indexation automatique,
- le problème de recherche d'information interactive et la rétroaction de pertinence (relevance feedback),
- l'architecture du système de recherche,
- l'utilisation du modèle vectoriel,

- le regroupement de documents (ou clustering).

Le système SMART fut réécrit dans les années 1970 et 1980 par E. Fox et C. Buckley. Il a été et est encore utilisé par de nombreux chercheurs pour des expérimentations en recherche d'information. Ce système est sans doute le système qui a eu le plus grand impact sur l'histoire de la recherche d'information.

TREC - Text REtrieval Conference, (D. Harman, 1992)

Cette série de conférences a pour objectif de tester des méthodes et des systèmes de recherche d'information avec des collections de tailles plus grandes. Elle est organisée annuellement. Les tâches (tracks) changent d'une année sur l'autre, mais elles reflètent bien les intérêts des chercheurs et les besoins réels. Par exemple : soumettre des requêtes sur une collection statique, le filtrage de l'information, la recherche d'information en d'autres langues que l'anglais (en espagnol, français, chinois) et translinguistique (trouver des documents dans une langue différente de celle de la requête), la question-réponse, la recherche d'information multimédia (vidéo et parole), etc....

Les conférences 'TREC' ont grandement contribué au développement récent de systèmes de recherche, en fournissant des collections de tests réalistes et en offrant une nouvelle méthodologie d'évaluation.

Systèmes de Recherche d'Informations

Un modèle de recherche d'information doit donner une représentation des documents et une représentation similaires doit être créée pour une requête, le modèle doit ensuite déterminer la relation entre un document et une requête à partir de leurs représentations. Ceci se fait souvent avec un calcul de similarité.

L'utilisation du modèle booléen a marqué les débuts de la recherche d'information à cause de sa simplicité et le fait qu'il soit intuitif. Cependant, dans sa version pure sans pondération, ce modèle souffre de graves lacunes (voir chapitre 1 de cette thèse). Ainsi, on proposait d'intégrer la pondération dans des modèles booléens étendus, par exemple, par l'utilisation de la théorie des ensembles flous. Le modèle vectoriel est populaire grâce à sa capacité d'ordonner les documents retrouvés, sa robustesse et ses bonnes performances dans des tests. Il est sans doute le modèle le plus souvent utilisé en recherche d'information.

Les recherches sur les modèles probabilistes ont commencé depuis le milieu des années 1970. C.J. van Rijsbergen, S. Robertson et K. Sparck Jones sont parmi les pionniers à proposer des modèles probabilistes. L'intérêt des modèles probabilistes a pris son envol à partir des années 1990, où les approches probabilistes se sont montrées performantes dans 'TREC' (par exemple, le système 'OKAPI').

Améliorations techniques

De nombreuses études ont porté sur des améliorations possibles de techniques d'indexation et de recherche. Parmi les tentatives les plus marquantes, on retrouve notamment :

- Rétroaction de pertinence (relevance feedback) : Cette technique vise à étendre la portée de la recherche en intégrant les termes issus des documents pertinents, ou des documents en tête de la liste de réponses trouvées automatiquement,
- expansion de requête : Cette technique vise à renforcer l'expression de la requête de l'utilisateur (qui est souvent très courte) par l'intégration des termes reliés (soit en exploitant un thésaurus, soit en utilisant un calcul basé sur des cooccurrences),
- regroupement (clustering) des documents : Il vise à créer une structure entre les documents selon leurs similarités. Cette structure peut aider à la fois la recherche et la présentation des résultats.

Ère Internet

Le domaine de recherche d'information fut créé à cause de l'explosion de l'information dans les années 1950. Mais cette explosion apporte de nouveaux problèmes dans le domaine de la recherche d'information.

- Sur le Web, on ne peut plus créer une collection statique. La collection (qui est le Web au complet) est une collection gigantesque qu'il est impossible (au moins pour le moment) de couvrir au complet,
- un système de recherche propose toujours des documents. Certains sont pertinents, mais noyés parmi beaucoup d'autres documents non pertinents. Plus notre collection contient des documents,

plus ce problème devient aigu. Il est de plus en plus demandé que la recherche soit plus précise, même si on doit accepter que certains documents pertinents ne soient pas retrouvés,

- l'existence des documents non textuels (image, son, vidéo, etc.) nécessite de nouvelles façons pour les indexer et les rechercher. Les méthodes traditionnelles de recherche sont surtout destinées aux textes et ne sont pas directement applicables à d'autres médias,
- l'utilisation des langues différentes pose un autre problème. Avec une requête en français, on ne peut retrouver que des documents en français. Or, la pertinence d'un document est souvent indépendante de la langue utilisée. Ainsi, nous avons besoin d'outils pour la recherche d'information translinguistique ou multilingue.

La francophonie de la recherche d'informations

Les développements en recherche d'information dans la communauté francophone ont commencé dans les années 1980. La première conférence organisée en France dédiée à la recherche d'information est 'RIO' – Recherche d'Informations Assistée par Ordinateur – tenue à Grenoble en 1985.

Dans les années 1980, peu d'équipes de recherche francophones étaient actives, l'équipe du laboratoire Génie Informatique (LGI) IMAG fut l'une des plus actives. Ce groupe a notamment développé le prototype 'IOTA' qui utilise un schéma d'indexation statistique combinant une analyse syntaxique simple du français [9].

Dans les années 1990, le domaine est devenu plus actif. Des équipes de recherche apparaissent en France, en Suisse, en Belgique, au Canada et ailleurs.

Quelques années après le lancement de 'TREC', l'INIST (Institut de l'information scientifique et technique, Nancy) a lancé un projet similaire – 'Amaryllis I' (1996-1997). Ce projet vise à tester les systèmes et les approches pour des collections de documents en français. En 2002, 'Amaryllis' s'est associé aux expérimentations de 'CLEF' – Cross-Language Evaluation Forum, organisées par la communauté européenne.

2 La recherche documentaire

Selon le vocabulaire de la documentation [10] la recherche documentaire est définie ainsi :
« *Action, méthode et procédures ayant pour objet de retrouver dans des fonds documentaires les références des documents pertinents.* »

Cette définition, ainsi que d'autres dérivées, sont bousculées par Internet qui permet à la fois :

- De rechercher des références de documents,
- de rechercher des documents entiers,
- de rechercher des informations (exploitation des documents).

Ce qui explique que l'on parle actuellement surtout de recherche d'information (RI). Elle permet de retrouver dans un fond documentaire, une information répondant à une question précise. Ce qui est donc visé c'est l'information plus que le document et le contenu plus que le contenant. C'est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information.

3 Qu'est-ce que l'information ?

Le journal Officiel définit l'information comme étant :

"Élément de connaissance susceptible d'être représenté à l'aide de conventions pour être conservé, traité ou communiqué" (définition du Journal Officiel, 28 octobre 1980)

Mais l'information est une notion polysémique, complexe, transversale : les journalistes, les informaticiens, les biologistes, les psychologues, les documentalistes... définissent, chacun à leur manière, d'information, dans des sens différents.

3.1. L'information documentaire

L'information documentaire renvoie à la connaissance (*knowledge*) et présente les caractéristiques suivantes :

- Elle apporte du nouveau, elle enrichit les connaissances d'un individu,
- elle a un sens pour l'utilisateur,

- elle sert à agir, à prendre des décisions : en général, on cherche de l'information pour éclairer une décision, une action, ...
- elle répond à des objectifs, à des besoins, plus ou moins bien définis,
- elle n'existe que si on l'interroge : il n'y a pas d'information "en soi", l'information est toujours relative à un sujet, un contexte, un besoin, ...

3.2. L'information spécialisée

L'information spécialisée est une des catégories de l'information documentaire, elle comprend notamment :

- **L'information professionnelle** : représentée par l'information financière, économique, sociale, technique, etc. ..., destinée à un ou plusieurs secteurs professionnels,
- **l'information juridique** : lois, règlements, information administrative,...
- **l'information scientifique et technique** : elle regroupe toute l'information et les documents, produits et diffusés par les chercheurs, dans toutes les disciplines scientifiques. Une thèse, un rapport de recherche, un article dans une revue scientifique, un brevet, un mémoire, des actes de congrès... sont des documents contenant de l'information scientifique, qu'il s'agisse de Littérature, de Sociologie, de Chimie ou de Mathématiques,...

4 Formes de l'information

Internet et les bases de données représentent des sources d'information très vastes et il est important de connaître les différents types d'information disponible, ainsi que ces formats afin que le processus de recherche d'information soit le plus efficace possible.

- **Informations textuelles** : tous les documents où prédominent le texte, l'écrit (livres, périodiques, etc.)
- **Informations non-textuelles** : tout document où prédominent l'image, le son ou la combinaison des deux ou des trois, image, son et écrit (documents iconographiques, audiovisuels, multimédia, sonores...).

5 Propriétés de l'information

Trois types d'informations peuvent être distingués :

Information structurée, information non structurée et information semi structurée.

5.1. Information structurée

C'est l'information stockée dans les bases et les banques de données de façon à être traitée automatiquement et efficacement par des logiciels. Dans l'exemple d'une base de données bibliographique, une notice bibliographique est structurée par des champs. Chaque champ contient des informations validées qui peuvent être de deux ordres :

- Les informations factuelles sont par exemple le champ « auteur », « organisme d'affiliation », ...
- les informations descriptives apportent une information analytique sur le contenu du document source. Ce sont par exemple les informations contenues dans les champs « titre », « résumé », ou bien le champ « mots-clés ».

5.2. Information non structurée

Toute l'information ne peut pas être de type structurée, par exemple les lettres, les courriels, les livres, les brevets, Dans certains documents textuels, on peut différencier plusieurs niveaux d'information structurée et celle non structurée, par exemple un courriel est sous la forme suivante :

Date :

De :

A :

Sujet :

Corps du message :

Sur cet exemple, on a un mélange d'information structurée représentée dans les champs : date, auteur et destinataire, le corps du message représente une information non structurée.

5.3. Information semi-structurée

ABITEBOUL dans [11] donne la définition suivante: « *Par **semi-structuré**, nous signifions que même si les données possèdent une structure, celle-ci n'est pas aussi rigide, aussi régulière ou complète que la structure requise par les systèmes de gestion de bases de données traditionnels.* » Et dans [12] « *Nous appelons [...] donnée semi-structurée la donnée qui n'est (d'un certain point de vue) ni une donnée brute ni une donnée strictement typée* ».

Ce concept recouvre, par exemple, le langage XML (eXtended Markup Language) qui permet de définir la structure et la présentation de documents et de données de tout type (texte, multimédia, dessins techniques, formules mathématiques ou chimiques, informations comptables ou financières, transactions commerciales... . C'est le cas de la majeure partie des informations que l'on peut trouver sur le Web, dans lesquelles on retrouve :

- Du texte libre,
- des informations structurées explicites descriptives comme les mots-clés que l'on peut indiquer dans une balise appropriée,
- des informations explicites, structurées et factuelles comme le nom de l'auteur ou le titre.

```
<?xml version="1.0" encoding="iso-8859-1" ?>
- <newsitem itemid="2286" id="root" date="1996-08-20"
xml:lang="en">
<title> MEXICO: Recovery excitement brings Mexican markets
to life.</title>
<headline> Recovery excitement brings Mexican markets to life.
</headline>
<byline> Henry Tricks </byline>
<dateline> MEXICO CITY </dateline>
- <text>
<p>Emerging evidence that Mexico's economy was back on the
recovery track sent Mexican markets into a buzz of excitement
Tuesday,
with stocks closing at record highs and interest rates at 19-month
lows.</p>
</text>
<copyright>(c) Reuters Limited 1996</copyright>
>
```

Figure 1. Extrait d'un texte en langage XML

6 Notions et définitions

6.1. La notion de ‘besoin’ dans la recherche d’information

La notion de ‘besoin d’information’ est centrale dans le domaine de la recherche d’information puisque elle est définie comme une interaction entre « un individu qui a besoin d’information » et « un document qui contient ou non la réponse à ce besoin » [13].

L’utilisateur doit donc formuler une requête, c’est-à-dire exprimer son besoin en information sous forme de descripteurs ou mots clés plus au moins liés, dont la relation est exprimée par la présence d’opérateurs entre eux. La requête peut s’effectuer sur l’ensemble des mots du texte, ou dans certaines zones précises du document, lorsque l’information est indexée et structurée selon différents champs (titre, auteur, ...).

6.2. La notion de pertinence

Pour être en mesure d’offrir aux utilisateurs les informations répondant le mieux à leurs besoins, tout système de recherche d’information s’appuie sur un modèle de calcul de pertinence qui, pour chaque requête, calcul le score de pertinence de chaque donnée (document). Celles qui auront le meilleur score de pertinence seront présentées à l’utilisateur.

Cette approche permet d’évaluer ce qu’on nomme la pertinence système, c’est-à-dire la pertinence que les systèmes de recherche d’information calculent. Or, La notion de pertinence est très complexe, elle est évaluée par les systèmes de recherche d’information et également liée au jugement des utilisateurs.

On distingue classiquement deux types de pertinence : la **pertinence utilisateur**, qui est le jugement apporté par l’utilisateur sur le document, en fonction de son besoin d’information, et la **pertinence système**, qui correspond à la valeur de correspondance entre le document et la requête, calculée par les systèmes. La satisfaction de l’utilisateur est liée à la correspondance entre ces deux pertinences.

Un étudiant en droit qui doit étudier un cas précis et qui dispose du corpus de toute la jurisprudence du droit français et ne disposant que d’un accès chronologique ou thématique aux documents, va chercher à identifier dans son besoin en information les critères qui peuvent cerner soit la période pendant laquelle des actes de jurisprudences qui lui sont pertinents ont pu être émis, soit la thématique traité dans sa requête. D’autres critères vont certainement intervenir dans

l'estimation de la pertinence d'un document. Certains documents ne seront pas utiles, car déjà connus, d'autres peuvent être éliminés puisque ils demanderaient trop de travail pour être utilisés. Cet exemple donne une idée sur la grande diversité des facteurs qui interviennent lorsqu'un utilisateur évalue la pertinence d'un document.

Il existe une distance plus ou moins grande entre les résultats d'un système de recherche d'information et les jugements de pertinence de l'utilisateur. L'utilisation d'un système de recherche d'information est plus généralement conçue comme un processus itératif visant à améliorer progressivement l'adéquation entre pertinence système et pertinence utilisateur. Pour ce faire, une nouvelle fonction est très fréquemment ajoutée au schéma fonctionnel classique : le bouclage de pertinence (relevance feedback). Une fois un premier ensemble de documents retrouvés, l'utilisateur peut émettre des jugements de pertinence sur ces documents, jugements qui sont pris en compte pour définir une nouvelle requête (reformulation de la requête).

6.3. Structures de stockage de l'information

Tout système d'indexation permet d'extraire, d'un corpus textuel, les termes qui le représentent, l'identifient au mieux et de les stocker dans un index. Ces termes-index sont comparés avec ceux de la question posée. Ensuite, la fonction de recherche fournit une réponse comprenant des informations triées.

Les structures de stockage de la plupart des systèmes de recherche d'information sont basées sur le comptage des fréquences des termes dans chaque document du corpus. Chaque document D est représenté par un vecteur de valeurs, la première valeur représente la fréquence du terme t_0 dans le document D_0 , la seconde représente la fréquence du terme t_1 dans le document D_0 , etc.... Sous cette approche les documents se présentent sous forme de tableau, où les lignes représentent les documents et les colonnes représentent les termes.

	T_0	T_1		T_m
D_0	0	1	0
D_1	2	1		0
.....			
D_n	0	1		3

Tableau 1. Exemple d'un index simple

Après avoir enregistré pour chaque document, la liste des termes qu'il contient, on crée un fichier inversé qui dresse, pour chaque terme, la liste des documents qui le contiennent.

Le fichier inversé est un index lexicographique, c'est-à-dire une table alphabétique de mots-clés accompagnés de références. Il permet à partir d'un mot-clé donné de trouver toutes ses occurrences au sein d'une collection de documents. Dans le cas général, il comporte, pour chaque terme d'indexation, une liste (appelée < *posting list* > ou parfois < *posting* >) contenant l'identifiant des documents dans lesquels il apparaît ainsi que sa fréquence d'apparition. Dans le cas où le fichier inversé mémorise en plus toutes les positions de chaque occurrence, le fichier inversé est dit : complet (*full inverted file*).

L'avantage de cette structure est qu'elle permet de représenter, avec efficacité, l'ensemble de la collection des documents. Ainsi, en conservant une seule occurrence de chacun des termes d'indexation, elle diminue l'espace mémoire nécessaire et elle accélère la recherche car elle supprime tout besoin d'accès aux documents d'origine : le fichier inversé contient toutes les informations utiles et la plupart des calculs numériques peuvent être effectués au moment de l'indexation.

6.4. L'utilisation d'une 'stop list'

Une 'stop list' est une liste de mots qu'on juge *inutiles* et qu'on va retirer du document. Typiquement, il s'agit des hapax (mots n'apparaissant qu'une seule fois), des N mots les plus fréquents (par exemple les 4 mots apparaissant plus 1000 fois dans le cas où la fréquence des mots varie de 1 à 1004).

Dans un corpus de documents rédigés dans une même langue, les mots vides sont principalement des mots caractéristiques de cette langue comme les prépositions, les articles, les pronoms d'où l'assimilation courante entre mots vides et mots grammaticaux. En français, des mots vides évidents pourraient être « le », « la », « de », « du », « ce », « ça », En anglais, ce sont des mots comme « and », « there », « some », « who », « of »,...

Cependant dans une collection de textes réunis autour d'un thème commun, certains mots peuvent respecter une distribution uniforme. Ce sont alors des mots vides pour cette collection bien qu'ils ne soient pas des mots grammaticaux. Par exemple, dans un corpus de textes légaux, le mot '*loi*' est un mot vide.

7 Différentes approches d'indexation

Le schéma fonctionnel classique pour les systèmes de recherche d'information comprend deux fonctions principales : l'indexation et l'interrogation.

Afin d'effectuer une recherche d'information efficace et pertinente, il apparaît comme nécessaire de donner une représentation mieux structurée et si possible normalisée du contenu des documents. Lors de l'interrogation, il faut également transformer la requête de l'utilisateur exprimée en langage naturel, en une représentation structurée et normalisée, qui va permettre d'apparier celle-ci avec la représentation du contenu des documents.

La fonction de correspondance est la fonction de recherche proprement dite. Le système met en correspondance les documents indexés avec la requête de façon à sélectionner un sous-ensemble des documents du corpus.

7.1. Définition de l'indexation

La définition proposée par l'AFNOR en 1993, est la suivante : « *l'indexation est le processus destiné à représenter par les éléments d'un langage documentaire ou naturel des données, résultat de l'analyse du contenu d'un document ou d'une question* ».

L'indexation a un double but de représentation :

- D'une part, elle consiste à identifier les informations caractéristiques du contenu d'un ou plusieurs documents,
- d'autre part, elle consiste à représenter ces informations sous une forme compacte, homogène (le plus souvent par un ensemble de termes empruntés à une langue naturelle ou un langage documentaire) et manipulable, c'est-à-dire utilisable par un Système de Recherche d'Information par exemple.

Le but général de l'indexation est d'identifier l'information contenue dans tout texte et de le représenter au moyen d'ensemble appelé index pour permettre la comparaison entre la représentation d'un document et d'une requête.

7.2. Les débuts de l'indexation dans la recherche d'information

La première difficulté rencontrée lors de l'indexation consistait à résoudre les problèmes linguistiques les plus visibles :

- L'ambiguïté : lorsqu'une phrase ou une expression possède plusieurs interprétations ou significations possibles, on parle d'ambiguïté.
 - Lorsque l'on ne sait pas à quoi rapporter une expression, on parle d'ambiguïté de référence. Par exemple, dans la phrase « je vois un homme avec un télescope », on ne sait si c'est moi qui regarde l'homme à l'aide d'un télescope ou si je regarde un homme qui possède un télescope,
 - lorsqu'une phrase possède plusieurs analyses syntaxiques, on parle d'ambiguïté structurale ou syntaxique. Par exemple, dans la proposition « la bonne cuisine », on ne sait si bonne l'adjectif ou bien c'est le nom.

- La synonymie : un même concept peut être exprimé par des mots différents
- La polysémie : un même mot peut renvoyer sur différents concepts. C'est une caractéristique très fréquente du langage courant, c'est la mise en discours qui permet parfois de lever l'ambiguïté. Par exemple, le terme « canard » peut signifier un journal, un oiseau,... par contre la phrase « j'ai lu dans le canard... » lève naturellement l'ambiguïté.

7.2.1. Indexation manuelle avec vocabulaire contrôlé

L'indexation manuelle consiste à représenter le contenu d'un document par une liste de groupes nominaux qui expriment les principaux thèmes traités dans le document. C'est un exercice très subjectif, étant donné qu'il dépend des connaissances du documentaliste sur le sujet traité dans le texte et par conséquent de la manière dont il va hiérarchiser les thèmes retenus.

Une fois la liste de groupes nominaux dressée, nous pouvons distinguer les descripteurs qui seront habilités à figurer dans l'index et les non-descripteurs ou les termes qui eux ne figureront pas dans l'index. Les descripteurs et les non-descripteurs seront liés par des relations sémantiques, l'ensemble de ces trois éléments constitue un graphe appelé Thésaurus.

Le thésaurus s'attaque à deux problèmes différents celui de la synonymie et celui de la polysémie. Le principe est de constituer pour chaque concept la liste des mots qui peuvent l'exprimer. L'un des mots est alors choisi comme descripteur, les autres mots sont des non-descripteurs et leur usage dans l'index est interdit. Concernant la polysémie, le problème à éviter c'est de récupérer des documents non pertinents lors de l'interrogation, ce qui est fort possible lorsqu'on utilise des mots ayant plusieurs sens, une solution serait d'utiliser comme descripteur celui qui n'est pas ambigu.

Outre l'aide qu'il apporte pour résoudre les ambiguïtés du langage, le thésaurus aide le documentaliste dans l'exhaustivité de la description par des relations sémantiques de « suggestion » de nouveaux termes à mettre dans l'index. Il existe deux types de relations sémantiques de suggestion :

Les relations hiérarchiques : (relations termes génériques et termes spécifiques), par exemple :
Europe → Terme spécifique → France
← Terme générique ←

La relation de terme associé : il peut s'agir de termes que l'on trouve fréquemment associés : relation entre un agent d'une action et l'action ou des termes co-occurents.

Si l'indexation manuelle sur vocabulaire contrôlé permet une recherche sur des thèmes assez généraux de manière assez efficace par un personnel formé, ses principaux inconvénients sont la perte importante d'information par rapport au texte intégral et par conséquent la difficulté de répondre à des questions très précises. De plus, l'indexation manuelle nécessite un cout financier non négligeable, il apparaît donc illusoire d'envisager son utilisation pour des gros volumes de documents.

7.2.2. Le texte intégral

C'est dans le milieu dans années 70 que sont apparues les premières bases de textes intégraux et notamment dans le domaine juridique. La réponse à une requête n'est alors plus constituée d'une référence mais du texte du document ce qui constitue un vrai progrès. La technique la plus utilisée consiste à prendre comme mot d'index chaque chaîne de caractères comprise entre deux blancs, à l'exception des mots vides. Ces derniers représentent à eux seuls près d'un tiers d'un texte.

Un des principaux problèmes de l'indexation en texte intégral est qu'elle ne tient pas compte des problèmes linguistiques (synonymie, polysémie). Le problème de la synonymie s'est plus aggravé, car on trouve dans le texte intégral les mots fléchis et dérivés, mal orthographiés ou encore pouvant accepter plusieurs variantes orthographiques. Il s'est donc avéré nécessaire de trouver des solutions mais ces dernières engendrent une complication non négligeable des techniques d'interrogation. L'utilisation d'opérateurs de troncature s'est trouvée être une solution pour trouver toutes les dérivations d'un même mot.

Un autre problème engendré par l'accès au texte intégral est la quasi-impossibilité de trouver des mots composés. En effet, les systèmes d'indexation en texte intégral ne disposant pas de connaissances linguistiques, ne possèdent aucune représentation interne précise de ce qu'est un mot, sauf celle d'une suite de caractères encadrés par des blancs, il est donc difficile de trouver des expressions figées ou des dates, des chiffres ou des acronymes. En réponse à ce problème, il a été créé en plus des opérateurs booléens, des opérateurs de proximité permettant de trouver des mots relativement proches les uns des autres. C'est le cas des guillemets qui permettent de trouver une expression dans son intégralité ou encore l'utilisation d'un opérateur d'adjacence ADJ qui impose que les deux mots soient dans un ordre donné et qu'ils ne soient séparés que par des mots vides.

Enfin, le dernier problème engendré par l'avènement du texte intégral, est que l'on se trouve vite confronté à la localisation des informations pertinentes dans un document. En effet, on est passé d'une recherche de document à une recherche d'information dans les documents. Il est donc apparu comme nécessaire de mettre en évidence les mots de la question dans les documents et de pouvoir passer d'une occurrence à l'autre pour faciliter le repérage des passages pertinents.

7.3. Les approches actuelles

Sous l'approche d'indexation basée sur un vocabulaire contrôlé, le contenu des documents est représenté par un ensemble de descripteurs, l'interrogation consistait alors à spécifier les descripteurs que l'on veut voir figurer dans l'index du document recherché. La question est alors exprimée sous la forme d'une fonction booléenne de descripteurs.

Le résultat de l'application de la fonction booléenne est une partition de la base en deux sous ensembles : l'ensemble des documents jugés pertinents et ceux non pertinents.

Dès le début des années 70, les chercheurs ont proposés d'établir une relation d'ordre de pertinence sur l'ensemble de la base par rapport à la question, plutôt de la diviser en deux ensembles, donc un mode de comparaison pondéré.

Plusieurs modèles statistiques ont été développés pour obtenir cette comparaison pondérée, dans la partie suivante nous présenterons les principaux modèles actuels.

8 Processus et architecture d'un SRI

Un système de recherche d'informations est un système informatique qui permet la recherche d'information dans un fond documentaire, cette recherche consiste à mettre en correspondance une représentation du besoin de l'utilisateur (requête) avec une représentation des contenus des documents au moyen d'une fonction de comparaison.

Les deux fonctions principales d'un système de recherche d'information sont donc :

- La fonction de représentation des contenus des documents,
- la fonction de comparaison (appariement) qui doit établir la correspondance et évaluer la pertinence des documents par rapport à la requête.

Avant que l'utilisateur puisse interagir avec le système de recherche d'information, il doit pouvoir lui fournir une information que le système comprendra. Le système traduit cette information en une requête et balayera par la suite l'ensemble des informations et extrait celles qui répondent le mieux à la question. L'ensemble des réponses est alors présenté à l'utilisateur.

Les cadres éclaircis sur la figure 3 marquent la prolongation du procédé de la recherche d'information suivant la rétroaction de l'utilisateur. Si ce dernier juge que les résultats présentés sont appropriés à son besoin d'information le processus de la recherche s'achève. Si le besoin n'est pas satisfait, l'utilisateur formulera une nouvelle question basée sur l'information obtenue dans les résultats précédents. Ce processus continue jusqu'à ce que le besoin de l'utilisateur soit satisfait.

Dans cette partie, nous reviendrons sur le modèle booléen dans lequel les documents sont représentés par un ensemble de termes non pondérés, les requêtes s'expriment à travers une expression booléenne et l'appariement ne se fait que s'il y a correspondance exacte.

Nous présenterons également le modèle vectoriel, l'un des modèles les plus couramment employés dans la recherche d'information. Les documents et les requêtes sont présentés par des vecteurs dans un espace d'information multidimensionnel et la pertinence d'un document par rapport à

une requête est relative à leurs positions respectives dans cet espace. En y présentent également quelques variantes et améliorations apportées au modèle vectoriel.

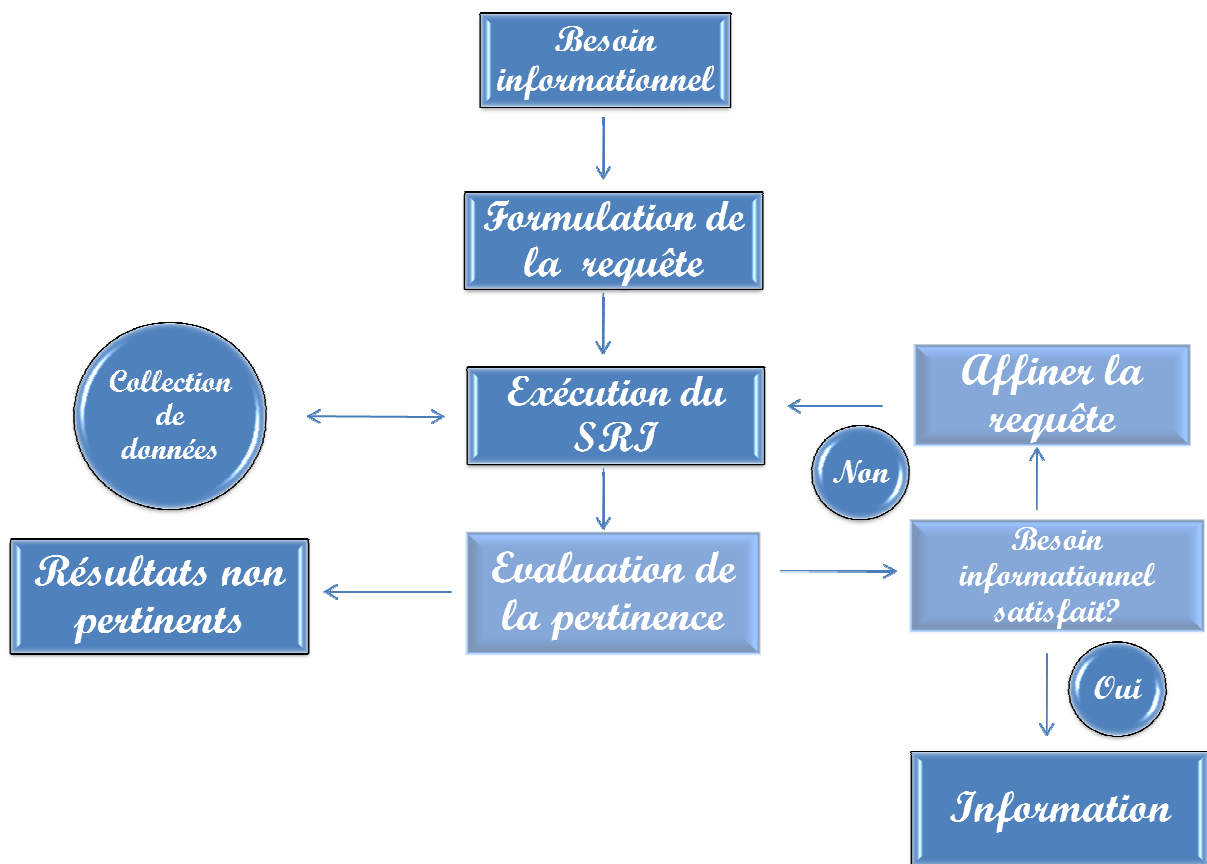


Figure 2. Architecture d'un Système de Recherche d'Information

9 Les Modèles de Recherche d'Information

1. Le modèle Booléen ou ensembliste

Le modèle booléen repose sur la manipulation des mots clés. D'une part, un document est représenté par une conjonction de mots clés, d'autre par une requête (R) est représentée par une expression logique composée de mots connectés par des opérateurs booléens (ET, OU, SAUF).

Le modèle booléen utilise le mode d'appariement exact, il ne restitue que les documents répondant exactement à la requête. Ce modèle est très largement utilisé, aussi bien pour les bases de données bibliographiques que pour les moteurs de recherche.

i. Formulation de la requête

A) Les opérateurs booléens : le modèle booléen tire son nom des opérateurs booléens utilisés pour formuler la requête de l'utilisateur.

- La conjonction (connecteur **ET**) : Indique la présence simultanée de plusieurs termes dans la réponse recherchée, pratique quand on veut limiter et affiner la recherche. Exemple : retrieval **AND** information **AND** internet, Les réponses contiendront obligatoirement les mots retrieval, information et internet. C'est la meilleure façon d'affiner un résultat.
- la disjonction (connecteur **OU**) : Exige qu'au moins un des termes soit présent dans les documents retrouvés, il permet d'élargir la recherche. Ainsi pour la requête retrieval **OR** information **OR** internet. Les réponses contiendront soit retrieval, soit information, soit internet. Cet opérateur n'est pas trop réducteur quant au nombre des réponses. Son avantage consiste à pouvoir utiliser deux synonymes dans une recherche.
- la négation (connecteur **SAUF**) : Permet d'éliminer les documents contenant un terme particulier. Le terme qui suit l'opérateur **SAUF** ne doit pas figurer dans les réponses.

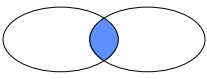
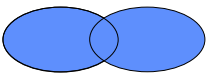
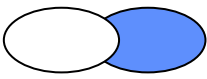
Recherche sur les deux éléments obligatoirement	Recherche sur soit un élément, soit l'autre élément	Recherche avec exclusion du second terme
		
ET	OU	SAUF

Tableau 2. Tableau explicatif des opérateurs logiques

B) La troncature :

Pour prendre en compte les variations morphologiques, la troncature permet de rassembler tous les termes de recherche qui commencent par une chaîne de caractères donné. La troncature est souvent symbolisée par l'astérisque (*).

La troncature est très utile quand on cherche les différentes formes d'un terme (singulier/pluriel, masculin/féminin) ou lorsqu'on recherche les termes ayant une même racine.

C) Les opérateurs de proximité :

Pour traiter le texte intégral, les opérateurs de proximité permettent d'associer plusieurs termes en contexte :

- Soit le champ.
- Soit la phrase.
- Soit le paragraphe.

Parmi les opérateurs de proximité on peut citer les suivants :

- Distance ordonnée entre termes (av, adj, ...)
- Termes présents dans la même phrase (same, phr, ...)
- Termes présents dans le même paragraphe (prg, with, ...)

ii. Les limites du modèle booléen

Le succès du modèle booléen dépend essentiellement du degré de la maîtrise de l'utilisation des opérateurs booléens et la facilité d'exprimer clairement les concepts sur lesquels porte la recherche. De plus, lors de l'interrogation, le poids relatif à ces concepts dans les documents et/ou dans les requêtes n'est pas pris en compte.

Les documents résultats ne sont pas classés et leur nombre est difficile à contrôler, ainsi, les documents importants qui ne sont pas indexés par l'ensemble des termes de la requête ne seront pas sélectionnés, par exemple une requête de type (t₁ ET t₂ ET t₃) donnera comme résultats

uniquement les documents contenant les trois termes, un document contenant un ou deux termes sera rejeté.

iii. Recherche booléenne pondérée

Le modèle booléen est largement répandu. La simplicité de sa mise en œuvre sur le plan informatique et la facilité de son utilisation (fonctions de comparaison) expliquent son succès.

Cependant la performance du modèle booléen est médiocre, c'est que la recherche booléenne est fondée sur le principe que l'utilisateur est toujours capable d'exprimer son besoin informationnel, cela explique qu'on a, pendant très longtemps, réservée cette technique à des spécialistes de la documentation. A l'heure où elle est mise à la disposition du grand public à travers les moteurs de recherche, plusieurs améliorations ont été développées. Certains auteurs ont mis au point une technique dite booléenne pondérée. Des poids sont préalablement attribués aux termes des documents et les termes de la requête peuvent également être pondérés. L'introduction des poids permet un classement des documents, cette pondération tient compte du nombre de documents qui contiennent le terme.

Selon FRIEDER dans [14], le poids W_i^D des termes t_i dans un corpus D peut être calculé de la manière suivante :

- soit N le nombre total des documents du corpus D ;
- soit Mdt_i le nombre de documents qui contiennent le terme t_i ;

$$W_i^D = \frac{\log\left(\frac{N}{Mdt_i}\right)}{\log(N)} \quad 1 \leq Mdt_i \leq N \quad (1)$$

Les termes auront un poids fort, s'ils sont peu fréquents dans le corpus (Mdt_i est alors faible). En particulier, si le terme n'apparaît qu'une seule fois, $Mdt_i=1$ et $W_i^d = 1$.

En revanche, si un terme t_i est présent dans l'ensemble des documents du corpus D , alors $Mdt_i = N$ et son poids est nul.

2. Le modèle vectoriel

Après le modèle booléen, le modèle ayant le plus influencé la recherche d'information est le modèle vectoriel, proposé au début des années 70 par Gérard SALTON [4], [15]. Ce modèle se base sur les propositions suivantes :

- Les requêtes et les documents du corpus sont représentés par des vecteurs de mots clés,
- ces mots clés sont extraits des documents lors de la phase d'indexation,
- la dimension de l'espace vectoriel est égale au cardinal de l'ensemble des mots d'index,
- un poids est attribué à chacun des termes d'indexation d'un même document. Un terme d'indexation n'appartenant pas à un document reçoit un poids nul pour ce document.

i. Vecteurs documents et vecteurs requêtes

Soit un terme t_i le i -ème terme d'indexation, un document D est représenté sous la forme suivante : $D = (wt_{1,D}, wt_{2,D}, \dots, wt_{n,D})$, dans laquelle chaque valeur ($wt_{i,D}$) indique la pondération associée au terme d'indexation (t_i) dans le document D . La requête (Q) est représentée suivant le même formalisme, soit $Q = (wt_{1,Q}, wt_{2,Q}, \dots, wt_{n,Q})$.

$$\vec{D} = \begin{pmatrix} wt_{1,D} \\ wt_{2,D} \\ \vdots \\ wt_{n,D} \end{pmatrix} \quad \vec{Q} = \begin{pmatrix} wt_{1,Q} \\ wt_{2,Q} \\ \vdots \\ wt_{n,Q} \end{pmatrix}$$

Ainsi, pour un corpus qui comporte 20 000 mots, chaque document sera représenté par un vecteur de dimension 20 000. Les éléments valent pour la plupart zéro : seuls ceux qui correspondent aux termes présents dans le document ne sont pas nuls.

La figure 3 illustre une représentation graphique de deux vecteurs documents et un vecteur requête dans un espace associé à trois termes t_1 , t_2 et t_3 .

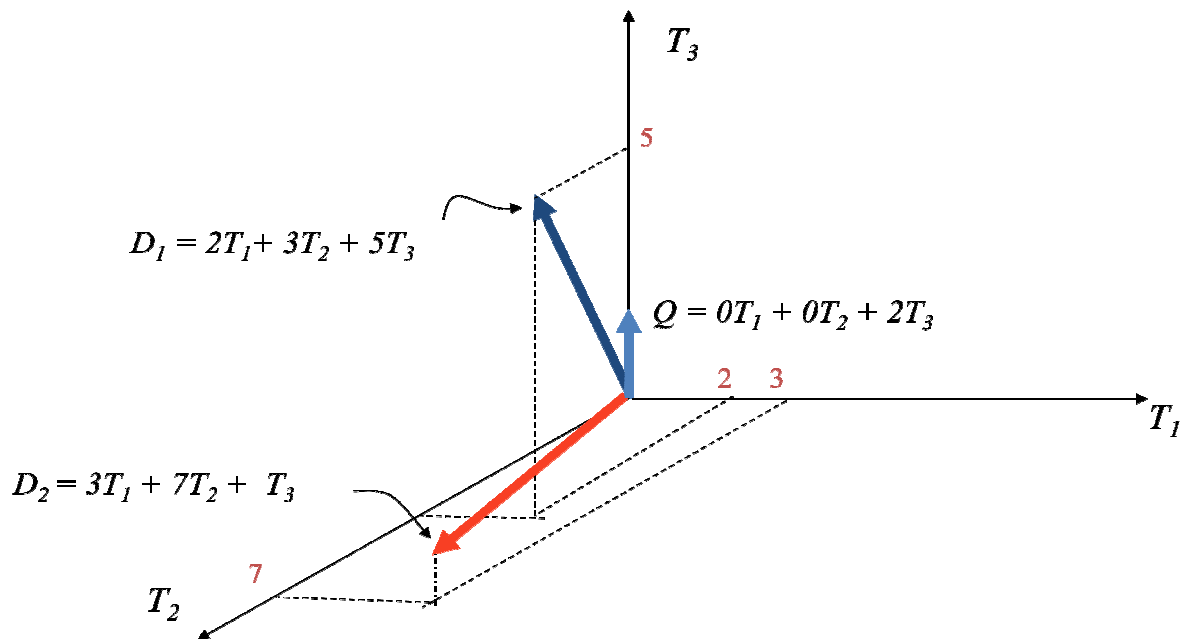


Figure 3. Exemple d'une représentation du modèle vectoriel avec deux documents et une requête.

ii. Les mesures de similarité

Dans le modèle vectoriel, le degré de similarité entre un document D et la requête Q est évalué à l'aide de la corrélation entre les vecteurs \vec{D} et \vec{Q} . Cette corrélation peut être quantifiée, par exemple, par un produit scalaire ou encore le cosinus de l'angle entre les deux vecteurs.

Le produit interne :

$$P = \sum_{k=1}^t (w_{ki} * w_{qk}) \quad (2)$$

Mesure du cosinus :

$$Cos = \frac{\sum_{k=1}^t (w_{ki} * w_{qk})}{\sqrt{\left[\sum_{k=1}^t (w_i^k * w_i^k) * \sum_{k=1}^t (w_q^k * w_q^k) \right]}} \quad (3)$$

Dans cette représentation, on admet que les axes représentant les termes forment une base orthogonale. Les termes sont donc considérés comme indépendants sémantiquement ce qui

constitue l'inconvénient majeur de cette modélisation, lorsqu'on essaye de prendre en compte des termes synonymes, car un même sens peut être décrit par différents termes qui ne seront jamais considérés comme identiques avec cette représentation.

iii. La sélection des termes d'indexation

Les termes d'indexation sont associés aux dimensions de l'espace vectoriel de représentation. Leur choix est donc primordial car ils déterminent la façon dont seront représentés les documents et les requêtes.

Les documents sont décomposés en unités linguistiques, puis les termes d'indexation sont sélectionnés parmi ces unités, la technique la plus simple est de considérer tous les mots et ignorer les mots outils du langage. Des techniques de lemmatisation peuvent également être utilisées pour réduire la taille de l'index en ne conservant que la racine des mots.

Une fois l'ensemble des unités linguistiques défini, il reste à déterminer quelles sont celles qui seront gardées comme termes d'indexation définissant les dimensions de l'espace vectoriel de représentation. Ces unités sont sélectionnées en fonction de leur pouvoir discriminant, c'est-à-dire leur capacité à différencier les documents.

Le critère de sélection des termes d'indexation le plus utilisé est la fréquence en documents, qui représente le nombre de documents contenant l'unité. D'autres techniques plus avancées de sélection des termes d'indexation peuvent être également utilisées, elles seront présentées dans le chapitre 4 de cette thèse.

iv. Les schémas de pondération

La pondération associée aux termes d'indexation dans un document donné prend en général en compte des facteurs de pondération locale, de pondération globale et de normalisation en fonction de la taille du document.

1. La pondération locale :

La pondération locale prend en compte les informations locales du terme qui ne dépendent que du document. Elle correspond en général à une fonction de la fréquence d'occurrence du terme dans

le document (notée tf pour *term frequency*), c'est-à-dire le nombre de fois ou le terme est utilisé dans le document. Les fonctions les plus utilisées sont :

- Facteur tf : correspond à la fréquence d'occurrence du terme dans le document.
- Facteur binaire : il vaut 1 si le terme est présent, 0 s'il ne l'est pas.
- Facteur logarithmique : ce facteur est une fonction logarithmique de la fréquence du terme dans le document valant : $1 + \log (tf)$.
- Facteur augmenté : ce facteur réduit les différences entre valeurs pour les différents poids accordés aux termes du document. Il accorde pour tous les termes présents dans le document une valeur minimale et en accordant aux termes présents plusieurs fois un poids ne dépassant pas une certaine valeur maximale, ce facteur vaut :

$$0.5 + 0.5 * (tf / \max tf) \quad (4)$$

Où $\max tf$ est le plus grand facteur tf des termes du document.

2. La pondération globale :

Contrairement à la pondération locale, la pondération globale prend en compte les informations concernant les termes et dépendant de la totalité de la collection des documents. Un poids plus important doit être donné aux unités linguistiques qui apparaissent moins fréquemment dans la collection : les termes qui sont utilisés dans de nombreux documents sont moins utilisés pour la discrimination que ceux qui apparaissent dans peu de documents. Le facteur de pondération globale est généralement fonction du facteur idf (*inverted document frequency*) valant pour une collection de document D :

$$idf_i = (\log(\frac{N}{df_i}) + 1) \quad (5)$$

df : représente la fréquence en documents du terme considéré.

Du fait de cette double pondération (locale et globale), le modèle vectoriel est souvent référencé sous le nom $tf*idf$.

Cette pondération modélise l'hypothèse qu'un terme est important pour un document donné, s'il apparaît souvent dans ce document et que peu de documents du corpus le contiennent.

$$wt_{i,D} = tf * idf(t_i, D) = tft_{i,D} * idft_i \quad (6)$$

Où $tft_{i,D}$ est le nombre d'occurrence « *term frequency* » du terme d'indexation t_i dans le document D , N est le nombre de documents du corpus. Et $idft_i$ « *inverse document frequency* » est une forme normalisée du nombre de documents qui contiennent t_i .

Cette formule combine l'importance du terme pour un document et son pouvoir discriminant. Ainsi, un terme qui a une valeur ***tf*idf*** élevée doit être à la fois important dans ce document et aussi doit apparaître peu dans les autres documents.

3. La normalisation :

Les pondérations locale et globale sont une bonne approximation de l'importance d'un terme dans un document, mais ne prennent pas en compte un aspect important du document : sa longueur, effectivement la taille d'un document joue un rôle dans le style et le vocabulaire. Les documents qui sont très longs auront tendance à utiliser les mêmes termes de façon répétée. Un document long peut comporter aussi pour des raisons stylistiques un grand nombre de synonymes d'un terme pour éviter les répétitions, ce qui influence le calcul des facteurs de pondération.

4. Combinaison des pondérations :

La donnée de ces trois fonctions de pondération locale, globale et de normalisation forme un schéma de pondération, repéré dans le système SMART [4] par un code de trois lettres. Les schémas de pondération sont exprimés par un double code de trois lettres, ou les trois premières lettres forment le schéma de pondération de la requête et les trois dernières lettres forment le schéma de pondération des documents.

La combinaison des différentes pondérations se fait dans le modèle vectoriel standard, par le produit des différents facteurs de pondération.

Pour un terme t apparaissant dans un document de son poids w est :

$$W(t, d) = w_1(d, t) * w_2(d, t) * w_n(d, t) \quad (7)$$

Avec :

$w_1(d, t)$: facteur de pondération locale

$w_2(d, t)$: facteur de pondération globale

$w_n(d, t)$: facteur de normalisation

Un document est alors représenté par un vecteur $(w_{i1}, w_{i2}, \dots, w_{it})$ où le poids associé au terme t_j est w_{ij} défini précédemment. Dans le cadre du modèle vectoriel, un corpus de documents C sera représenté par une matrice E définie comme suit :

$$E = \begin{pmatrix} D_1 \\ D_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1t} \\ w_{21} & w_{22} & \dots & w_{2t} \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}$$

(8)

v. Prise en compte des dépendances dans modèle vectoriel

On a évoqué précédemment que dans le cadre d'une représentation vectorielle simple, on suppose l'indépendance des termes d'indexation. Ce qui signifie l'orthogonalité des axes de l'espace de représentation. Cette technique pose problème lorsqu'on essaye de prendre en compte des termes synonymes, car un même sens peut être décrit par différents termes qui ne seront jamais considérés comme identiques ni même proche sémantiquement dans cette représentation.

Plusieurs solutions ont été proposées pour tenir compte des dépendances sémantiques dans le cadre du modèle vectoriel. Elles consistent à construire un espace dans lequel les axes ne sont plus orthogonaux en se basant sur les dépendances calculées pour les différents termes.

Nous allons voir dans la suite de ce chapitre deux modèles qui, en se basant sur le modèle vectoriel simple, essaient de palier à ce problème de dépendance en prenant en compte l'aspect sens des documents et des requêtes. Le premier modèle est le modèle *LSI* qui se base sur la décomposition des termes d'indexation, puis nous continuerons avec le modèle *DSIR* qui se base sur l'hypothèse de sémantique distributionnelle et se focalise sur les cooccurrences des termes.

3. Le modèle LSI

Le modèle *Latent Semantic Indexing (LSI)* [5] est un modèle algébrique de recherche d'informations basé sur la décomposition des termes d'indexation à travers l'espace vectoriel. C'est une autre variante du modèle vectoriel qui prend en compte, pour les représentations des documents la structure sémantique des unités linguistiques, qui sont implicite (latent) représentées par leurs dépendances cachées.

Pratiquement, la technique *LSI* construit un espace sémantique à partir d'un corpus de textes et par le biais d'une analyse statistique simple. Cet espace sémantique est construit en prenant en compte le contexte de chaque mot. Comme le sens d'un mot peut être donné par tous les mots qui sont proches de lui dans les différents paragraphes ou il apparaît. Cet espace est réduit et c'est dans cette réduction que réside la puissance de la méthode. En effet, c'est ce processus qui induit les similarités sémantiques entre mots. Cette réduction est au cœur de la méthode car elle extrait les relations sémantiques, un mot peut être considéré comme proche d'un autre mot sans jamais apparaître dans le même texte, de même, deux documents peuvent être proches sans avoir aucun mot en commun. Le cadre d'utilisation de la méthode est fortement dépendant de la collection de documents indexés. Elle obtient de bons résultats sur des masses documentaires très hétérogènes, où elle permet de dégager la sémantique sous-jacente de documents apparemment différents.

LSI utilise une matrice X (terme*documents) qui est composée des vecteurs mots-clés des requêtes et des documents comme pour le modèle vectoriel standard. Ensuite, une décomposition de la matrice X est effectuée. Cette décomposition en valeur singulière (appelée aussi SVD : Singular Value Decomposition) de cette matrice X permet de créer un nouvel espace vectoriel tout en gardant le plus d'information possible.

4. Le modèle DSIR

Le modèle *DSIR -Distributional Semantics Information Retrieval-* [16] ou recherche documentaire à base de sémantique distributionnelle, est un modèle hybride qui intègre des connaissances sémantiques dans la représentation vectorielle des documents en prenant en compte les occurrences des unités linguistiques mais également leurs co-occurrences. L'utilisation des co-occurrences repose sur la notion de sémantique distributionnelle.

L'hypothèse principale de la sémantique distributionnelle est que la sémantique d'un mot est reliée à l'ensemble des contextes dans lesquels il apparaît. Sa démarche est décomposée en trois étapes :

- Définition du contexte d'un mot dans un corpus, donc identifier les mots qui co-occurrent avec ce mot,
- la représentation des mots selon leurs contextes,
- la définition d'une mesure de similarité entre les représentations des mots qui est identifiée avec la mesure de la similarité entre les contextes,

Afin d'identifier les types de relations sémantiques, plusieurs sortes de contextes ont été définis :

- Les contextes positionnels : fenêtres de n mots,
- les contextes syntaxiques : les contextes dépendent de la structure syntaxique de l'unité textuelle,
- les contextes à l'intérieur d'un document : les contextes sont définis selon les unités textuelles à l'intérieur d'un document (paragraphe, section, chapitre).

5. Modèle probabiliste

Le modèle probabiliste se base sur le « principe de classement probabiliste » « Probability Ranking Principle » [17], il doit estimer de manière aussi précise que possible la pertinence d'un document en fonction des informations et données disponibles. Ce modèle considère la recherche d'information comme un espace d'événements possibles. Un événement peut être le jugement de pertinence par l'utilisateur sur un document par rapport à une requête.

Le modèle probabiliste calcul la probabilité de la pertinence d'un document D par rapport à une requête Q . La fonction de similarité de ce modèle tente de séparer les documents pertinents des non pertinents au sein d'une collection. L'idée de base du modèle est de déterminer les probabilités $P(R|D)$ et $P(NR|D)$ pour une requête donnée. Cette probabilité signifie : si on retrouve le document D , quelle est la probabilité qu'on obtienne l'information pertinente et non pertinente.

i. Représentation des documents et des requêtes

Dans le modèle probabiliste seulement l'absence ou la présence des termes est prise en considération dans la représentation des documents et des requêtes. Ainsi, les termes considérés ne sont pas pondérés, mais prennent seulement les valeurs 0 (absent) ou 1 (présent).

Ainsi, tout document (requête) peut être représenté (é) par un vecteur binaire :

$D = (X_1, X_2, \dots, X_n)$, $X_i = 0$ ou 1 indique l'absence ou la présence des termes considérés.

ii. Fonction de correspondance

On suppose deux événements mutuellement exclusifs :

P = document pertinent.

NP = document non pertinent.

On suppose également qu'on a une requête fixe et on tente de calculer $P(P|D)$ et $P(NP|D)$ ainsi, on établit la probabilité qu'un document D soit jugé pertinent par rapport à une requête Q .

$$O(D) = \frac{P(P / Q, D)}{P(NP / Q, D)} \quad (9)$$

Pour calculer cette probabilité, il faut tenir compte des dépendances entre les différents événements, car plusieurs paramètres interviennent :

- Les documents et leurs représentations,
- les requêtes,
- les termes d'indexation, etc.

iii. Prise en compte des dépendances dans le modèle probabiliste

Comme dans la plupart des modèles de recherche d'information, l'hypothèse d'indépendance des termes est présente dans le modèle probabiliste pour permettre de réduire la complexité des calculs de correspondance de ce modèle. En effet, l'hypothèse de dépendance des événements pour la théorie de probabilité implique que les événements sont liés entre eux. Il en est de même pour les termes d'indexation. En effet, on peut penser que la présence d'un terme comme

« ordinateur » dans un document, implique la présence du terme « informatique » avec une très grande probabilité.

6. Le modèle logique

Dans une représentation logique un document est jugé pertinent par rapport à une requête si son contenu sémantique implique logiquement celle-ci. La notion de pertinence est alors considérée comme une inférence logique. Ce modèle permet de formaliser les paramètres intervenant dans un processus de recherche d'informations et de définir correctement la correspondance entre un document et la requête de l'utilisateur. Il permet aussi de définir la formulation automatique d'une requête, ainsi que la mesure de pertinence associée aux réponses données par le système.

i. Représentation des documents et requêtes

Le modèle booléen est un exemple simple qui met en œuvre l'implication logique, un document est modélisé par une proposition logique formée de la conjonction de ses mots clés. On considère ces mots clés comme des propositions atomiques dans les modèles de la logique des propositions. La requête est une expression logique quelconque. L'idée de base de ce modèle est la suivante : étant donné un document D et une requête Q , D est pertinent vis-à-vis de Q si D implique Q , noté mathématiquement par $D \rightarrow Q$.

ii. Fonction de correspondance

Une condition nécessaire exprimée dans la plupart des modèles existants est que le document doit satisfaire « exactement » la requête. En logique, ceci signifie : étant donné un document D , la requête Q doit être totalement satisfaite et l'implication (document \rightarrow requête) doit être évaluée à vrai. C'est-à-dire la probabilité $P(D \rightarrow Q) = 1$. Ceci reste une mesure de correspondance attachée au modèle booléen.

7. L'évaluation des Systèmes de Recherche d'Information

L'évaluation des Systèmes de Recherche d'Information (SRI) porte à la fois sur les résultats d'expérimentation, de discussion sur les méthodologies et les critères d'évaluation et de la pertinence des différentes approches.

Les premières évaluations datent de 1953 selon Lancaster [18], ce domaine a suscité de nombreux travaux de recherches, par exemples les synthèses suivants [19], [20], [21].

Deux pistes de réflexion se distinguent dans ce domaine. La première favorise les études portant sur les usagers avec comme objectif l'analyse et la modélisation des comportements des usagers. Cette piste s'inscrit dans un « paradigme cognitif » qui se fonde sur la satisfaction des usagers et non pas sur la performance des systèmes.

La deuxième piste privilégie l'évaluation quantitative des Systèmes de Recherche d'Information et correspond actuellement au modèle dominant. Il s'agit des campagnes d'évaluation 'TREC' (« Text REtrieval Conference) financées aux Etats-Unis depuis 1993 par la DARPA (*Defence Advanced Research Projects Agency*) et organisées par le NIST (National Institute of Science and Technology).

On peut citer également le programme français 'Amaryllis' [22], [23], les programmes japonais 'IREX' et 'NTCIR', ou encore le projet 'CLEF' financé par la commission européenne.

Cet axe de recherche est celui des chercheurs de l'informatique documentaire il s'inscrit dans une approche appelée « paradigme système ».

Les conférences 'TREC' construisent à ce jour l'initiative la plus importante dans le domaine de l'évaluation de SRI. Dès l'origine, un aspect important de 'TREC' a été la volonté de tester les systèmes sur les corpus de documents de grande taille.

Dans ces différentes méthodologies d'évaluation, la comparaison des réponses d'un système pour une requête avec les réponses idéales permet d'évaluer les deux métriques suivantes :

7.1. Le rappel : calculer l'exhaustivité de la recherche

Le rappel mesure la capacité du système à retrouver tous les documents pertinents pour une requête q , c'est-à-dire le ratio entre le nombre de documents pertinents retrouvés et le nombre total de documents pertinents dans la base. Il est fréquemment exprimé en pourcentage.

$$rappel_q = \frac{\|R_q \cap P_q\|}{\|P_q\|}$$

Où R_q est l'ensemble des documents retrouvés par le système pour la requête q et P_q est l'ensemble des documents pertinents de la collection pour cette requête.

Lorsque l'utilisateur interroge la base il souhaite voir apparaître tous les documents qui pourraient répondre à son besoin d'information. Si cette adéquation entre la requête de l'utilisateur et le nombre de documents présentés est importante alors le taux de rappel est élevé. A l'inverse si de nombreux documents intéressants n'apparaissent pas on parle de silence. Le silence s'oppose au rappel.

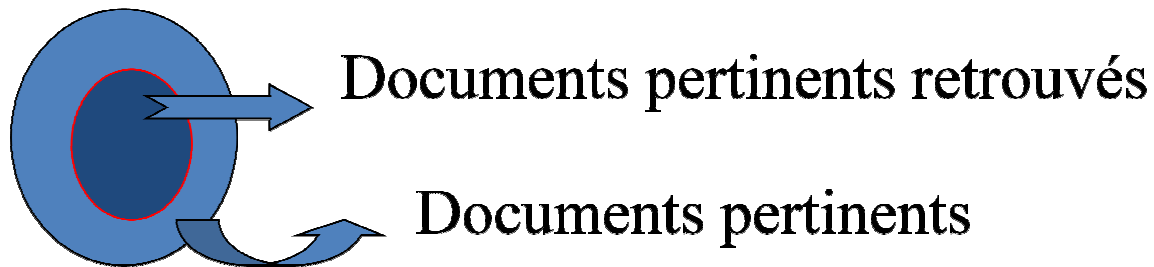


Figure 4. Principe du rappel dans la recherche d'information

7.2. La précision : combien de non pertinent ?

La précision mesure la capacité du système à ne retrouver que les documents pertinents pour une requête q , *c'est-à-dire* le nombre de documents pertinents retrouvés rapporté au nombre de documents (pertinents ou non) retrouvés.

$$precision_q = \frac{\|R_q \cap P_q\|}{\|R_q\|}$$

Si l'utilisateur interroge une base de données, il souhaite ne voir que les documents qui répondent à sa requête. Tous les documents superflus constituent du bruit. La précision s'oppose à ce bruit documentaire. Si elle est élevée, cela signifie que peu de documents inutiles sont présentés par le système.

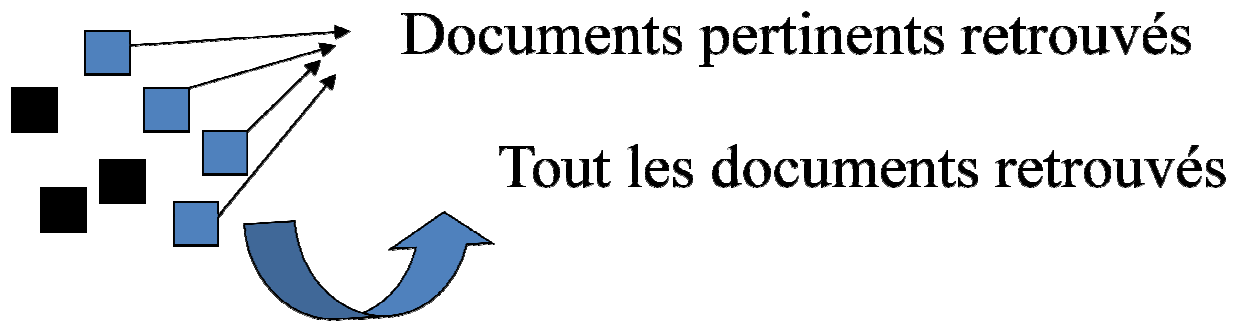


Figure 5. Principe de la précision dans la recherche d'information

7.3. Combiner précision et rappel

Si un système de recherche retrouve tous les documents dans une base de données, il a un rappel de 100% et une précision très basse, donc la précision et le rappel sont inversement liés. Quand le taux de rappel augmente, le taux de précision se détériore. Inversement, quand on précise la requête pour améliorer le taux de la précision le rappel diminue. L'idéal serait que les deux mesures atteignent 100%, ce qui signifie que tous les documents pertinents de la base ont été retrouvés et seulement ceux-ci.

CHAPITRE 2 : Modélisation et visualisation des données textuelles

Les systèmes de recherche d'information présentés dans le premier chapitre se basent sur une approche de modélisation appelée 'sac de mots'. Cette approche ignore les positions d'apparition des mots dans un document dans le calcul de sa pertinence, elle exprime cette dernière, par rapport à une requête, par la plus grande fréquence des mots de la requête dans le document.

Dans ce deuxième chapitre nous allons présenter cette approche de modélisation ainsi que les différents outils informatiques de traitement et de visualisation classiques de données textuelles basée sur cette approche. Nous présenterons également un nouveau concept dans la visualisation basé sur la modélisation spectrale.

Introduction

Ces dernières années ont vu d'importants changements à la fois dans la nature des collections de données disponibles et également dans les besoins des utilisateurs. Les corpus sont devenus beaucoup plus volumineux et sont souvent composés de données hétérogènes aussi bien dans leurs formes que dans leurs contenus.

De nouveaux standards de représentation des données textuelles se sont développés en liaison avec le Web en particulier avec la proposition du langage XHTML (World Wide Web Consortium, 1998). Paradoxalement, la recherche d'information apporte peu d'outils pour le traitement de cette quantité de données, les systèmes de recherche d'information actuelles ont été conçus pour des représentations de données textuelles plates et homogènes, ils ne sont pas adaptés au traitement simultané de la structure et du contenu des documents.

La nécessité actuelle donc est de lier dans un système de recherche d'information, la structure logique et le contenu d'un document pour le calcul de sa pertinence. Un texte sera ainsi représenté par une information plus complète que son simple contenu textuel, il sera accessible à différents niveaux de description.

L'accès simultané à la structure des documents et à leur contenu permet d'envisager de nouveaux modes d'exploitation de l'information textuelle. Celle-ci nécessite la création de nouveaux outils et de nouveaux modèles capables d'exploiter de telles données.

1. Modèles de représentation des données textuelles

1.1 Approche 'sac de mots'

Les techniques classiques représentent un document par un ensemble de mots, cette approche de représentation ignore la position relative des mots, est souvent surnommée modèle 'sac de mots ou bag of Words'. Cette approche suppose que les mots utilisés dans un document suffisent pour donner le sens de son contenu général. Elle suppose également que la probabilité d'utiliser un mot est indépendante de sa position et de celle des autres mots. Sous cette approche le texte est considéré comme une large masse d'informations non structurée.

L'approche 'sac de mots' est l'une des premières méthodes de représentation des textes utilisée dans les systèmes de recherche d'information actuels vue sa simplicité. Cependant, plusieurs de ces systèmes amplifient cette approche par la suppression des variations grammaticales et ignorent les mots vides dans la représentation des textes. Un des premiers systèmes qui se fonde sur cette approche est la première version de SMART [4]. D'autres systèmes récents améliorent cette approche de 'sac de mots' par l'utilisation de l'information sur les cooccurrences des mots dans le but d'amplifier la vue sémantique du mot, comme le système *DSIR* ou *LSI* étudiés dans le chapitre précédent.

La simplicité et l'efficacité de cette approche sont les raisons principales de sa popularité, mais parfois elle est beaucoup trop simple pour certaines tâches de recherche et d'interrogation. Si un texte est considéré comme une collection désordonnée de mots, toutes les informations concernant sa structure est détruites. Les systèmes utilisant cette approche ne peuvent pas rechercher des segments pertinents dans un document, ni d'analyser le flot de la discussion dans le texte. Ces capacités sont pourtant importantes pour des systèmes de segmentation des textes, des systèmes de visualisation, etc. ...

1.1.1 Identification des termes d'indexation

L'approche 'sac de mots' utilise des méthodes classiques afin de retrouver les termes d'indexation valides parmi les termes candidats, généralement ces méthodes sont basées sur des calculs statistiques simples. Nous allons exposer ci-après les méthodes les plus utilisées.

1-Indice T_f :

La première mesure est la fréquence d'un terme T_j relatif à un document quelconque D_i pris dans un corpus de référence C .

L'idée est de supposer que l'importance d'un terme va de paire avec son nombre d'occurrences dans D_i par rapport à son utilisation dans le reste du corpus. L'indice T_f mesure cette fréquence :

$$T_f(D_i, T_j) = \frac{N(D_i, T_j)}{N(T_j)} \quad (10)$$

Où : $N(D_i, T_j)$ est le nombre d'occurrence de T_j dans D_i et $N(T_j)$ est le nombre d'occurrence de T_j dans le corpus C .

2-Indice I_{df} :

Cet indice est nommé I_{df} pour '*Inverse Document Frequency*'. Il caractérise la répartition d'un terme dans le corpus, en partant du principe que l'importance d'un terme est inversement proportionnelle au nombre de documents du corpus dans lequel il apparaît. Cet indice permet de retrouver les termes caractéristiques d'un corpus donné.

$$I_{df}(T_j) = \log \frac{N}{n_i(T_j)} \quad (11)$$

Où: N est le nombre de documents dans le corpus C et $n_t(T_j)$ le nombre de documents distincts de C contenant T_j .

3-Indice $T_f * I_{df}$:

Cet indice est le juste compromis entre termes fréquents et termes bien répartis dans le corpus. L'expérience montre que les termes sélectionnés uniquement à partir des indices T_f et I_{df} ne sont pas nécessairement de bons descripteurs, par contre ils servent de référence pour pouvoir comparer avec d'autres méthodes de pondération.

Un moyen simple de retrouver des termes fréquents mais présents dans peu de documents est finalement de combiner les deux critères de pondération qui sont T_f et I_{df} , on obtient alors l'indice $T_f * I_{df}$:

$$T_f * I_{df} = \frac{N(D_i, T_j)}{N(T_j)} \log \frac{N}{n_i(T_j)} \quad (12)$$

1.1.2 Méthodes d'analyse de l'information

Les modèles de représentation et de visualisation de l'information utilisent différents outils d'analyse statistique. Parmi ces méthodes d'analyse, on distingue : les méthodes uni ou bidimensionnelles comme le calcul de fréquence ou de cooccurrences et les méthodes multidimensionnelles, qu'il faut également distinguer en deux grandes classes : les méthodes factorielles et les méthodes classificatoires.

Nous allons présenter certaines de ces méthodes, à savoir, l'analyse en composante principale, le positionnement multidimensionnel (MDS) [93] et les cartes topologiques de Kohonen. (Self-

Organizing Maps) [94] Tous ces outils et autres sont largement utilisés dans la réduction de dimension des données facilitant ainsi leur présentation (carto) graphique.

Dans ce qui suit nous allons présenter certains de ces outils et leur utilisation dans le traitement et la visualisation des données textuelles.

1.1.2.1 Les méthodes factorielles

Les méthodes factorielles s'appliquent sur des tableaux de données numériques volumineux et permettent de représenter graphiquement ces données de façon synthétique.

Le principe des méthodes factorielles est le suivant : elles réalisent une projection des données, avec un minimum de déformation selon une métrique choisie, du nuage de points dans un espace de dimensions réduites (plan par exemple). L'analyse en composantes principales (ACP) et les analyses des correspondances simples ou multiples (AFC: Analyse Factorielle des Correspondances) sont les méthodes factorielles les plus utilisées.

a)- Analyse en composante principale

L'analyse en Composante Principale (ACP) est une technique largement répandue pour opérer une réduction dimensionnelle sur les données d'origine.

Étant donnée une matrice (documents-termes) de dimension $(n*m)$. L'ACP utilise les K premiers vecteurs propres (axes principaux d'inertie) de la matrice de covariance $(n*n)$ pour projeter les données initiales dans un espace de dimension K réduite et qui retient le plus d'information possible.

L'efficacité de l'ACP réside dans sa capacité de réduire le bruit lié aux mots vides, la redondance et l'ambiguïté. L'ACP convient à l'analyse des documents mais les dimensions de l'espace réduit manquent de signification sémantique.

b)- L'Analyse Factorielle des Correspondances

Cette méthode est particulièrement populaire dans la communauté de l'analyse de données textuelle. Elle permet par exemple de révéler une certaine proximité entre les mots à l'intérieur d'un corpus de documents. L'analyse factorielle des correspondances a été développée par BENZECRI dans le début des années 1980 [25].

Sur un tableau de n données sur p dimensions, l'AFC va déterminer les K premiers axes d'un système d'axes orthogonaux résumant le plus la variance du nuage. Les structures résultantes (cartes par exemple) permettent d'interpréter les données. L'idée fondamentale est d'éliminer la redondance dans les données originales en essayant de « résumer » les variables de départ en un nombre plus faible.

1.1.2.2 *Le positionnement multidimensionnel (MDS)*

Le positionnement multidimensionnel (MDS) est un ensemble de techniques mathématiques qui permet de découvrir les structures cachées dans les données. Les applications possibles incluent le traitement des données dans la psychologie, la sociologie, l'économie et la visualisation des textes de documents.

Le (MDS) est une méthode d'analyse d'une matrice de proximité (similarité ou dissimilarité) établie sur un ensemble de données (documents dans le cas des données textuelles). Le (MDS) a pour objectif de modéliser les proximités entre les données de façon à pouvoir les représenter le plus fidèlement possible dans un espace de faible dimension (généralement 2 dimensions). Cette proximité peut être obtenue de différentes manières, par exemple, en calculant le coefficient de corrélation ou la distance euclidienne dans une représentation vectorielle des documents.

1.1.2.3 *Les cartes auto-organisantes de Kohonen SOM (Self-Organizing Maps)*

Le réseau de *Kohonen* est à la fois une méthode de classification et de visualisation sous forme de carte hypertextuelle. Le réseau se répartit en deux couches : une couche d'entrée formée par les vecteurs et une couche de sortie qui est la couche de *Kohone* et une distance euclidienne est utilisée comme fonction de transfert entre les deux couches.

Les nœuds de la couche de *Kohonen* (les documents en l'occurrence) sont répartis uniformément sur une grille en deux dimensions qui sera représentative d'une carte de navigation sur les documents. La carte positionne les classes sur chaque nœud et associe chaque classe à un groupe de documents. La carte est graphique, elle permet une navigation documentaire à partir des champs sémantiques identifiés.

1.1.3 Modèles de visualisation : la cartographie des données textuelles

La quantité de données textuelles disponibles sur Internet augmente de façon exponentielle. De nombreux outils existant aujourd'hui permettent à la fois aux utilisateurs d'exprimer l'objectif réel de leur besoin et de visualiser et d'interagir avec les résultats retournés, ceci afin d'aider ces utilisateurs à mieux appréhender le contenu d'importants ensembles de documents.

La cartographie de données textuelles est l'un de ces nouveaux moyens d'accès à l'information par la visualisation. Nous allons tout d'abord présenter dans cette partie les motivations qui ont entraîné la création de tels outils cartographiques, nous ferons ensuite un état de l'art des outils de visualisation et de cartographie existant, tout en présentant les principes de fonctionnement de chacun.

1.1.3.1 *Les limites des outils traditionnels de recherche d'information*

La quantité et la grande diversité (diversité en forme, qu'en genre ou en thématique) des documents disponibles de nos jours sur Internet ne permettent plus aux moteurs de recherche traditionnels de proposer aux utilisateurs un service très efficace. En effet, de tels outils proposent en réponse à une requête de l'utilisateur (se limitant à quelques mots tentant de décrire l'objet de la recherche), une liste de liens vers des pages Web. Ces liens sont triés selon une certaine pertinence des pages qui désigne par rapport à la requête de l'utilisateur. Les listes de liens ainsi retournées sont certes très simples à interpréter par l'utilisateur (les liens situés en tête de liste sont plus pertinents que ceux situés en queue de liste), mais contiennent souvent un très grand nombre de liens (souvent plusieurs dizaines, voire des centaines ou des milliers), alors que l'utilisateur se contentera la plupart du temps de ne visiter que les tout premiers.

Le classement des liens orchestré par les moteurs de recherche est souvent assez complexe et critiquable. Les techniques utilisées sont propres à chaque moteur et se basent généralement sur la présence de graphiques et d'URLs dans les pages parcourues par le moteur.

Deux critiques peuvent alors être adressées à ces outils traditionnels de recherche. Tout d'abord de tels outils mettent en œuvre des traitements, certes très élaborés, mais généralistes et retournant aux utilisateurs des résultats n'étant pas toujours en rapport avec l'objet réel de leur recherche.

La seconde critique est liée aux listes de liens retournés à l'utilisateur en réponse à sa recherche. Ces listes ne permettent pas à l'utilisateur d'avoir une vue globale des résultats retournés, ni même d'observer précisément des proximités entre pages Web (par exemple, des similarités thématiques entre certaines pages).

Cette notion de proximité de contenu est particulièrement intéressante dans le cadre d'outils de fouille de textes : le simple regroupement de plusieurs liens désignant des pages aux contenus similaires en un groupe de liens pointant vers une information particulière, peut être très utile afin de permettre à l'utilisateur d'avoir une idée sur la grande quantité d'informations qui lui est retournée.

La proposition des outils permettant à la fois de prendre en considération le besoin réel de l'utilisateur et proposant des représentations visuelles interactives des résultats obtenus est donc nécessaire dans des outils de fouille de texte mais également dans des outils de veille documentaire.

1.1.3.2 *Etat de l'art : Les travaux de recherche exploitant la visualisation de données textuelles*

Depuis plusieurs années, un grand nombre de chercheurs s'intéressent à la problématique de l'accès au contenu de grands ensembles de documents. Pour cela, certains ont mis en œuvre des techniques de visualisation de l'information. Dans cette partie, nous nous intéressons aux principales méthodes de visualisation.

Dans le domaine de fouille de textes, SALTON propose dans [26] de projeter sur le périmètre d'un cercle les pages obtenues en réponse à une requête. Nous pouvons ainsi obtenir la représentation suivante :

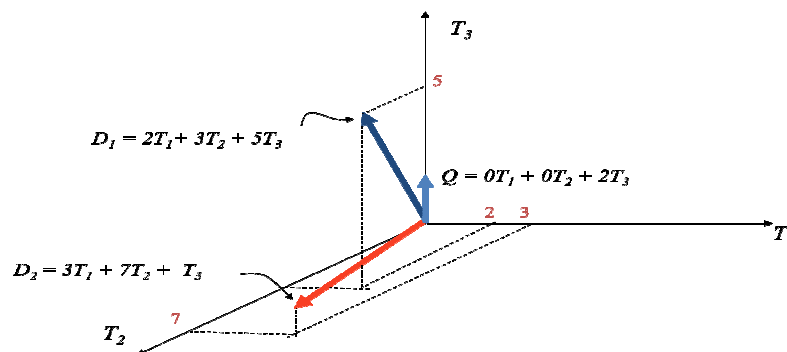


Figure 6. Tour de Salton

Sur cette figure, la requête de l'utilisateur est désignée par la lettre Q et les documents retournés en réponse à cette requête sont désignés par les lettres D1 et D2. La distance angulaire entre Q et D1 est la plus petite. Le document D1 est le plus approprié pour répondre à la requête D.

Toujours dans ce domaine, HEARST propose dans [27] en réponse à une recherche un ensemble de rectangles correspondant chacun à un document jugé pertinent par le système. Dans ces rectangles quadrillés, chaque ligne correspond à un mot-clé de la requête et chaque colonne est grisée en fonction de la fréquence du mot-clé au sein du segment de document qui lui est associé. (Figure 7).

Dans des tâches de parcours rapide d'un ensemble de documents, d'autres techniques ont été proposées. Certaines de ces techniques présentent un ensemble de documents sous forme de hiérarchies en 3 dimensions (tel le Cone Trees (Figure 8) [28]) ou d'arbres hyperboliques (l'Hyperbolic Tree [29]).

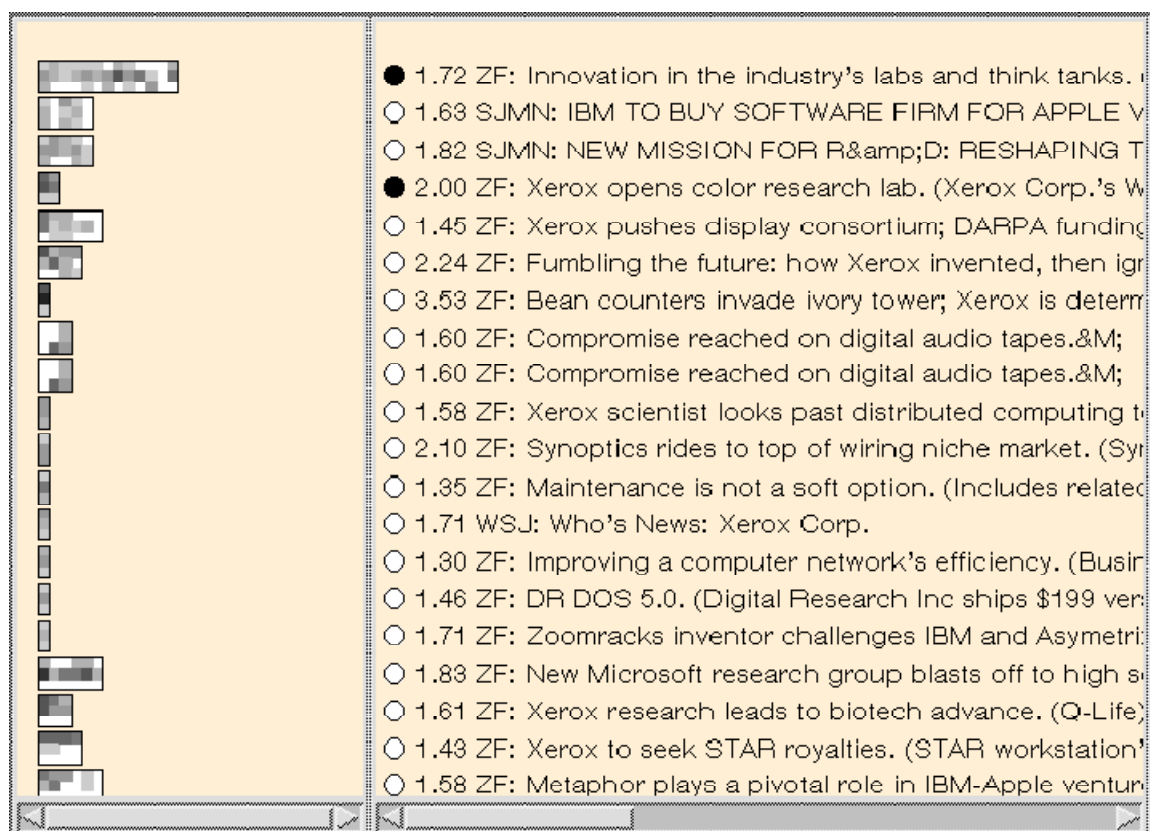


Figure 7. Les « TilesBars » de Hearst

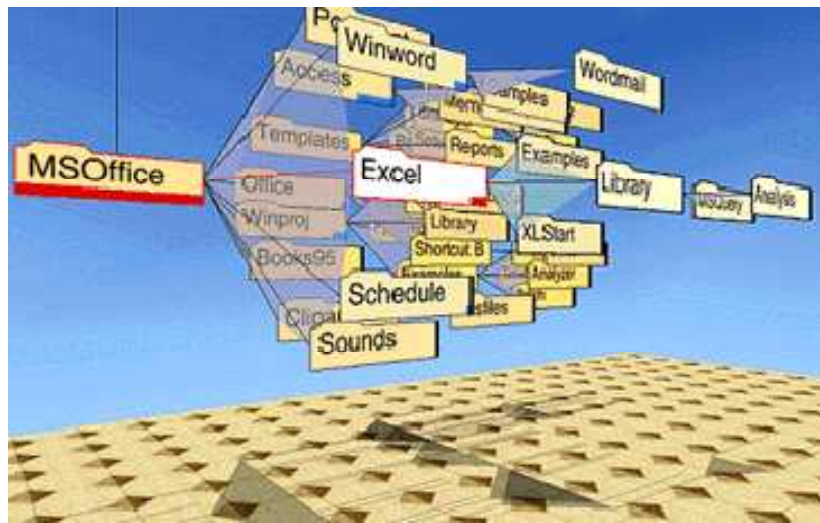


Figure 8. Un exemple de Cone Tree

Une technique de visualisation particulière, appelée cartographie, est également exploitée dans certains travaux. Une carte d'un ensemble de documents met alors en évidence des proximités et des liens entre entités textuelles (comme pas exemple des mots, des documents, des auteurs, ...) au sein de cet ensemble.

Dans une tache de veille documentaire, Abdenour MOKRANE propose dans [30] d'utiliser une technique de cartographie afin de visualiser les liens entre les principaux termes présents dans un ensemble de dépêches d'agences de presse (figure 9).

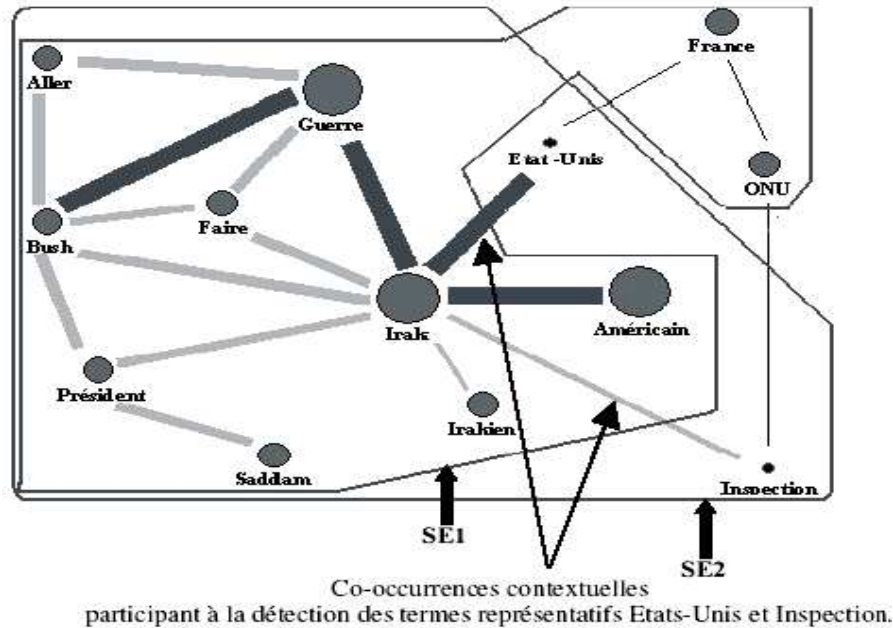


Figure 9. La cartographie d'un corpus de dépêches d'agences de presses selon MOKRANE [30]

Devant l'intérêt croissant des techniques de visualisation, de nombreux logiciels d'analyse de données textuelles proposent des méthodes de visualisation de résultats d'analyse d'un corpus des données. Parmi ces logiciels, nous pouvons citer SAMPLER et Tétralogie.

Dans une tâche de veille concurrentielle, SAMPLER [31], [32] permet de visualiser l'ensemble du lexique sous la forme d'une représentation graphique (réseaux de clusters) après une étape de clustérisation caractérisée par le regroupement des mots constitutifs du lexique par familles homogènes. A l'intérieur d'un cluster, les mots sont reliés entre eux par des liens plus ou moins forts calculés selon les cooccurrences relatives des mots dans les textes. Il est possible, en sélectionnant un mot, de retourner aux documents qui lui sont liés (figure 10).

SAMPLER peut être associé au développement de plate-forme de veille stratégique permettant l'élaboration, le traitement et la diffusion de l'information ou à la réingénierie de systèmes documentaires (constitution ou rénovation de référentiels terminologiques).

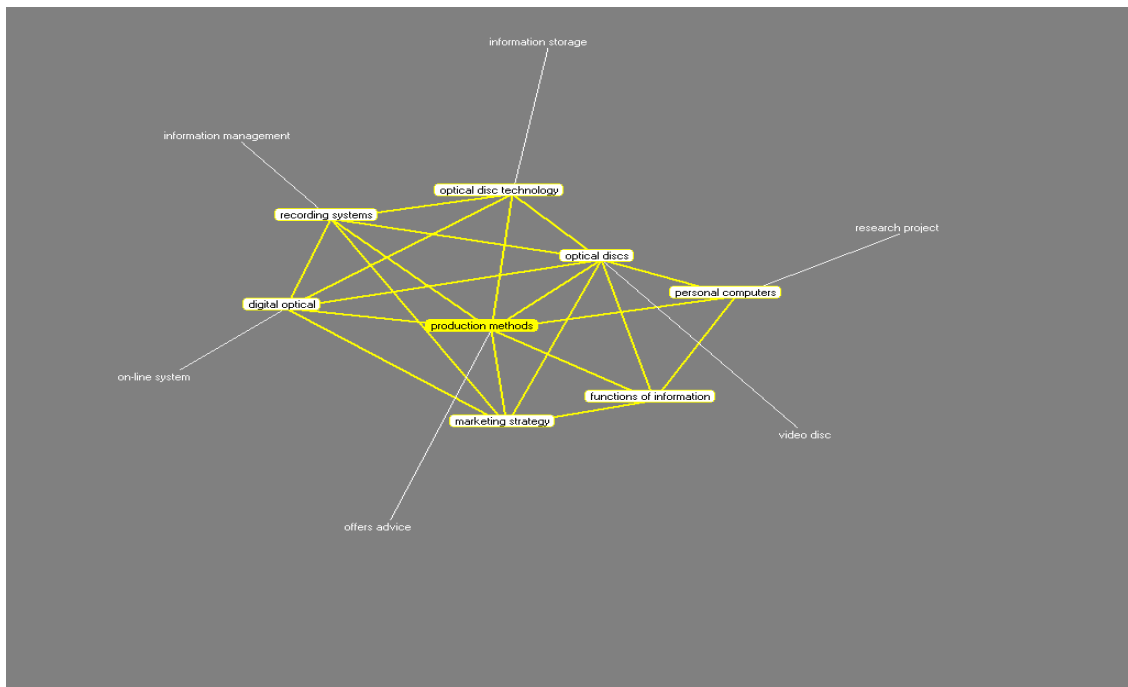


Figure 10. Un cluster extrait d'un corpus « armement nucléaire » réalisé sous SAPMLER

Tétralogie [33], [34] permet par l'intermédiaire de méthodes statistiques d'analyse de données évolutives visualisées en quatre dimensions, de mettre en évidence: l'identité des acteurs et leurs relations, leurs lieux d'action, l'émergence et l'évolution des sujets et des concepts, les éléments stratégiques de propriété industrielle (brevets), les domaines virtuellement porteurs, que lire et où publier, avec qui collaborer, etc....

Tétralogie est un des éléments essentiels de la station bibliométrique 'ATLAS' élaborée conjointement grâce aux aides du CEDOCAR¹ et du SGDN² et qui fait intervenir de nombreux partenaires nationaux (INIST³, CRRM⁴, IRIT⁵, différents ministères) afin de proposer sur un même support l'ensemble des méthodes opérationnelles à l'heure actuelle dans le domaine. L'outil Tétralogie est alimenté par des banques de données bibliographiques, textuelles ou factuelles (en ligne ou sur CD/Rom).

¹ Centre d'Information et de Documentation de l'Armement

² Le secrétariat général de la défense nationale

³ Institut de l'Information Scientifique et Technique

⁴ Centre Recherche Rétrospective de Marseille

⁵ Institut de Recherche en Informatique de Toulouse

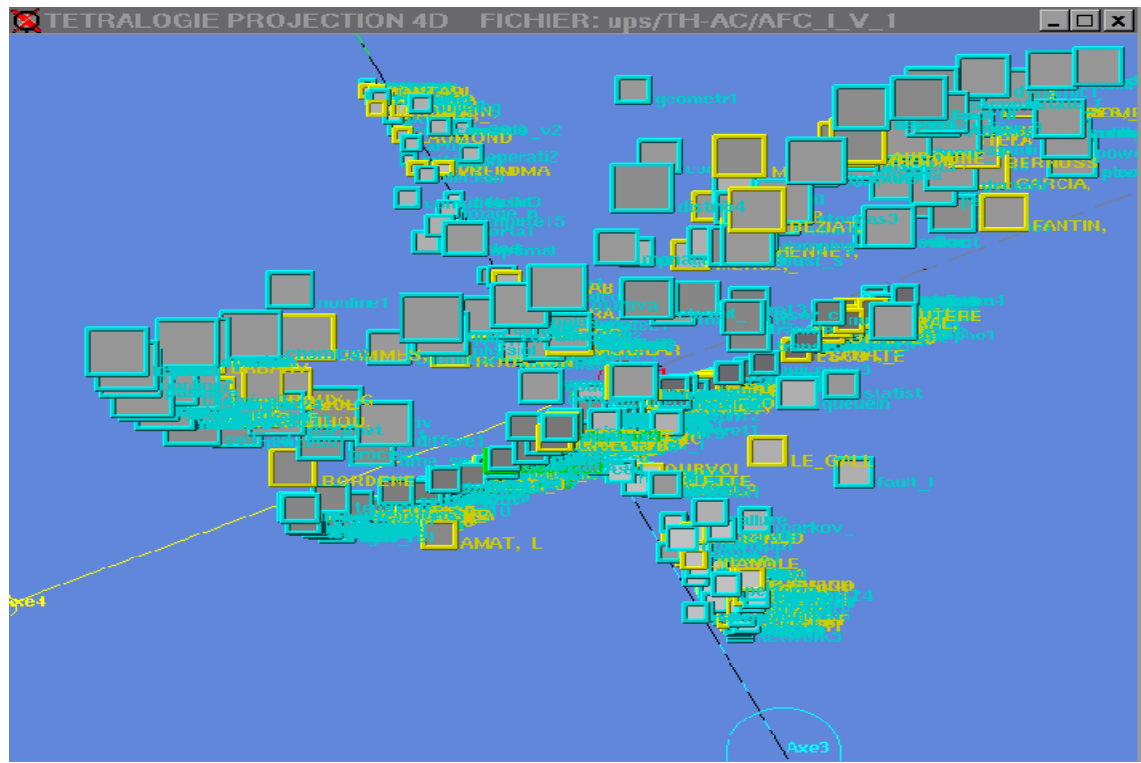


Figure 11. Analyse multidimensionnelle selon Tétralogie

1.2 Approche de document structuré

Un document n'est pas seulement un 'sac de mots', mais un ensemble structuré de termes et d'expressions, un ensemble qui permet de communiquer des informations sur un ou plusieurs sujets à la fois. Cette richesse et cette complexité doivent être prises en compte dans la conception des systèmes de recherche d'information.

Baeza_Yates [35] proposent une typologie des documents textuelles basée sur la structure qui comporte trois éléments : structures fixes, structures hypertextes et structures hiérarchiques.

1- Les structures larges fixes représentent le document sous forme d'une liste d'unités indépendantes dont chacune est un sac de mots. Les courriels par exemple se composent des champs suivants : l'adresse du destinataire, le sujet et le message. Les articles techniques sont représentés par un titre suivi d'une liste de paragraphes. Ces larges structures ont l'avantage de garder la simplicité de l'approche de sac de mots et elles permettent également de rechercher les unités pertinentes dans un document.

Salton et Buckley dans la version récente de SMART qui combine l'approche sac de mots et celle d'unités structurées [36] mesurent la similarité entre un document et une requête en deux temps, premièrement avec tout le document (sac de mots) et en second lieu avec les paragraphes du document indépendamment. Si un ou plusieurs paragraphes sont pertinents par rapport à la requête, ils seront retournés par le système, sinon tout le texte peut être retourné s'il est jugé pertinent. D'après leur expérience cette combinaison d'approches donne des résultats plus pertinents par rapport à l'approche sac de mots.

2- A la différence des structures larges, les structures hypertextuelles représentent un document par un ensemble d'unités reliées entre elles. Les pages HTML est un exemple idéal de ces structures.

Cette structure est adoptée par des moteurs de recherches sur Internet mais seulement quelques moteurs l'utilisent pour mesurer la pertinence de leurs résultats, par exemple, la mesure PageRank de Google mesure l'importance d'une page web, en calculant le nombre de pages pointant vers cette page.

Google est un bon exemple de la façon dont la structure hypertextuelle (liens entre les documents) peut fournir une ressource riche pour les systèmes de recherche d'information. Cependant, les structures de ce type sont habituellement chères à traiter.

3-Les structures hiérarchiques représentent un compromis entre la complexité de l'aspect hypertextuelle et la simplicité du texte libre. Les livres sont un exemple de structure arborescente, contenant des chapitres, qui contiennent des sections, qui peuvent éventuellement contenir des sous sections.

Beaucoup de système de recherche d'information permettent à des utilisateurs d'indiquer une sous-section particulière dans un document pour préciser leur besoin de recherche. Cette sous-section est appelée champ, par exemple, recherche des mots clés dans titre, le résumé ou dans le corps d'un document.

L'approche de document structuré permet de chercher et de trouver des segments pertinents dans un document, mais elle apporte peu par rapport à l'approche de sac de mots puisque chaque segment est encore représenté comme un ensemble non ordonné de mots et l'information concernant le contexte local du mot dans un document reste absente.

1.3 Le contexte local d'un mot dans un texte

Le contexte local d'un mot est l'ensemble des mots qui l'entoure dans un texte, ce contexte immédiat aide généralement le lecteur à limiter une signification possible du mot, ainsi que les sujets présents. Par exemple, le mot '**avocat**' peut porter plusieurs significations, y compris la profession d'avocat et le fruit d'avocat, mais l'expression de '**tribunal**' immédiatement à côté du mot avocat clarifie et désambiguïse le contenu.

Plusieurs systèmes définissent un contexte local comme l'ensemble des mots qui précède un mot choisis, d'autres le définissent comme l'ensemble des n mots qui précèdent et suivent le mot. Reconnaisant l'importance du contexte local plusieurs moteurs de recherche tels que Altavista et Google permettent à l'utilisateur d'effectuer la recherche avec une certaine proximité entre les mots.

Le contexte local d'un mot permet de maintenir des informations partielles sur l'ordre des mots dans le texte, il est utilisé dans certaine approche de proximité. Cependant cette approche ne fournit pas une vue globale du texte mais seulement une vue partielle, ainsi elle est insuffisante pour représenter et décrire le contenu d'un document.

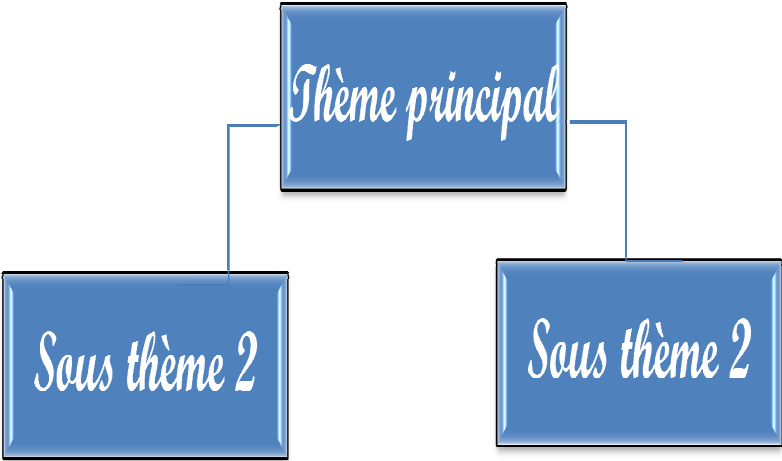
1.4 Les thèmes dans un document

Dans un texte l'auteur peut développer un ou plusieurs sujets ou idées (appelées thèmes), sans aucune supposition sur leurs organisations. L'avantage de cette approche est sa flexibilité dans le sens ou les thèmes peuvent être interrompus et réintroduits plusieurs fois dans le texte. Cependant, cette flexibilité provoque la perte de relations hiérarchiques entre le thème principal et ses sous thèmes. Une autre approche plus générale présente le texte avec un ensemble de thèmes principaux discutés parallèlement et les sous thèmes se présentent linéairement consécutifs.

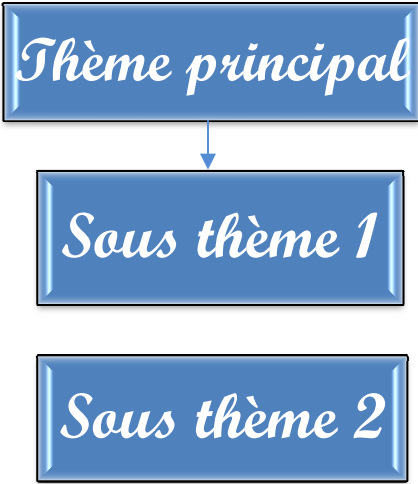
Définition du thème

Jean-Marie SCHAEFFER dans [37] souligne la difficulté de définir la notion de thème. Les auteurs citent tout de même deux définitions du thème. La première, considère le thème comme « *un principe concret d'organisation, un schéma (...) autour duquel aurait tendance à se constituer et à se déployer un monde* ». La seconde définit le thème tel « *un signifie individuel, implicite et concret* ».

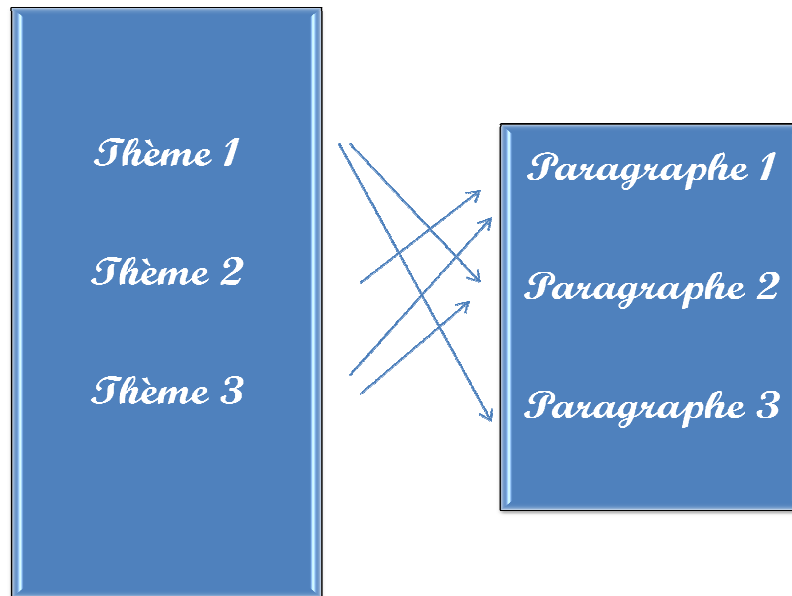
François RASTIER dans [38] considère les définitions précédentes comme étant plutôt d'ordre philosophique et que d'ordre linguistique, il propose de définir le thème comme « *une structure stable de traits sémantiques (ou sèmes), récurrente dans un corpus et susceptibles de lexicalisations divers* ». Dans le tableau 2 on montre quelques possibilités d'organisation thématique dans un texte.



(a)



(b)



(c)

Tableau 3. Différentes organisations des thématiques dans un texte

1.5 Visualisation multidimensionnelle spectrale

Miller propose dans [2] une approche de visualisation et d'exploration des textes non structurés. Cette technologie appelée 'Topic-O-Graphy', applique une Transformée en ondelettes sur un signal construit à partir des mots d'un texte. La propriété de Multi résolution des Ondelettes permet d'analyser les caractéristiques du flot narratif dans le texte tel que la détection des points de rupture et de changement thématiques dans un document.

'Topic-O-Graphy' permet à l'utilisateur de visualiser rapidement les thèmes principaux d'un document électronique selon divers modes.

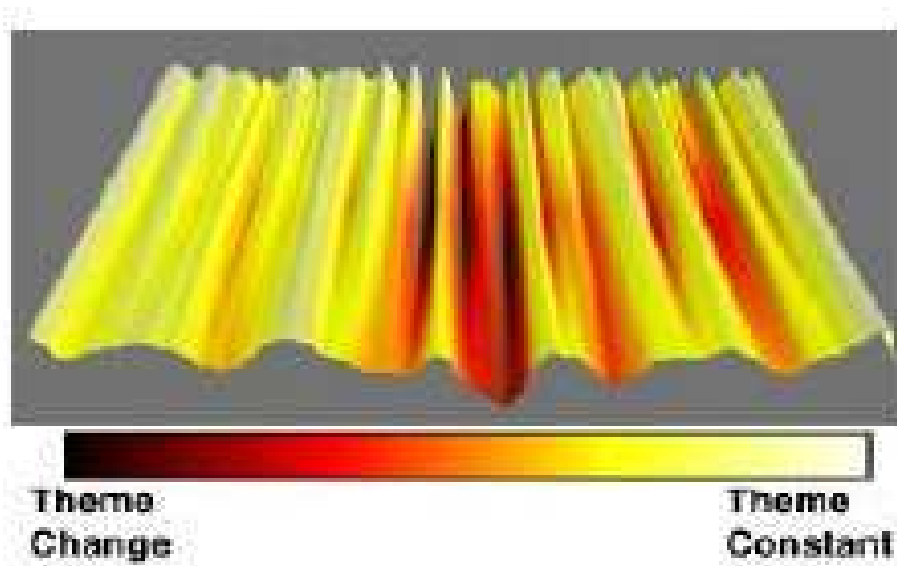
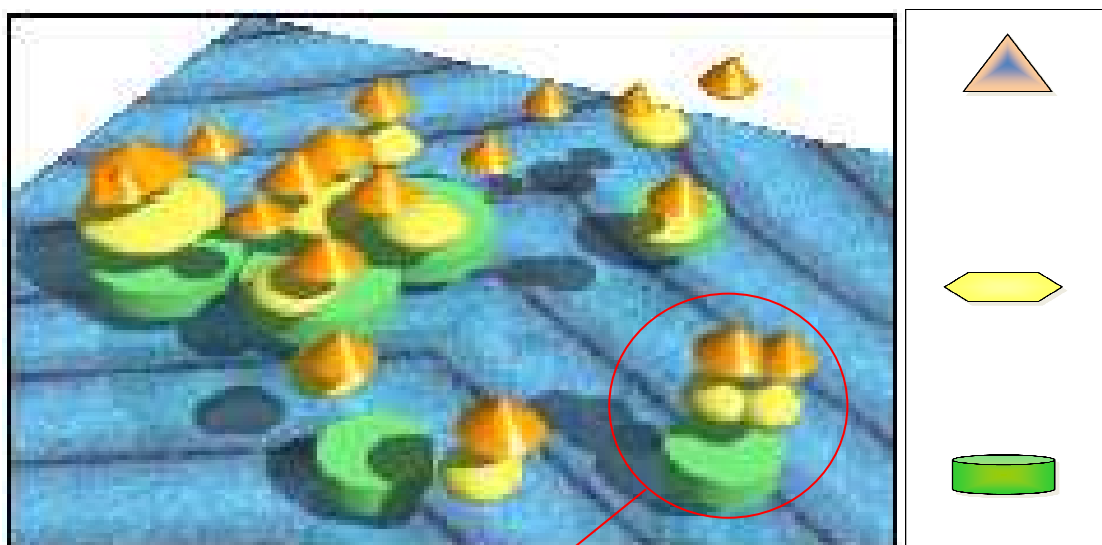


Figure 12. Prototype de visualisation Topic-O-Graphy: forme 'Wave':



Thèmes proches ou communs
aux 3 échelles de résolution

Représentation de 3
échelles de
résolution

Figure 13. Visualisation thématique multidimensionnelle Topic- O- Graphy

Mise en couleur des thèmes du texte (échelle de résolution = MRL)

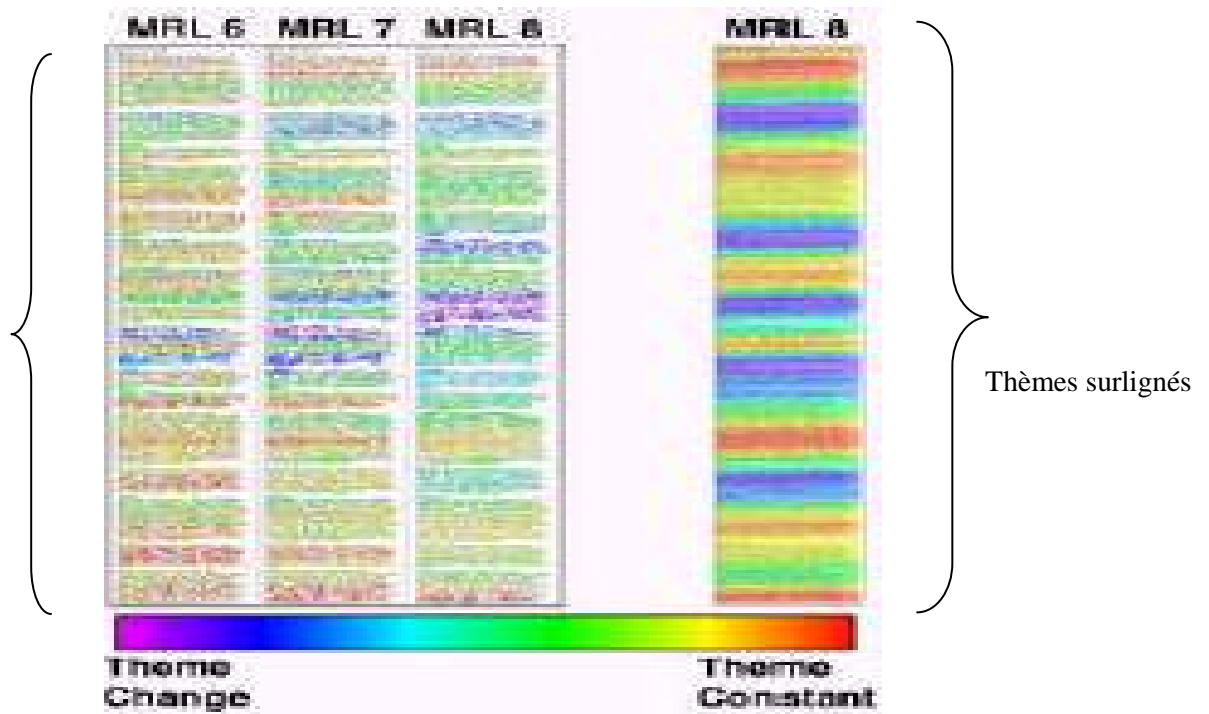


Figure 14. Visualisation de texte par coloriage

Le processus de modélisation décrit par Miller [39] commence par l'identification des mots porteurs de sens dans un texte. Pour cela, les auteurs utilisent les mesures d'identification proposées par Bookstein [40] : Les mots ayant un poids plus élevé qu'un seuil fixé sont considérés comme *des thèmes*. Ceux avec un poids faible, mais supérieur à un second seuil, inférieur, sont appelés *des mots croisés*. Tous les mots restants avec des poids inférieurs sont rejetés.

Les mesures de sélection de Bookstein consistent à appliquer des calculs de comparaison entre l'occurrence du mot dans le texte et son occurrence aléatoire prévue.

Bookstein définit deux mesures statistiques :

- la tendance du mot à se regrouper en masse, dans une unité textuelle étroite (telle qu'un paragraphe),
- la tendance d'un mot de se reproduire au moins une fois dans plusieurs unités textuelles consécutives dans un document.

Ces méthodes de mesure montrent qu'une telle information sur l'occurrence du mot améliore la qualité d'indexation, une fois comparée à la fréquence inverse du document. Cette expérience indique également un lien entre les mots « porteurs de sens » et les mots qui ont tendances à se regrouper. Malheureusement les auteurs ne précisent pas comment employer ces mesures pour caractériser les thèmes présents dans un document, ou quels sont les mots du vocabulaire associés à ces thèmes.

Brewster dans [39] définissent un canal de thème (signal) comme étant le reflet de la pertinence des mots du texte par rapport à un thème donné. Ils appliquent par la suite une Transformée en ondelettes sur la collection des canaux résultante afin de construire un signal appelé signal d'énergie composée, l'analyse de ce signal vas permettre d'identifier les ruptures thématiques à travers le document et ces ruptures peuvent être analysés à différents degrés de détails. Précisons qu'aucun résultat expérimental concernant le calcul de l'énergie composée, n'a fait l'objet d'une publication à l'heure actuelle.

L'utilisation de cette application reste pour l'instant assez 'confidentielle', il existe peu de publications relatives à ce sujet.

CHAPITRE 3 : Les Transformées en ondelettes et leurs utilisation actuelle

Le prototype Topic-O-Graphy présenté dans le chapitre précédant fait appel à l'Ondelette de Haar et ses propriétés spécialement la propriété de multi résolution afin de générer les différentes formes de visualisations de données textuelles proposées en [2].

Dans ce troisième chapitre nous exposons de manière théorique différentes méthodes d'analyse des signaux issus de la Théorie du signal, à savoir la fameuse Transformée de Fourier, ses propriétés ainsi que ses limites.

Nous exposons également les principes théoriques majeurs de l'analyse par ondelettes et ses récentes applications.

1. Pourquoi a-t-on besoin de Transformées?

Les transformations mathématiques sont appliquées aux signaux pour obtenir davantage d'informations, qu'il n'y en a apparemment de disponibles dans le signal brut. Il existe un grand nombre de transformations qui peuvent s'appliquer à un signal. Parmi celles-ci les Transformées de Fourier sont de loin les plus populaires.

Dans la pratique, la plupart des signaux sont des signaux dépendant du temps sous leur format brut. Cela signifie que, quoique le signal mesure, c'est une fonction du temps. En d'autres termes si on représente le signal, un des axes est le temps (variable indépendante) et l'autre axe (variable dépendante) est ordinairement l'amplitude. Donc une représentation temps- amplitude. Cette représentation n'est pas toujours la meilleure pour la plupart des applications de traitement de signal.

Dans beaucoup de cas, l'information pertinente est cachée dans la composante de fréquence. Le spectre de fréquence d'un signal indique quelles sont les fréquences qui existent dans le signal, nous savons que la fréquence indique le changement d'une variable, si la variable change rapidement, nous dirons qu'elle est de haute fréquence, quand une variable change lentement, nous dirons qu'elle est de basse fréquence.

Comment allons-nous mesurer la fréquence d'un signal? La réponse est la Transformée de Fourier (TF). Si on prend la TF d'un signal du domaine temporel, on obtient la représentation fréquence-amplitude de ce signal (nous obtenons un graphe dont l'un des axes est la fréquence et l'autre l'amplitude). Ce graphe nous indique la quantité de chacune des fréquences qui existent dans le signal. La TF est la plus populaire des transformées utilisées, elle n'est pas la seule. Les ingénieurs et les mathématiciens utilisent bien d'autres transformées par exemple : les Transformées de Hilbert, la Transformée de Fourier fenêtrée, Transformée de Radon et plus récemment les Transformées en ondelettes. Chacune de ces techniques de transformation possède son champ d'application, ses avantages et ces inconvénients.

Pour mettre en évidence le besoin de la Transformée en ondelettes (WT) nous allons étudier d'un peu plus près la Transformée de Fourier.

1.1 Naissance de la Transformée de Fourier

Le problème qu'eut à résoudre *Jean Baptiste Fourier*, dans le cadre de son traité sur la chaleur [41] fut le suivant :

Trouver les solutions $V(x, y)$ de l'équation aux dérivées partielles

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0 \quad (13)$$

Décrivant les transferts de chaleur, il les résolut en les développant en une somme infinie de fonctions trigonométriques.

En effet lorsque l'on considère un signal quelconque, il est indispensable d'avoir présent à l'esprit deux représentations possibles pour ce signal, une représentation temps, c'est-à-dire une représentation de la forme $y = f(t)$ dans laquelle la variable indépendante t est la durée qui s'écoule et une représentation fréquences de la forme $\gamma = f(\nu)$ dans laquelle la variable indépendante est la fréquence (dont la dimension est l'inverse du temps).

1.1.1 Transformée de Fourier des fonctions périodiques

L'analyse de Fourier décompose les signaux en fonctions élémentaires périodiques comme des fonctions sinus et cosinus.

Etant donnée une fonction $\chi(t) = f(t)$, supposée périodique pour simplifier, c'est-à-dire $f(t + T) = f(t)$, on montre que $\chi(t)$ peut s'écrire

$$\chi(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos \frac{2\pi}{T} nt + b_n \sin \frac{2\pi}{T} nt \right) \quad (14)$$

Soit

$$\chi(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos 2\pi\nu_0 nt + b_n \sin 2\pi\nu_0 nt \right) \quad (15)$$

avec : $\nu_0 = \frac{1}{T}$

La somme ci-dessus est a priori infinie elle comporte une infinité de termes. Les nombres a_0, a_1, b_1, \dots donnent le poids de chacune des sinusoïdes dans $f(t)$ et sont appelés les « Coefficients de Fourier », ils se calculent par les expressions suivantes :

$$a_n = \frac{2}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} x(t) \cos 2\pi\nu_0 n t dt \quad (16)$$

$$b_n = \frac{2}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} x(t) \sin 2\pi\nu_0 n t dt \quad (17)$$

Si l'on pose

$$x(n\nu_0) = \frac{1}{2} (a_n - j b_n) \quad (18)$$

On a
$$x(n\nu_0) = \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} x(t) e^{-2\pi\nu_0 n t} dt \quad (19)$$

$x(n\nu_0)$ est le spectre de fréquence, grandeur en général complexe, qui peut se décomposer en Spectre d'amplitude

$$x(n\nu_0) = \frac{1}{2} \sqrt{a_n^2 + b_n^2} \quad (20)$$

Et Spectre de phases

$$\varphi(n\nu_0) = \text{Arctg}\left(-\frac{b_n}{a_n}\right) \quad (21)$$

Réciproquement, on aura

$$x(\nu) = \sum_{n=-\infty}^{\infty} X(n\nu_0) e^{-j\phi(n\nu_0)} \quad (22)$$

$$x(t) = \sum_{n=-\infty}^{\infty} e^{2\pi j n \nu_0 t} \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} x(\sigma) e^{-2\pi j \nu_0 n \sigma} d\sigma \quad (23)$$

Les valeurs négatives de n sont introduites pour rendre les équations plus simples ; on vérifie aisément que si $x(t)$ est réel

$$a_{-n} = -a_n \quad \text{et} \quad b_{-n} = -b_n$$

1.1.2 Transformée de Fourier des fonctions non périodiques

On peut considérer cette non périodicité comme résultat d'une extension à l'infini de la période T .

L'intervalle de fréquences $\nu_0 = \frac{1}{T}$ tend alors vers zéro et le spectre devient alors une fonction

qui peut être continue.

On a

$$x(t) = \int_{-\infty}^{+\infty} e^{2\pi j \nu t} d\nu \int_{-\infty}^{+\infty} x(\sigma) e^{-2\pi j \nu \sigma} d\sigma \quad (24)$$

$$x(\nu) = \int_{-\infty}^{+\infty} x(t) e^{-2\pi j \nu t} dt \quad (25)$$

$x(\nu)$ est appelée la Transformée de Fourier de $x(t)$.

$x(\nu)$ est une fonction de ν , en général complexe, qui comprend donc une partie réelle (ou en phase) et une partie imaginaire (ou en quadrature).

$$\Re[X(\nu)] = \int_{-\infty}^{+\infty} x(t) \cos 2\pi\nu t dt \quad (26)$$

$$\ell[X(\nu)] = \int_{-\infty}^{+\infty} x(t) \sin 2\pi\nu t dt \quad (27)$$

Le spectre d'amplitude est

$$X(\nu) = \sqrt{\Re[X(\nu)]^2 + \ell[X(\nu)]^2} \quad (28)$$

Le spectre de phases est

$$Q(\nu) = \text{Arctg}\left(-\frac{\ell[X(\nu)]}{\Re[X(\nu)]} \right) \quad (29)$$

1.2 Signification physique de la Transformée de Fourier

$x(\nu)$ et $x(t)$ représentent la même grandeur physique, mais dans une représentation différente. Si l'on considère $x(t)$, le point représentatif se déplace dans le domaine (amplitude- temps). Si l'on considère $x(\nu)$ le point représentatif se déplace dans le domaine (amplitudes- fréquences). Lorsqu'on cherche la valeur de $x(\nu)$ pour une valeur ν_0 de ν , cela signifie que l'on cherche dans toute l'histoire de $x(t)$ ce qui correspond à la fréquence ν_0 . Ceci correspond à un filtrage infiniment sélectif.

L'analyse en fréquence de $x(t)$ nécessite donc sa connaissance totale. Si, comme c'est le cas physiquement, la fonction $x(t)$ n'est connue que dans l'intervalle $[0, T]$ il est illusoire de chercher à définir son spectre de fréquences avec une finesse supérieure à $1/T$. De même, si l'on veut retrouver $x(t)$ à partir de $x(\nu)$, il faut connaître le spectre pour toutes les fréquences jusqu'à l'infini et la formule montre que c'est la même opération de filtrage infiniment sélectif qui intervient, les variables temps et fréquences étant permutées. Cela signifie que pour connaître parfaitement la valeur de $x(t)$ à un instant t , il faut disposer d'une bande passante infinie. Tout

ceci n'est qu'une autre forme de la relation d'incertitude qui exprime l'impossibilité pour l'observateur humain à appréhender la réalité sans la déformer ou la rendre floue.

1.3 Quelques applications de la Transformée de Fourier

Nous ne donnerons pas ici l'ensemble des applications existantes de cet opérateur mathématique devenu classique, mais quelques exemples d'applications les plus importantes. Ces applications peuvent être classées en trois grands domaines.

1. Applications aux signaux monodimensionnels

Quelques applications importantes dans le cas des signaux monodimensionnels Le filtrage. - La modulation. - L'échantillonnage. - L'analyse des signaux vocaux. - L'analyse des résultats de la résonance magnétique nucléaire appliquée à l'étude des structures chimiques.

2. Applications aux signaux bidimensionnels

Description de sinusoïde bidimensionnelle. - Traitement d'images. - Filtrage de bruit ; mise en évidence des contours.- Utilisation de la transformée en cosinus en codage JPEG.

3. Applications fondées sur la propagation des ondes électromagnétiques

La propagation des ondes les interférences. - Traitement d'antenne. - Holographie. - Interférométrie et imagerie astronomique.- Rayons X, cristallographie. – Tomographie.

1.4 Limites de la Transformée de Fourier

Malgré son immense succès, cette technique a plusieurs défauts, en particulier son manque évident de localisation temporelle. En effet, l'analyse de Fourier permet de connaître les différentes fréquences excitées dans un signal, c'est-à-dire son spectre, mais ne permet pas de savoir à quels instants ces fréquences ont été émises. Cette analyse donne une information globale et non locale, car les fonctions d'analyse utilisées sont des sinusoïdes qui oscillent indéfiniment sans s'amortir. Cette perte de localité n'est pas un inconvénient pour analyser des signaux dont la structure n'évolue pas ou peu (statistiquement stationnaires), mais devient un problème pour l'étude de signaux non stationnaires.

De plus l'analyse de Fourier ne permet pas l'étude de signaux dont la fréquence varie dans le temps. De tels signaux nécessitent la mise en place d'une analyse « temps -fréquence » qui permettra une localisation des périodicités dans le temps et indiquera donc si la période varie d'une façon continue ou si elle disparaît puis réapparaît par la suite.

1.4.1 Analyse temps- fréquence

L'analyse temps- fréquence repose sur la combinaison des deux variables temps et fréquence dans une même représentation, fournissant ainsi une signature de l'évolution temporelle du contenu spectral. Pour cela il existe différentes approches la plus intuitive consiste à limiter temporellement et fréquentiellement les éléments de la famille d'analyse, puis à déplacer en tous points du plan temps- fréquence les atomes d'analyse ainsi définis, avant d'évaluer le produit scalaire avec le signal analysé.

Pour pouvoir sélectionner un certain intervalle de temps et analyser les composantes de fréquence dans cet intervalle, nous devons observer le principe d'incertitude d'Heisenberg.

1.4.2 Principe d'incertitude d'Heisenberg

Le principe d'incertitude vient de la mécanique quantique, mais il joue un grand rôle dans le traitement du signal. Il stipule que l'on ne peut localiser aussi précisément que l'on veut en temps et en fréquence un signal.

La variation temporelle σ_t et la variation fréquentielle σ_w d'un signal f satisfait

$$\sigma_t^2 \sigma_w^2 \geq \frac{1}{4} \quad (30)$$

Ce théorème montre qu'il possible de sélectionner une position temporelle spécifique dans le signal et calculer la composante fréquentielle correspondante.

1.5 Transformée de Fourier Fenêtrée

La solution initiale pour résoudre le problème d'analyse des signaux transitoires était la Transformée de Fourier Fenêtrée (TFF). Elle est similaire à la Transformée de Fourier dans le sens où elle convertit un signal dans le domaine de fréquence, mais elle divise le signal en segments suffisamment petits pour que sur ces portions, le signal puisse être considéré comme stationnaire. A cet effet, on choisit une fonction de fenêtrage (w). La largeur de cette fenêtre doit être égale à la longueur du segment où il est considéré comme stationnaire.

La fonction de fenêtrage est d'abord positionnée au tout début du signal, elle est donc placée à $t = 0$. Supposons que la largeur de la fenêtre soit T . A cet instant, $t = 0$, la fenêtre recouvre les $T/2$ premières secondes. La fonction fenêtre et le signal sont alors multipliés. On sélectionne ainsi les $T/2$ premières secondes du signal et on les pondère. Si, par exemple, la fenêtre est un rectangle d'amplitude 1, alors le produit est égal au signal. Le produit obtenu est alors considéré comme un autre signal, dont on peut calculer la transformée. Autrement formulé on calcul la TF du produit de la même manière qu'on calcul la TF d'un signal quelconque (stationnaire).

Le résultat de ce travail est la TF des $T/2$ premières secondes du signal. Si cette portion du signal est stationnaire. L'étape suivante consiste à décaler la fenêtre de t_1 secondes vers un autre emplacement, de la multiplier avec le signal puis de calculer la TF du produit obtenu. Cette procédure se répète, en déplaçant la fenêtre d'un intervalle de t_1 secondes, jusqu'à atteindre la fin du signal.

La définition de la STFT 'Short-Time Fourier Transform' suivante résume ce qui vient d'être exposé

$$TFT(t, f) = \int_t [x(t) \cdot w^*(t - t')] \cdot e^{-j2\pi ft} dt \quad (31)$$

Avec

$x(t)$ est le signal lui-même, $w(t)$ est la fonction fenêtre et w^* son complexe conjugué.

Le problème de la TFT, réside dans le principe d'incertitude. Comme pour le principe d'Heisenberg (l'impossibilité de connaître, au même instant, le moment et la position d'une

donnée élémentaire) il est impossible d'obtenir simultanément les informations de temps et de fréquence d'un signal. On ne peut donc obtenir une exacte représentation temps/fréquence d'un signal, c'est-à-dire, on ne peut savoir quelles composantes spectrales existent à un instant donné. Tout ce que nous pouvons connaître ce sont les intervalles de temps pendant lesquels une certaine bande de fréquences existe. Le problème de la TFT est un problème de résolution.

Cette résolution dépend de la largeur de la fonction de fenêtrage utilisée. Pour être techniquement correct, cette largeur de fenêtre est désignée par le terme *support* de la fenêtre. Si la fonction de fenêtrage est étroite, on la dit à *support compact*. Cette terminologie est davantage utilisée dans le monde des Ondelettes comme nous le verrons plus loin dans ce chapitre.

1.6 La Transformée en Ondelettes

Les deux approches mathématiques présentées dans les précédemment sont adaptées aux processus stationnaires. De nouvelles méthodes élaborées et mises au point ces dernières années, unifient et généralisent les idées et les pratiques développées précédemment et permettent d'analyser des signaux non-stationnaires. La Transformée en ondelettes fait partie de ces nouvelles méthodes, son principe est de décrire l'évolution temporelle d'un signal à différentes échelles de temps.

La théorie des ondelettes est apparue au début des années 1990 [44], elle touche de nombreux domaines des mathématiques, notamment le traitement du signal et des images [45], [42].

Cette section présente un rapide aperçu des fondements théoriques des Ondelettes, pour aller plus loin sur cette théorie du traitement du signal à l'aide des Ondelettes, le lecteur pourra se reporter au livre de Mallat [43].

Malgré une origine aux nombreuses racines, on attribue le point de départ de l'utilisation des Ondelettes au géophysicien Jean Morlet, qui envisageait de les utiliser pour l'analyse de sismogrammes utilisés dans la recherche de pétrole sous terre.

Dans la transformation par Ondelettes, comme dans l'analyse de Fourier, on cherche à transformer un signal quelconque en une série de nombres que l'on pourra ensuite utiliser pour reconstruire au mieux le signal d'origine. Cependant dans la transformation par Ondelettes, on utilise plusieurs niveaux de résolution pour examiner le signal et faire ressortir les différentes variations.

L'analyse multi résolution donne un ensemble de signaux d'approximation et de détails d'un signal de départ en suivant une approche fin-à-grossier. On obtient une décomposition multi-échelle du signal de départ en séparant à chaque niveau de résolution les basses fréquences (approximation) et les hautes fréquences (détails) du signal.

Cette approche a un sens quand le signal a des composantes à haute fréquence pour des courtes durées et des composantes de basse fréquence pour de longues durées. Pour accomplir une telle tâche une Ondelette sera employée au lieu d'une fonction de fenêtrage, la Transformée en ondelettes est capable de fournir les informations de temps et de fréquence simultanément et donc une représentation temps – fréquence du signal.

1.6.1 Définition

La Transformée en ondelettes est une représentation multi-résolution, qui exprime les variations d'un signal à différentes résolutions. Une Ondelette ψ est une fonction oscillante, comme les fonctions sinus et cosinus, mais localisée. Cela se traduit par le fait qu'elle est intégrable de valeur 0.

$$\int_{-\infty}^{+\infty} \psi(x) dx = 0 \quad (32)$$

Le signal est étudié aux échelles $1/2, 1/4, \dots, 2^j$, avec $j \in \mathbb{Z}$ et $j \leq -1$

$\forall x \in \mathbb{R}, \psi_2^j(x) = 2^j \psi(2^j x)$ est la dilatation de l'Ondelette $\psi(x)$ à l'échelle 2^j .

1.6.2 Les propriétés des Ondelettes

Avant de calculer une transformation en Ondelettes sur des données recueillies, nous devons d'abord choisir une Ondelette parmi les nombreuses variétés qui existent. Nous pouvons citer les plus reconnues : Daubechies, Shannon, Battle-Lemarié, Meyer, ...

Les deux facteurs principaux qui influencent le choix d'une Ondelette sont le nombre de moments nuls et la taille de support.

1. Les moments nuls

Pour le stockage et la transmission des signaux, la meilleure approche consiste à employer le minimum d'espace de stockage et de la moindre quantité de largeur de bande.

Dans le domaine de la recherche d'information, une compression optimale de données signifie l'utilisation d'un espace minimum, permettant la facilité d'accès et la recherche d'information, nous devons représenter le maximum d'information dans le plus petit espace possible.

En langage de transformée en ondelette cela signifie la production d'un maximum de coefficients d'Ondelette proches de 0. Le moment k de la fonction $f(t)$ est défini comme suit

$$\int_{-\infty}^{+\infty} t^k f(t) dt \quad (33)$$

On dit qu'une Ondelette possède n moments nuls si l'équation (33) vaut 0, pour les valeurs de k suivantes $0 \leq k < n$.

2. La taille du support

Le second paramètre qui affecte le choix de l'une Ondelette est la taille de son support. Le support de la fonction f est le domaine (l'intervalle) ou la fonction est non nulle. Une fonction f à un support compact si son support est limitée. Par exemple la fonction carrée à un support compact dans l'intervalle $[0, 1]$.

$$f(t) = \begin{cases} 1 & 0 \leq t \leq 1 \\ 0 & \text{sinon} \end{cases}$$

Si une fonction f a une singularité au point t_0 et si t_0 est dans le support compact de $\psi_{j,n}(t) = 2^{-j/2}\psi(2^{-j}t - n)$, alors $\langle f, \psi_{j,n} \rangle$ peut être de grande amplitude [43]. Si ψ a un support compact de taille k , il existe, à chaque échelle 2^j , k Ondelette $\psi_{j,n}$ dont le support compact contient t_0 .

Pour minimiser le nombre de coefficients de grande amplitude, on peut diminuer la taille du support de ψ et donc moins de composantes significatives à prendre en considération dans les calculs pour le stockage des signaux.

Dans le contexte de la modélisation et la recherche d'information, la taille de l'index sera plus compacte, si nous choisissons une Ondelette avec un support de petite taille.

1.6.3 L'Ondelette de Haar

L'Ondelette de Haar est l'Ondelette dont le support est le plus petit, cela implique que sa transformée du signal nécessitera le minimum d'espace de stockage.

Soit h la fonction, dite de base de Haar, définie sur \mathcal{R} par

$$h(x) = \begin{cases} 1 & \text{si } 0 < x < 1/2 \\ 0 & \text{sinon} \end{cases}$$

Supposons que nous avons un signal défini sur l'intervalle $[0,1]$. Pour avoir une approximation discrète du signal nous allons calculer ses valeurs dans deux points, quatre points, huit points et ainsi de suite ; le diviser en deux fonctions, de 0 à 1/2 et de 1/2 à 1, puis en quatre fonctions, de 0 à 1/4, de 1/4 à 1/2, de 1/2 à 3/4, et de 3/4 à 1 etc....

On obtient différentes résolutions et pour chacune on peut avoir une représentation dans l'espace des fonctions à l'aide d'un système de fonctions de base, nommées fonctions de base multi-résolutions ou multi-échelles.

Les Ondelettes sont des fonctions de base multi - échelles qui assurent le passage cohérent entre les différentes résolutions, la décomposition et la reconstitution de la fonction représentée. Si on utilise les Ondelettes comme système de fonctions de base, à chaque niveau on dispose des approximations (moyennes) de la fonction initiale et des informations de détails.

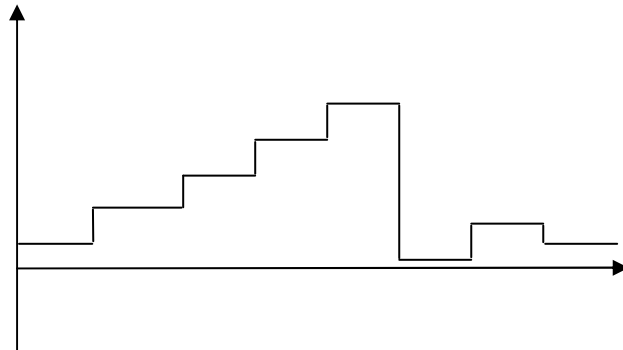
1.6.4 Exemple de calcul

La transformée de Haar de la fonction $f(x) = [y_1 y_2 y_3 \dots y_n]$ génère

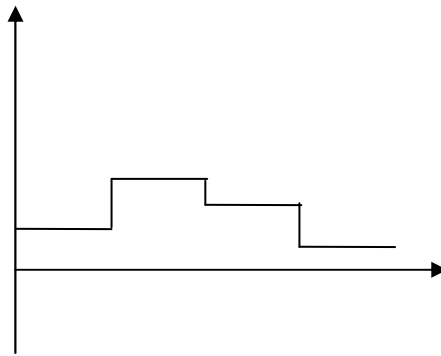
- Des approximations $[a_1 a_2 a_3 \dots a_{n/2}]$ qui sont les moyennes des valeurs initiales de la fonction prises deux par deux $a_1 = (y_1 + y_2) / 2 \dots$
- Coefficients de détail ou les différences $[d_1 d_2 d_3 \dots d_{n/2}]$, avec $d_1 = y_1 - a_1$, $d_2 = y_3 - a_2$

Considérons un signal monodimensionnel composé de quatre échantillons

$$S = [2 \ 4 \ 8 \ 12 \ 14 \ 0 \ 2 \ 1]$$



Pour calculer sa transformée de Haar, moyennons d'abord les paires de valeurs voisines pour obtenir $[3 \ 10 \ 7 \ 1.5]$



Afin de récupérer le signal initial nous devons également enregistrer d'autres valeurs représentant la perte d'information.

$$[-1 \ -2 \ 7 \ 0.5]$$

Le signal peut donc être représenté par sa résolution inférieure et le signal de détail.

En appliquant ce procédé, récursivement sur le signal on aboutit à sa transformée de Haar, à la fin signal est représenté par un seul coefficient de moyenne du signal et l'ensemble de coefficients des signaux de détails successifs.

Résolution	Moyenne	Détails
8	[2 4 8 12 14 0 2 1]	
4	[3 10 7 1.5]	[-1 -2 7 0.5]
2	[6.5 4.25]	[-3.5 2.75]
1	[5.375]	[1.125]

Tableau 4. Transformée de Haar du signal $S = [2 \ 4 \ 8 \ 12 \ 14 \ 0 \ 2 \ 1]$

Observons la transformée de Haar ainsi obtenue, en plus du coefficient de moyenne du signal, les coefficients de détails expriment les variations du signal aux différentes résolutions. A une même échelle, plus le coefficient est grand en valeur absolue, plus ces variations sont importantes. Le signal original sera présenté par $[5.375 \ 1.125 \ -3.5 \ 2.75 \ -1 \ -2 \ 7 \ 0.5]$.

Pour effectuer cette transformée, deux filtres ont été utilisés [46] le filtre d'échelle $[1 \ 1]$ qui permet de calculer le signal de résolution inférieure et le filtre d'Ondelette $[1 \ -1]$ qui permet de calculer le signal de détail.

Appliquer le filtre d'Ondelette au signal revient à calculer le produit scalaire du signal et de la fonction Ondelette ψ .

$$\Psi(x) = \begin{cases} 1 & \text{pour } 0 \leq x < \frac{1}{2} \\ -1 & \text{pour } \frac{1}{2} \leq x < 1 \end{cases}$$

1.6.5 L'utilisation actuelle des Ondelettes

La théorie des Ondelettes est apparue au début des années 1990, elle touche de nombreux domaines notamment le traitement du signal et des images [47]. Aujourd'hui son utilisation s'est élargit à différents domaines de recherche particulièrement la fouille des données [48] (Data Mining), ce domaine représente un secteur de recherche très important dans le domaine scientifique ainsi que dans l'industrie. La fouille de données a vu le jour dans les années 80, quand les professionnels ont commencé à se soucier des grands volumes de données informatiques inutilisables.

Le Data Mining consistait essentiellement à extraire de l'information de gigantesques bases de données de la manière la plus automatisée possible ; contrairement à aujourd'hui où le Data Mining consiste à la sélection, l'exploration, la modification et la modélisation de grandes bases de données afin de découvrir des relations entre les données jusqu'alors inconnues.

La théorie des Ondelettes joue un rôle important dans ce domaine, les Ondelettes ont beaucoup de propriétés favorables telles que les moments nuls, la décomposition multi résolution et la représentation hiérarchique, les coefficients de corrélation ainsi que tout un ensemble de fonctions de base. Ces propriétés pourraient fournir des représentations des données qui permettent de rendre le processus d'extraction et de traitement plus efficace et précis.

Les Ondelettes ont pu être incorporées dans beaucoup d'algorithmes de recherche d'information, principalement sur des données qui ont des localités temporelles/spatiales, par exemple, les séries chronologiques, le flux des données, ...

Dans cette partie, nous essayerons de donner un survol sur les secteurs où les Ondelettes sont récemment employées.

1.6.5.1 La gestion des données

L'objectif de la gestion des données est de trouver des méthodes de stockage et d'organisation permettant un accès facile, rapide et efficace aux données. La Transformée en ondelettes fournit une structure hiérarchique et une représentation de données multidimensionnelle, ce qui favorise son utilisation dans la gestion de données.

SHAHABI et ses col dans [49], [50] présentent deux structures arborescentes TSA-Tree et 2d TSA-Tree, (Trend and Surprise Abstractions Tree). Les deux structures sont utilisées pour détecter les changements des données dans les séries chronologiques et améliorer le processus d'interrogation.

(TSA-Tree) se base sur la Transformée en ondelette Discrète, la racine de l'arbre (figure 16) est représentée par les données originales de la série chronologique et chaque niveau de l'arbre correspond à une étape de la décomposition en Ondelettes.

Au premier niveau de décomposition, les données originales sont décomposées en deux parties une première pour les basses fréquences (tendances) et une autre pour les hautes fréquences (surprises) et le processus se répète jusqu'au dernier niveau de décomposition.

(2D-TSA-Tree) est une extension bidimensionnelle de (TSA-Tree), elle applique une Transformée en ondelettes sur des données 2D afin d'obtenir l'ensemble des coefficients d'approximations et de détails.

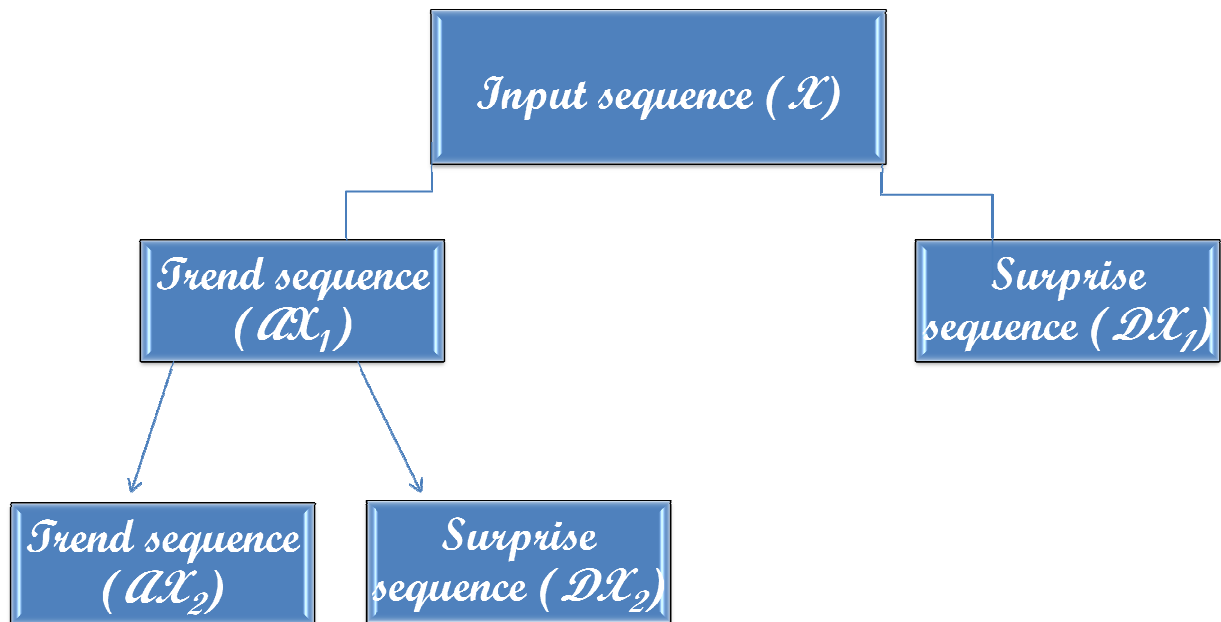


Figure 15. La structure de TSA-tree : X représente les données originales, AX_i et DX_i représente la tendance et la surprise au niveau i

Avec la popularisation de la quantité d'images numériques, la gestion des bases de données des images devient de plus en plus difficile. Venkatesan et col. dans [51] proposent une nouvelle technique d'indexation des images (fonction de hachage). Le processus d'indexation est le suivant : ils calculent d'abord une décomposition d'une image on Ondelette et chaque sous bande de l'image (segment) est aléatoirement couverte de petits rectangles.

Les statistiques (variance et moyenne) de chaque rectangle seront calculées et présentées comme valeur d'entrée pour l'étape de codage afin de choisir la clé de hachage. Les expériences ont prouvé que le hachage des images est robuste contre tout traitement d'image commun et modification malveillante.

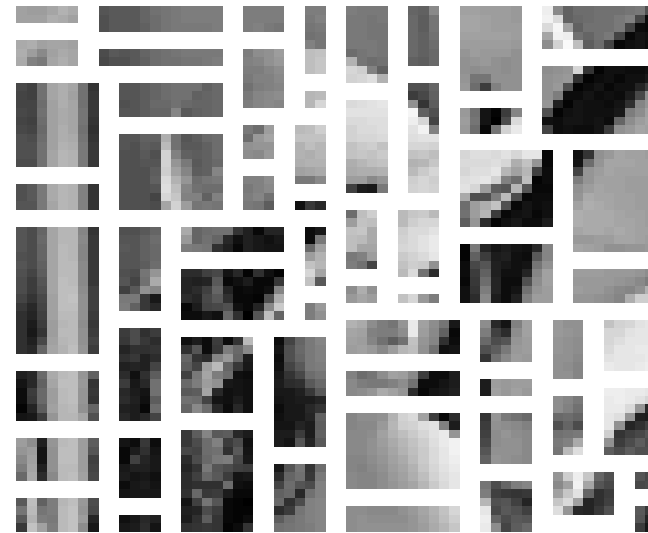


Figure 16. Hachage aléatoire de l'image de Lena (décomposition en 3 niveaux avec l'Ondelette de Haar)

1.6.5.2 Le débruitage

Le bruit est une erreur aléatoire d'une variable mesurée. Il existe plusieurs raisons possibles pour générer des données bruitées, tel que les erreurs de mesures pendant l'acquisition des données, les erreurs humaines, ou les erreurs de machines lors de la saisie des données. On peut définir, donc, le débruitage comme le processus d'identification des données optimales parmi les données bruitées disponibles.

L'idée principale est de transformer les données en différentes bases d'Ondelette ou les grands coefficients représentent l'information utile, tandis que les plus petits représentent le bruit.

Donoho et Johnstone [52] présentent une technique appelée WaveShrink (Shrinkage Methods méthodes de rétrécissement). La technique applique une Transformée en ondelette sur les données originales, les coefficients d'ondelettes dont la valeur absolue est au dessous d'un certain seuil seront placés à zéro.

1.6.5.3 La réduction de dimension

La réduction de dimension consiste à projeter des données d'un espace de n dimensions dans un autre espace de dimension k avec $k \ll n$ avec ou sans perte d'informations.

Les Ondelettes offrent deux options différentes pour la réduction de dimension des données la première consiste à garder les plus grands k coefficients d'Ondelette et approximer le reste à 0. La seconde consiste à garder que les k premiers coefficients d'ondelettes et approximer le reste à 0.

1.6.5.4 La clusterisation

Le problème de clusterisation des données surgit dans beaucoup de disciplines et applications. On peut décrire le concept de clusterisation comme suit :

Soit w un ensemble de n points dans un espace multidimensionnel, il s'agit de trouver une partition de w en un certains nombre de classes tels que les points dans chaque classes seront semblables.

La propriété de multi résolution des Ondelettes à permet aux chercheurs d'améliorer leurs algorithmes de clusterisation. WaveCluster [53] représente une approche de clusterisation pour des larges bases de données, ou le partitionnement de l'espace de données par une grille réduit le nombre de données à considérés.

Dans une perspective de traitement de signal, si la collection de données est considérée comme un signal de dimension n , les parties du signal de hautes fréquences correspondent aux régions de l'espace ou il y'a un changement rapide de distribution des données (les frontières des clusters). Les parties de basses fréquences du signal correspondent aux secteurs de l'espace ou les données sont concentrés (les clusters).

Appliquer la Transformée en ondelettes sur un signal le décompose en bandes de fréquences, par conséquent, l'identification des clusters revient à trouver les composantes reliées dans l'espace transformé et à différents niveaux de résolution.

1.6.5.5 La visualisation

La visualisation est une tache de description dans le data Mining, elle permet à l'utilisateur d'explorer et de mieux comprendre les données en exploitant au maximum le graphisme en détriment du texte et aux nombres. Cependant, pour les larges bases de données, même une simple opération de visualisation est difficile à réalisée et la Transformée en ondelettes à permis l'accès progressif aux données volumineux.

Dans le deuxième chapitre on a présenté une nouvelle approche de visualisation et exploration des textes non structurés basée sur les Ondelettes [2]. La technologie de base applique la Transformée en ondelettes sur un signal construit à partir des mots du document et l'énergie

résultante de l'Ondelette employée sert à analyser les caractéristiques du flot narratif du texte dans le domaine de fréquence.

Rik Littlefield chercheur au PNNL 'The Pacific Northwest National Laboratory' présente cette technologie ainsi: "*This technology could help people who are overloaded with information, such as teachers, researchers and lawyers*". Ce laboratoire a testé cette technologie sur les discours de Fidel Castro au cours de ces 30 dernières années. Ces tests ont détecté le thème principal de chaque discours, ainsi que l'ordre dans lequel apparaissent les différents sous thèmes.

1.6.5.6 La similarité requête / donnée

Le concept de la similarité consiste à trouver des données qui répondent à une requête précise, en se basant sur une certaine mesure de similitude. Cette tâche prend tout son sens dans le traitement des séries chronologiques, les images, les textes, ...

Pour les textes par exemple, il faut trouver l'ensemble des documents pertinents, à partir d'un ensemble de mots clés correspondant à une requête.

Jacobs et col. [55] présentent une méthode basée sur l'utilisation d'une image comme une requête, la requête peut être potentiellement de basse qualité par rapport à l'image recherchée.

La méthode '*image querying metric*' se base sur le calcul des singularités (coefficients de décomposition), elle compare essentiellement le nombre de coefficients d'Ondelette significatifs que la requête a en commun avec les cibles potentielles.

Natsev et col. [56] proposent l'algorithme WALRUS (WAveLet-based Retrieval of User-specified Scenes) pour la recherche de similitude dans des bases de données d'images. Dans le même domaine, Ardizzoni et col. [57] décrivent *Windsurf* (Wavelet-Based Indexing of Images Using Region Fragmentation), une nouvelle approche pour la recherche d'images, *Windsurf* utilise l'Ondelette de Haar pour extraire les caractéristiques de couleur et de texture, il applique également des techniques de clusterisation pour partitionner les images en régions

Wang et col. dans [58] décrivent *WBIS* (Wavelet-Based Image Indexing and Searching), un nouvel algorithme d'indexation et de recherche d'image. *WBIS* applique l'Ondelette Daubechies-8 pour chaque composant de couleur, les coefficients de basse fréquence et leurs variances sont enregistrés sous forme de vecteurs. Pour accélérer la recherche, le processus

s'effectue en deux étapes : d'abord une sélection brute basée sur les variances, puis raffinement de la recherche en exécutant une comparaison de vecteurs de caractéristiques entre les images choisies et l'image requête.

Dans le domaine d'aide à la décision, Chakrabarti et col. [59] démontrent que le traitement heuristique des requêtes sur d'énormes volumes de données avec des Ondelettes semble fournir des résultats intéressants avec des temps de réponses très courts.

Conclusion

Nous avons vu que pour l'analyse des signaux non stationnaires, une solution consiste à utiliser une variante de la Transformée de Fourier classique : la Transformée de Fourier Fenêtré. Cependant, cette solution a des limites, notamment dans le choix de la taille de la fenêtre d'analyse qui détermine si nous nous concentrons sur une analyse fréquentielle du signal ou sur une analyse des événements temporels. La Transformée en ondelettes représente une alternative à ce problème.

Comme la Transformée de Fourier, la Transformée en ondelette fournit une représentation temps-fréquence du signal, mais avec une résolution variable. Nous pouvons ainsi effectuer une analyse multi-résolution du signal et étudier plus finement les détails du signal en l'observons à différentes échelles.

Dans ce chapitre nous avons donné un survol sur les bases mathématiques de la théorie des Ondelettes, nous avons présenté également certaines applications récentes des Transformées en ondelettes à savoir : la visualisation, la gestion des données, la classification, ...

Dans le chapitre suivant nous allons utiliser et appliquer la Transformée en ondelettes dans le domaine de la modélisation et la recherche d'information textuelle.

CHAPITRE 4 : Modélisation Spectrale des données textuelles : vers un Système de Recherche d'Information Spectral

Après avoir présenté les principaux systèmes de recherche d'informations, leurs principes de fonctionnement, leurs avantages ainsi que leurs limites, nous avons également présenté le nouveau concept de la modélisation spectrale dans la visualisation des grandes masses de données textuelles non structurés.

Dans ce chapitre, nous allons construire une approche de modélisation et de recherche spectrale pour les données textuelles. L'hypothèse exprimée dans ce chapitre est que le texte peut être considéré comme un signal.

Pour valider notre modèle de recherche d'information, nous allons dans une première expérimentation appliquer la méthode de modélisation spectrale sur un corpus de données textuelles, nous allons construire les signaux thématiques correspondant aux données et effectuer une comparaison des signaux dans le but d'une comparaison documentaire.

Dans une seconde expérimentation nous allons valider notre algorithme de recherche d'information sur un corpus. Nous allons construire les signaux correspondant aux termes d'une requête et calculer la pertinence des données par rapport à cette requête. En dernier et afin de juger l'efficacité de notre système de recherche d'information, nous allons comparer nos résultats avec les résultats obtenus à l'aide de la présentation vectorielle.

Introduction :

Les systèmes de recherche d'information traditionnels se basent sur une représentation dite plate 'sac de mots' où les documents sont représentés simplement par la présence, l'absence ou la fréquence d'apparition des mots dans le texte.

Dans le deuxième chapitre on a rappelé les différentes améliorations apportées à cette approche de représentation plate et nous avons constaté que l'information de position des mots dans le texte est ignorée. Si on examine les deux documents suivants [95]:

Document 1: « The smell of the bakery first appeared at five in the morning. The maroon delivery trucks passed through the town ten minutes later »

Document 2 :«The smell of the diesel first appeared at five in the morning. The bakery delivery trucks passed through the town ten minutes later ».

Et étant donné la requête suivante: « bakery trucks ».

En utilisant le modèle vectoriel chacun de ces deux documents aura le même score de pertinence par rapport à la requête. Pour les distinguer on doit observer les positions des mots dans les deux textes.

Les travaux réalisés, au cours de l'étude suivante, ont pour objectif d'évaluer l'impact de l'ordre d'apparition des termes dans un texte pour la recherche d'information dans un corpus de documents textuels.

Considérons les deux documents précédents :

Dans la représentation vectorielle, un document est représenté par un vecteur en fonction de la fréquence des mots ou de leur présence et/ou absence dans le document. Dans cet exemple nous allons considérer les fréquences d'apparition des mots dans les textes.

Le lexique de ces deux documents est constitué des mots suivants: « *smell, bakery, first appeared, five, morning, maroon, delivery, trucks, passed, through, town, ten, minutes, later, diesel* ».

Dans le premier document, le mot *smell* apparaît une fois, le mot *bakery* apparaît une fois et ainsi de suite, le vecteur V_1 comportera ces informations $V_1 = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0\}$.

Le vecteur V2 comportera les valeurs suivantes, $V2=\{1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$.

Si on considère la requête « bakery trucks », le vecteur correspondant comportera les valeurs suivantes : $V3= \{0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0\}$.

En utilisons la mesure de similarité de cosinus (équation 34), l'angle formé entre le vecteur requête et ceux des documents est le même donc les deux documents sont similaires (dans cette représentation) par rapport à la requête.

$$\text{Cos}\alpha = \cos(V1, V2) = \frac{V1.V2}{\|V1\|\|V2\|} \quad (34)$$

En effet, dans la représentation vectorielle classique [61], [62], lors de la création des vecteurs documents, seule les fréquences d'apparition des termes sont présent en considération mais les positions de ces termes sont ignorées. Certains travaux de recherche ont essayé d'inclure l'information de position des termes [63], [64], [65], ces travaux ont donné naissance à plusieurs méthodes de proximité qui calculent la pertinence d'un document en se basant sur la distance entre les termes de la requête dans le document. Ces méthodes fournissent une précision élevée lors de la recherche mais les essais de prise en compte des occurrences des mots dans le calcul de la pertinence augmentent le temps d'interrogation ainsi que l'espace de stockage des données de façon significative.

Dans notre travail nous n'allons pas proposer une nouvelle amélioration de la représentation vectorielle mais une nouvelle représentation de l'information textuelle basée sur la modélisation spectrale, en considérant à la fois les notions d'occurrences mais également de l'ordre d'apparition des mots dans le texte.

L'originalité de cette représentation repose sur la façon de représenter un document, non plus simplement comme un ensemble de vecteurs mais comme un ensemble de signaux décrivant le contenu du document. Cette nouvelle forme de représentation nous permettra par la suite d'appliquer de nombreux outils mathématiques connus en théorie du signal, tel que les Transformées en ondelettes et jusqu'à aujourd'hui inutilisés dans le domaine de la recherche d'information textuelle.

Exemple

Considérons l'exemple suivant:

“The emergence of library networks is discussed. Management issues involving network structure, economics, and applications of computer technology are considered.”(Base de données LISA)

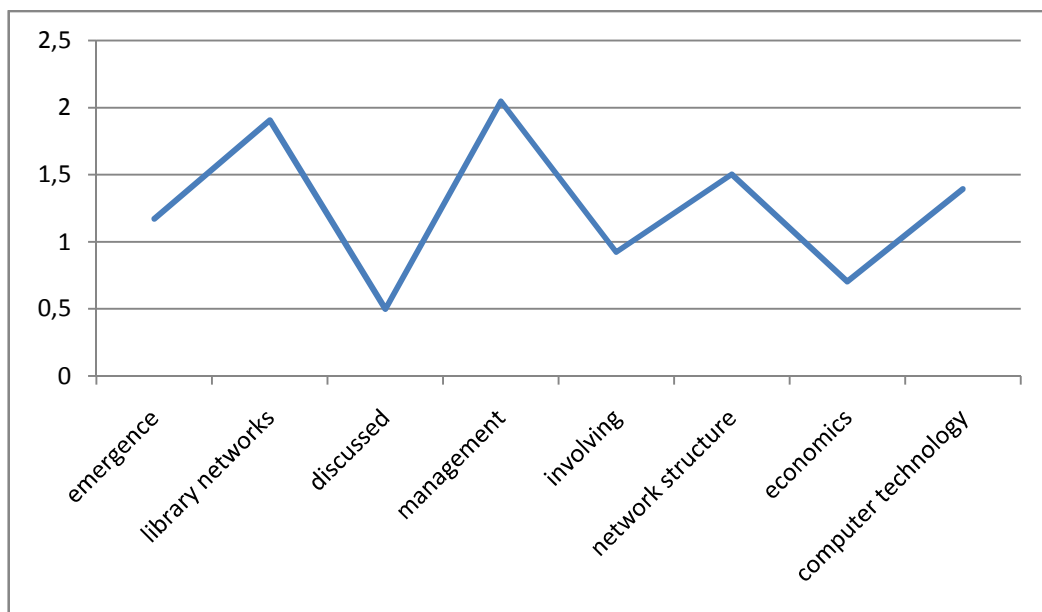


Figure 17. Signal de pertinence correspondant à l'exemple 1

Le signal sur la figure 17, représente la modélisation du texte de l'exemple 1. L'axe des abscisses correspond à l'ensemble des termes du texte dans leur ordre d'apparition, sur l'axe des ordonnées on retrouve les niveaux de pertinence correspondants à ces termes présentés dans le tableau 5.

termes	Pertinence thématique
emergence	1,17098566035656
library networks	1,90603103013786
discussed	0,498887802
management	2,04604692374826
involving	0,923831045475437
network structure	1,50197887939799
economics	0,703624242926058
computer technology	1,39283440997292

Tableau 5. Pertinence des mots de la phrase exemple 1. Les mots vides sont éliminés.

Cette modélisation des niveaux de pertinence représente le flot du thème « Library networks » dans le texte, sous forme d'un signal ; le flot de chaque thème pourra être représenté de la même manière. Les signaux sont des formes puissantes et universelles utilisées dans des applications de traitement de son, des images, des discours, etc.,...la qualité du signal dépend de la mesure de pertinence utilisée dans le calcul de la pertinence thématique des termes du texte.

Brewster et.col dans [39] représentent également des documents en fonctions de signaux, ils définissent les thèmes comme les mots ayant un poids plus élevé qu'un seuil donné .les auteurs appliquent par la suite une Transformée en ondelettes sur la collection des signaux résultante afin de construire un signal appelé signal d'énergie composée. Ce signal représente le contenu du document et c'est la base de leur prototype de visualisation et de segmentation Topic-O-Graphy.

1. Pourquoi une modélisation spectrale

Le recours à la modélisation de l'information textuelle en signaux paraît un choix judicieux en considérant les arguments suivants :

- Peu ou pas d'améliorations au niveau de la modélisation des données textuelles, la majorité des systèmes actuels se basent sur la représentation vectorielle classique, les améliorations se portent essentiellement sur les mesures de similarité utilisées,
- la modélisation en signaux a déjà prouvé son utilité dans divers domaines, notamment le traitement d'images, de la parole et du son,
- une telle approche permet de faire appel à différents outils mathématiques empruntés à la théorie du signal et jusque là inexploités dans un contexte de recherche d'information et comparaison textuelle.

2. Notions et fonctions

Nous devons tout d'abord donner une vue globale des notions et fonctions définies et utilisés dans le modèle que nous proposons.

Notion de fréquence :

La fréquence tf d'un terme w (mot ou mot composé) dans un corpus de documents correspond au nombre d'occurrences du terme w dans la base. Sa fréquence Idf correspond au nombre de documents contenant le terme w .

Notion de cooccurrence :

Dans la recherche d'information on définit généralement des cooccurrences entre des termes et des documents contenant ces termes. Ou entre des termes et les thématiques présentes dans un document.

De deux termes i et j on dit qu'ils co-occurrent ou qu'ils sont associés s'ils sont utilisés ensemble pour décrire un même document.

Notion d'ordre chronologique :

Dans un document textuel, l'auteur peut traiter un ou plusieurs sujets. L'ordre chronologique dans un texte représente l'ordre d'apparition des termes qui construisent la discussion autour de ces sujets.

Notion de Signal Thématique :

Nous désignons par *Signal Thématique* : la modélisation du flot de la discussion autour d'un thème donné dans un texte. Le signal relie les termes de la discussion dans leur ordre d'apparition dans le texte.

Notion de Thème :

Dans le contexte de la modélisation spectrale, un thème représente l'idée ou le sujet développé dans un texte. Un document peut traiter d'un ou plusieurs thèmes à la fois sans aucune supposition sur leur organisation. Traiter la thématique d'un texte revient donc pour nous à mettre en évidence les principaux sujets abordés dans ce dernier.

Il existe plusieurs méthodes pour le choix des thématiques représentatifs d'une base de documents textuels. Nous expliquerons dans ce qui suit ces différentes méthodes.

La détection des thèmes dans un document :

L'ensemble des mots clés indique généralement les principaux thèmes présents dans un document. Certains mots clés sont présents dans la partie « *KeyWords* » du document, ou ils peuvent être reconnus en utilisant des extracteurs terminologiques [Annexe 1]. L'ensemble des mots extraits est appelé « *vocabulaire* » et ils sont employés pour construire et faire évoluer les discussions autour des thématiques du texte.

La détection des thèmes traités dans un document est donc fonction de deux paramètres principaux qui sont le vocabulaire et les méthodes de calcul utilisées. Le vocabulaire définit l'ensemble des mots candidats caractéristiques d'un thème. Classiquement, c'est sur la base de l'ensemble des mots constituant le vocabulaire que la plupart des méthodes fondent leurs principes de détection.

- *Le vocabulaire:*

Dans le domaine de la recherche d'information, la majorité des méthodes de détection des thèmes utilise le mot comme unité de représentation du document. La question qui se pose est de savoir quels sont ces mots.

Damereau [66] définit l'ensemble du vocabulaire comme étant: *“a list of content words not necessarily complete that would characteristically be used in talking about a particular subject, say education, as opposed to the list of words used to talk about, say aviation”*.

Il teste deux approches pour l'extraction des mots candidats d'un domaine donné dans un corpus de documents, la première approche consiste à créer une liste de tous les mots du texte avec leurs fréquences, il élimine les mots vides (*Stop Words*) et il élimine, également, tous les mots ayant une fréquence minimum. La seconde approche consiste à créer deux listes, dans la première il extrait tout les mots du dictionnaire concernant un domaine donnée. La seconde liste index les mots du domaine utilisés dans le texte. On crée ensuite une liste finale des mots communs aux deux listes.

Néanmoins, lorsque l'on observe le nombre d'occurrences d'un terme dans un document donné, on remarque que celui-ci est proportionnel à l'intérêt de ce terme pour le document. Cela provient du fait qu'un auteur répète naturellement les termes importants à la compréhension de son article. Bien entendu, la polysémie et la synonymie tempèrent l'étendue de la portée de cette constatation, mais de manière générale un terme fréquent dans un document est plus important qu'un terme rare donc la fréquence d'un terme est le plus souvent utilisée pour montrer l'importance de celui-ci dans la présentation du contexte.

Un des problèmes de l'extraction terminologique est qu'un document, même après filtrage, contient toujours des mots sans intérêt qui ne sont pas représentatifs du discours. Ces mots communs appelés « mots vides » foisonnent dans les corpus documentaires et dépendent du corpus étudié.

Il existe dans la littérature plusieurs méthodes de calcul permettant de trouver cet ensemble de termes représentatif des thématiques dans un document, nous présentons ici les méthodes les plus étudiées :

1. La fréquence des mots T_f : on calcul pour chacun des termes sa fréquence d'apparition, le vocabulaire sera composé de termes avec les fréquences les plus élevés.

2. La fréquence de documents des mots I_{df} :

Dans ce cas, on ne prend pas en compte la fréquence des termes mais le nombre de documents dans lesquels chaque terme est apparu. Le vocabulaire résultant sera composé des termes apparus dans le plus grand nombre de documents.

3. Indice $T_f^* I_{df}$:

Il s'agit de combiner les deux mesures précédentes afin de trouver des termes fréquents mais présents dans peu de documents.

Ces trois notions ont été abordées avec plus de détails dans le chapitre 2, section « Identification des termes d'indexation ».

4. Information mutuelle:

La mesure de l'Information mutuelle quantifie le lien existant entre un terme et un thème. Plus précisément elle évalue l'influence qu'a sur le thème d'un texte, la présence d'un terme dans ce texte, elle est évaluée de la façon suivante :

$$I(x, y) = \log \frac{p(x, y)}{p(x) * p(y)} \quad (35)$$

avec: $p(x, y)$ est la probabilité que les deux mots x et y apparaissent ensemble, $p(x)$ et $p(y)$ représentent la probabilité d'apparition de x et y respectivement.

Une information mutuelle élevée entre un thème et un terme est le signe d'un lien de pertinence fort entre ces deux éléments.

5. Le gain d'information :

Egalement appelé information mutuelle moyenne [67] permet, tout comme la mesure d'information mutuelle, de quantifier le lien existant entre un terme et un thème mais ne prend pas seulement en compte l'influence qu'a l'apparition d'un terme sur un thème, mais tiens compte également de sa non apparition. La mesure de gain d'information se calcul de la façon suivante :

$$IG(w,T) = \sum_T \sum_w p(w,T) \log \frac{P(w,T)}{P(w)P(T)} \quad (36)$$

6. Mesure de diffusion de Church :

Church dans [68] emploie la notion de « la diffusion de mot » pour montrer que les termes candidats s'adaptent. Autrement dit, la probabilité de changement d'occurrences de ces termes est basée sur le contenu et le contexte du document.

Church divise chaque document en deux parties : la première est la partie historique (History) la seconde est la partie teste (test), il prouve que : quand un mot candidat apparaît dans la partie historique, la probabilité d'apparition de ce mot dans la seconde partie augmente de manière significative. Si un mot est utilisé le long du texte, alors il se produira dans les deux parties du document. Avec l'idée de Church de segmenter un document en 2 parties, la mesure de pertinence (IM) devient :

$$\left. \begin{aligned} M^{DF_2}(w, t) &= \log \frac{p(w, t)}{p(w) * p(t)} \\ M^{DF_2}(w, t) &= \log \frac{DF_2(w, t)}{DF_2(w, db)} + \log \frac{| db |}{| t |} \end{aligned} \right\} \quad (37)$$

Avec :

$DF_2(w,t)$: le nombre de documents du thème T contenant le mot w dans les deux parties du texte.

$DF_2(w,db)$: le nombre de documents dans le corpus contenant le mot w dans les deux parties du document. $|t|$: le nombre de documents dans le corpus dont le thème principal est T. $|db|$: le nombre de documents dans le corpus.

- *Les méthodes de détection des thèmes dans un corpus:*

Le second paramètre dans cette approche est la méthode de détection de thème, qui définit la façon dont les termes présents dans les textes sont exploités.

Plusieurs travaux se sont intéressés à la détection automatique de thèmes en s'appuyant soit sur le repérage d'indices linguistiques [70], [71], soit sur des notions telles que la cohésion lexicale [6], [72], [73], ces travaux réalisent pour la plupart simultanément la caractérisation de thèmes et une segmentation du texte en fragments thématiquement cohérents.

TextTiling [6] réalise le découpage d'un texte en groupes de paragraphes successifs portant sur un même thème, en se basant sur une mesure de similarité lexicale entre les séquences consécutives des mots. Tous les 20 mots, l'algorithme calcule la « ressemblance » entre les listes des 100 mots apparaissant à gauche et à droite du point de focus. Un minimum local de cette mesure est considéré comme l'indice d'une zone de changement thématique et une frontière est alors définie à l'emplacement de la limite de paragraphe le plus proche. Les mots ayant joué un rôle important dans le maintien à une valeur élevée de la mesure de cohérence lexicale entre deux minima sont mis à profit pour caractériser le thème de la région considérée.

Le couple de traitements *SEGCOHLEX* / *SEGAPSITH* présenté dans [72] est pour sa part basé sur une idée similaire mais réalise une segmentation de données beaucoup plus fine, la mesure de cohésion lexicale étant calculée pour chaque occurrence de mot et à l'aide de fenêtres de 10 mots à gauche et à droite du point de focus. Cet indice faisant usage d'une quantité de mots bien plus faible que le précédent, un corpus de 45 millions de mots est employé au préalable et exploité dans la mesure de proximité lexicale pour réaliser une première segmentation des textes étudiés (module *SEGCOHLEX*). À l'aide de cette segmentation, le système définit des « signatures thématiques » qui forment la base d'une seconde segmentation (module *SEGAPSITH*) et fournissent une caractérisation indirecte mais — contrairement à *TextTiling* — systématique des thèmes détectés.

Nous allons étudier trois différentes méthodes différentes à savoir, le classifieur *TF*IDF* considéré comme la plus ancienne des méthodes de détection des thèmes, la méthode unigramme une des plus performante, ainsi qu'une très récente méthode à savoir Les Machines à Vecteur Support '*SVM*'.

a. Le classifieur *TF*IDF* :

La classifieur *TF*IDF* [73] est la référence dans le domaine, c'est un des premiers modèles à avoir été développé. Son principe est le suivant : un terme sera d'autant meilleur pour représenter le

contenu d'une classe s'il est à la fois fréquent dans cette classe et rare dans l'ensemble des classes à analyser.

Dans ce modèle, chacun des éléments des vecteurs de document est pondéré par un facteur reflétant la proportion de thèmes dans lequel le terme est présent. Ensuite, une distance cosinus est calculée entre le vecteur représentant le document et celui de chacun des thèmes. Le thème correspondant à la distance la plus faible sera celui affecté au document.

$$sim(D_j, D_i) = \frac{\sum_{k=1}^n d_{jk} d_{ik}}{\sqrt{\sum_{k=1}^n (d_{jk})^2 \sum_{k=1}^n (d_{ik})^2}} \quad (38)$$

b. Le modèle unigramme :

Dans le modèle unigramme [74], une distribution de probabilités des mots est calculée pour chaque thème. Ensuite, la probabilité de chaque thème est calculer selon l'équation 39 et le thème correspondant à la probabilité a posteriori la plus élevée sera le thème retenu.

$$P(T_j / W_1^N) = \frac{P(T_j) P(W_1^N / T_j)}{\sum_{k=1}^J P(T_k) P(W_1^N / T_k)} \quad (39)$$

c. Les machines à vecteur support (SVM):

La méthode SVM [75] oppose le thème en cours de traitement à l'ensemble des autres thèmes. Sur une représentation dans un espace donné, des documents du thème ainsi que de l'ensemble des autres documents, on recherche l'hyperplan optimal séparant les deux ensembles de données. L'originalité des (SVM) est qu'elles cherchent à maîtriser l'erreur en généralisation. Pour traiter le cas ou plus de deux thèmes sont utilisées, une étape de recombinaison des scores est ensuite nécessaire pour retrouver le thème d'un document donné.

3. La mise en œuvre du Système de Recherche d'Information Spectrale

La conception de notre Système de Recherche d'Information Spectrale (SRIS) s'appuie sur deux axes : le premier concerne une modélisation des documents '**représentation thématique spectrale**' qui consiste à modéliser un document par un ensemble de signaux reflétant les variations thématiques présentes dans ce document. Cette approche permettra d'importantes améliorations dans la compression et la comparaison de grandes masses de données textuelles à différents niveaux de résolution.

Le second axe qu'on présente concerne la modélisation des requêtes '**représentation spectrale des requêtes**' qui permettra dans un système de recherche d'informations de prendre en compte la position relative aux mots de la requête dans les documents dans le calcul de la pertinence de ces derniers.

3.1 Modélisation thématique spectrale des documents

Les principales étapes de la modélisation spectrale des documents (figure 19), se résument principalement en deux phases à savoir : une première phase qui réalise le traitement du corpus, l'analyse statistique et l'élimination des mots vides, la seconde phase consiste à la construction des signaux thématiques.

Phase 1 : La première étape consiste à éliminer les mots vides présents dans les textes (articles, pronoms, prépositions,...) (sous forme d'anti-dictionnaire, En français, les mots « de », « un », « les », ... sont les plus fréquents. En anglais, ce sont les termes tels que « of », « the »,

Phase 2 : Construction des signaux thématiques

Nous allons modéliser un texte par un ensemble de signaux qu'on a appelé *signaux thématiques*. Le signal de flux thématique reflète la variation de l'intensité de la discussion liée à un thème donnée à travers le texte.

Les thèmes fortement discutés dans un texte tendent à se reproduire dans la majeure partie du texte. Cependant, les thèmes discutés moins fortement concernent des parties limitées du texte, ou qui sont souvent interrompus, décalant de ce fait le centre de la discussion. Le vocabulaire de ces thématiques est moins présent ce qui indique une présence plus faible du thème dans le texte.

Ainsi, construire un signal de flux thématique dans un texte revient à représenter la pertinence du vocabulaire lié à ce thème dans le texte.

Une représentation idéale du flot thématique devrait permettre la présence de plusieurs thèmes simultanément à un point donné dans un texte, cette flexibilité peut être réalisée en représentant indépendamment les thèmes les uns les autres.

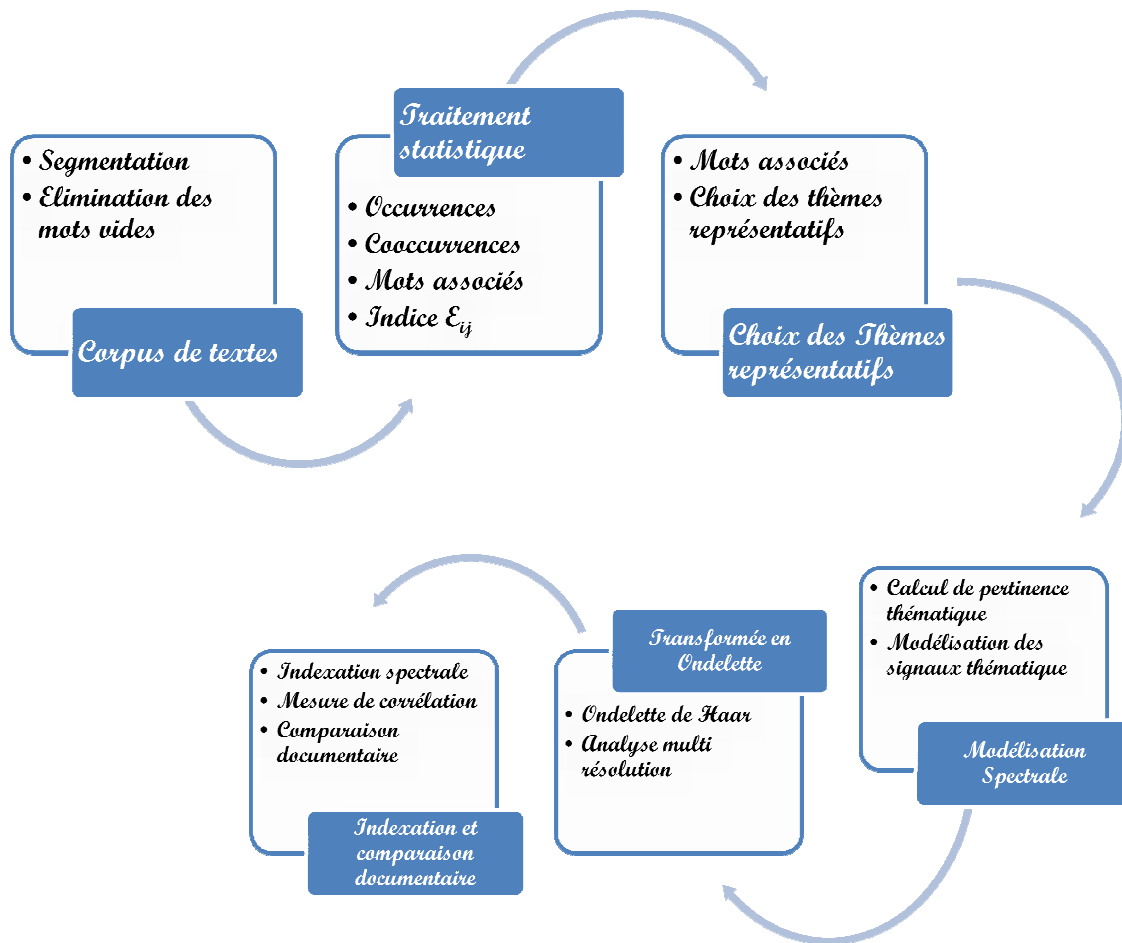


Figure 18. Processus des différents traitements du corpus en vue d'une modélisation spectrale des documents.

3.1.1. Algorithme de construction des signaux thématiques

Supposons qu'on puisse mesurer la pertinence d'un mot w par rapport à un thème T dans le texte en utilisant une mesure $P(T, w)$. Si on trace la courbe qui représente ces différents niveaux de pertinence des mots par rapport à ce thème dans leur ordre d'apparition dans le texte, on peut observer le flot de ce thème (sa variation de pertinence). (Exemple figure. 17)

L'algorithme de construction du signal parcourt le texte et pour chaque position du terme w , il attribue pour la position correspondante dans le signal la valeur de sa pertinence thématique. Ce processus de construction du signal thématique est décrit dans l'algorithme 1.

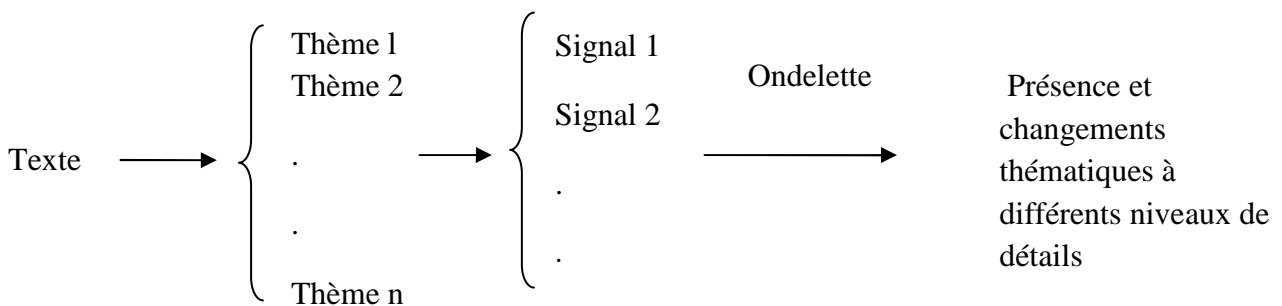


Figure 19. Processus de construction des signaux thématiques

Nous allons commencer par décrire le processus de construction du signal thématique, ce processus va générer des signaux thématiques basiques, qui seront décomposés par la suite et analysés à différents niveaux de détails appelés niveaux de multi résolution. Dans la partie suivante nous allons étudier et analyser ces niveaux de multi résolution dans un système de comparaison documentaire.

Algorithme 1 : Construction du signal thématique S_t pour un thème t dans un document D

Entrée : Document D ;

$||D||$ = le nombre des positions des mots dans le document D ;

L = liste des thèmes T ;

M = liste des mots candidats avec leur pertinence thématique $P(T, w)$;

Sortie : signal thématique S_t pour tout thème t dans le document D ;

Pour (position-texte=0 ; position-texte < ||D|| ; position-texte++) faire

$$S_t [\textit{position-signal}] = P(t, w) ;$$

Position-signal ++ ;

Notons que la qualité du signal est sensible à la mesure de pertinence utilisée dans sa construction. Dans la partie suivante, nous allons appliquer cet algorithme sur un exemple. Et déterminer notre choix de méthode de sélection et détection des thèmes dans ce corpus de documents.

3.1.2 Expérimentation : la modélisation spectrale

3.1.2.1 Les données

Pour cette première expérimentation nous allons utiliser la base de données LISA (Library and Information Science abstracts). Un index international conçu pour les professionnels de l'information. LISA couvre 23 catégories, on peut citer : la recherche en science de l'information, l'intelligence Artificiel, management, recherche d'information, Télécommunications, ainsi que des articles de recherche de 'IRWI' : Information Research Watch International Database. Elle couvre près de 500 périodiques (depuis 1969) sur une soixantaine de pays et en 20 langues différentes.

Artificial intelligence
Book reviews
Computer science applications
Information centres
Information management
Information science
Information Storage
Information technology

Tableau 6. Exemple des thématiques de la base LISA

3.1.2.2 Processus de l'expérimentation

Dans ce qui suit nous allons détaillées les étapes de traitement et d'indexation du corpus, en verra par la suite les étapes de calcul de pertinence thématique et de modélisation spectrale des documents. Pour cela nous allons utiliser l'algorithme 1 de construction de signaux thématique décrit dans la section 1.1.1 précédente.

3.1.2.2.1 Traitement du corpus

Le corpus de notre expérimentation contient 6004 résumés d'articles sous format texte, nous allons commencer par appliquer à notre corpus textuel l'extracteur terminologique du logiciel SAMPLER qui va établir un lexique de termes avec leurs occurrences dans le corpus. Pour cela SAMPLER [29] utilise un ensemble de grammaires et de dictionnaires ainsi que des listes de mots vides.

L'extracteur terminologique de Sampler est composé d'un lexique, d'une liste de patrons et de règles de désambiguïsation contextuelles. Le lexique est constitué d'unitermes étiquetés par leur catégorie grammaticale (adjectif, verbe, préposition, article, nom, conjonction, ...).

Les paramètres d'extraction sont fixés comme suit :

Fréquence minimale d'extraction : 4

Nombre de mots par expression : 2

MOT	FREQ	vis	#
aacr 2	45	*	1
university abstract 2seel	5	*	2
university abstract journals	8	*	3
university abstracting journals	10	*	4
university abstracting services	13	*	5
updated ver abstracts of papers	13	*	6
urban areas abstracts of the individual	15	*	7
urgent need academic institutions	9	*	8
useful info academic librarian	32	*	9
useful tool academic libraries	204	*	10
user demand academic staff	4	*	11
user educat academy of sciences	13	*	12
user group acquisition policies	15	*	13
user needs adult education	30	*	14
user satisf adult population	4	*	15
user servic adult services division	6	*	16
user studie advanced education	10	*	17
user traini advisory committee	13	*	18
users of in aesthetic education	5	*	19
usrr academ african institute	9	*	20
valuable in african library association	5	*	21
variety of agricultural information	24	*	22
various asp agricultural librarians	8	*	23
various cou agricultural libraries	17	*	24
various methods	11	*	1524

Index - trié alphabétiquement : affichage par paquet de 1500 lignes

Figure 20. Extrait de l'index réalisé sous SAMPLER

SAMPLER permet ensuite de visualiser l'ensemble de ce lexique sous forme de représentation graphique après une étape de clustérisation correspondant à un regroupement des mots constitutifs du lexique par familles homogènes. Au sein d'un *cluster*, les mots sont reliés entre eux par des liens plus ou moins forts calculés en fonction des cooccurrences relatives des mots dans les textes.

Cependant les auteurs ne précisent pas comment employer ces mesures pour détecter et caractériser les thèmes présents dans un document, ou quels sont les mots du vocabulaire associés à ces thèmes. Le processus de sélection étant décrit partiellement, il s'avère difficile à implémenter dans la pratique.

Dans notre modèle nous allons utiliser la méthode des mots associés, basée sur les occurrences et les cooccurrences des mots candidats dans le corpus, afin de pouvoir déterminer l'ensemble des thèmes représentatifs dans le corpus.

- *Méthode des mots associés (co-word analysis)*

La méthode des mots associés est une technique développée au début des années 80 par le CSI (Centre de Sociologie de l'Innovation) de l'Ecole des Mines et le CDST (Centre de Documentation Scientifique et Technique) du CNRS ⁶ [82]. Cette méthode est née des problèmes posés par l'application des méthodes statistiques usuelles aux données utilisées par la sociologie des sciences. Elle propose d'identifier les mots les plus fortement associés.

L'approche des mots associés concerne l'analyse de cooccurrences de termes indexés et permet l'expression des convergences d'intérêts entre acteurs, actants, de faire apparaître des agrégats, des agencements, de les hiérarchiser selon la nature et la taille du corpus. L'association de deux mots-clés se mesure en fonction de leur nombre d'apparitions communes dans les documents qu'ils indexent. Des récents travaux [83] illustrent la pertinence de cette méthode.

Nous avons utilisé la méthode des mots associés pour calculer l'intensité des liens qui existe entre les mots clés de l'index et générer un classement des ces mots. Cette intensité nous à permettait d'identifier les thèmes représentatifs de notre corpus.

Afin de réaliser ces calculs d'intensité on a fais appel au module de clusterisation de l'outil SAMPLER [32], qui utilise l'information sur la fréquence et la cooccurrence des termes dans le corpus, mais cette dernière avantage les termes qui co-occurrent un grand nombre de fois. Mais l'emploi d'un indice statistique permet de normaliser cette mesure.

Ils existent plusieurs méthodes de calcul d'un coefficient d'association. Le coefficient utilisé dans la méthode des mots associés implantée dans SAMPLER est l'indice d'équivalence E_{ij} défini par :

⁶ CNRS : Centre National de la Recherche Scientifique

$$E_{ij} = \frac{C_{ij}^2}{C_i * C_j}$$

Où :

(40)

- C_{ij} le nombre de co-occurrences des mots i et j.
- C_i le nombre d'occurrence du terme i.
- C_j le nombre d'occurrence du mot j.

On va ainsi mesurer l'intensité de l'association existante entre les mots i et j. Si $E_{ij} = 1$, cela signifie que la présence d'un terme entraîne la présence de l'autre, si $E_{ij} = 0$, cela signifie que la présence d'un terme exclu la présence de l'autre dans le corpus.

production methods
subject libraries
cooperative activities
library networks
technology transfer
local authorities
automation project
eastern europe
documentation center
containing information
role of libraries

Tableau 7. Extrait de l'ensemble des thèmes représentatifs du Corpus LISA

3.1.2.2.3 Exemple de calcul de pertinence thématique :

Pour chaque texte du corpus on va modéliser un signal représentant la présence et le changement de chaque thème de la liste des thèmes représentatifs extraite lors de l'étape précédente.

Nous allons utiliser la mesure d'Information Mutuelle qui quantifie la pertinence existante entre un mot m et un thème T. Elle est évaluée de la façon suivante :

$$\begin{aligned}
 IM(T, m) &= \log \frac{p(T, m)}{p(T) * p(m)} \\
 IM(T, m) &= \log \frac{f(m, T)}{f(m, C)} + \log \frac{\|C\|}{\|T\|}
 \end{aligned}
 \tag{41}$$

Où :

$f(m, T)$ est la fréquence du terme m dans les documents du thème T

$f(m, C)$ est la fréquence de terme m dans l'ensemble du corpus C

$\|T\|$ est le nombre de documents du thème T

$\|C\|$ est le nombre de documents dans le corpus C

Prenant par exemple le texte suivant de notre corpus, on va calculer la pertinence de ses mots selon la thématique « Library network ».

Exemple 1: le document 651 du corpus, sous le titre ‘Analysis of Library Networks.’

“The emergence of library networks is discussed. Management issues involving network structure, economics, and applications of computer technology are considered. A variety of library network models are reviewed, including both analytical and simulation models. Typical problems in applying models to the analysis of library networks are discussed.”

Après élimination des mots vides de ce texte nous avons calculé les pertinences thématiques des différents mots du texte selon la thématique ‘Library network’ et dans leur ordre chronologique.

termes	Pertinence thématique
emergence	1,17098566035656
library networks	1,90603103013786
discussed	0,498887802
management	2,04604692374826
involving	0,923831045475437
network structure	1,50197887939799
economics	0,703624242926058
computer technology	1,39283440997292
Library network models	2,34707691941225
analytical	0,884678921513289
simulation models	2,34707691941225
typical problems	2,34707691941225
applying	1,02485762467833
models	1,06264618556773
analysis	0,258940830711694
library network	1,90603103013786

Tableau 8. Pertinence des termes du texte 1 par rapport à la thématique « Library network »

3.1.2.2.4 Construction des signaux thématiques

On a défini dans le signal thématique comme étant la modélisation de la variation de la discussion d'un thème donné dans le texte. Cette modélisation prendra en considération les fréquences des mots ainsi que leurs positions dans le texte, donc leurs ordres d'apparition dans le texte. La suppression des mots vides dans le document peut éventuellement perturber les

positions des mots restants dans le document et de se faire perturber la modélisation des signaux thématiques.

Lors de la modélisation en signaux, les mots qui se produisent dans un certain segment de texte avant la suppression des mots vides, doivent respecter et apparaître dans le même segment après la suppression des mots vides. Nous proposons de remplacer les positions des mots vides, avant de les supprimer, par un mot VIDE qui peut être sans risque éliminer lors de la modélisation du signal correspondant.

Le processus de construction du signal thématique se déroule ainsi :

Soit un plan orthogonale (X, Y) et un segment donné dans un texte du corpus :

Le processus parcourt le segment dès le début, la position du premier mot du texte dans le signal est *Signal* [0]. Pour chaque mot dans le segment, pour sa position correspondante dans le signal, le processus attribue sa valeur de pertinence thématique selon l'algorithme 1. Si le mot est VIDE (utilisé pour remplacer les positions des mots vides) il sera rejeté. Sur l'axe des abscisses du plan sera représenté les positions des mots dans le segment et sur l'axe des ordonnées leurs pertinences correspondantes.

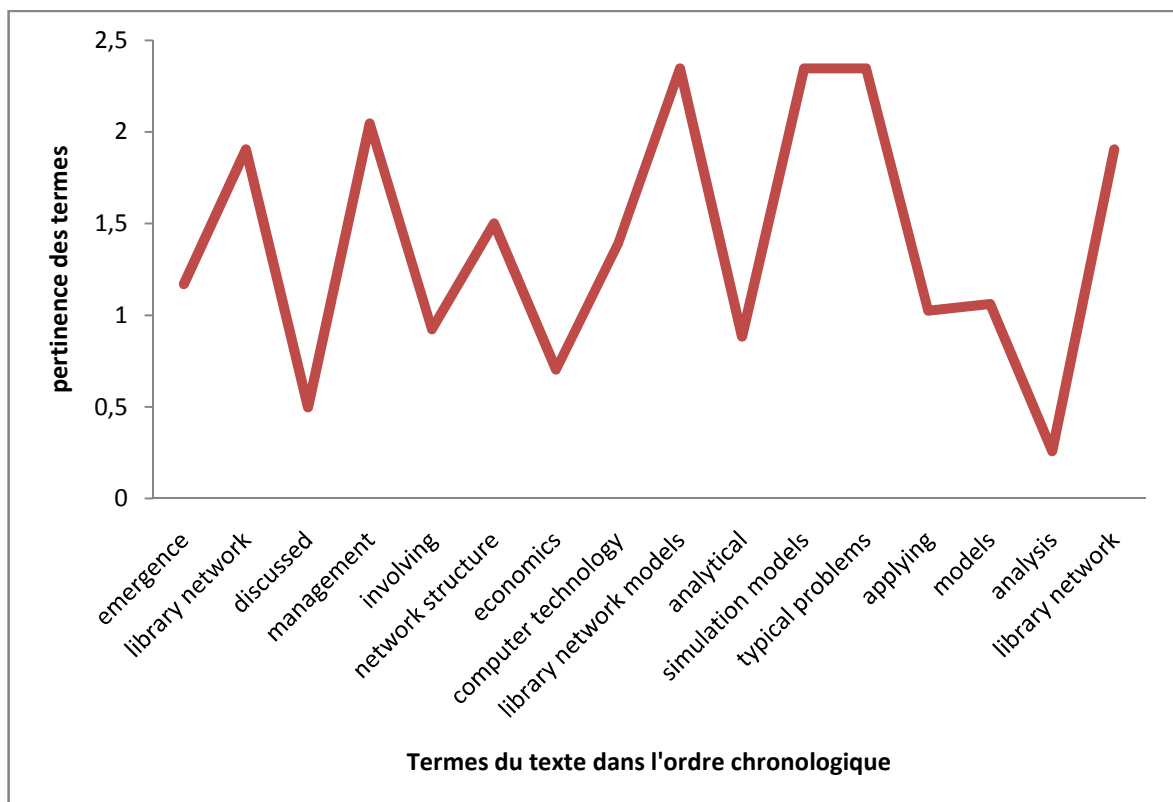


Figure 22. Signal thématique 'Library network' dans le texte 1

Considérons un second document et suivant le même principe de modélisation, nous avons obtenue les valeurs de pertinence (tableau 9), ainsi que la représentation thématique (figure 23) qui permet de modéliser le thème « Library network » dans ce second document.

Exemple 2: Le document 4399 du corpus, sous le titre ‘library networks in the federal republic of Germany’. “*Describes cooperative and network activities in West Germany. An attempt is made to define the requirements of network data bases. The following topics Are discussed' data base structure, hierarchical relationships, authority Files, local files, local data, access points, and data manipulation control.*”

termes	Pertinence thématique
cooperative	2,37425083
network activities	3,34707692
west germany	2,37447418
attempt	2,3638367
define	2,38806053
requirements	2,36606426
network data bases	3,34707692
topics	2,35442986
discussed	2,36126132
data base structure	2,84707692
hierarchical relationships	3,34707692
authority Files	2,45818803
local files	2,84707692
local data	2,47207692
access points	2,43798601
data manipulation control	3,34707692

Tableau 9. Pertinence des mots du texte 2 par rapport à la thématique « Library network »

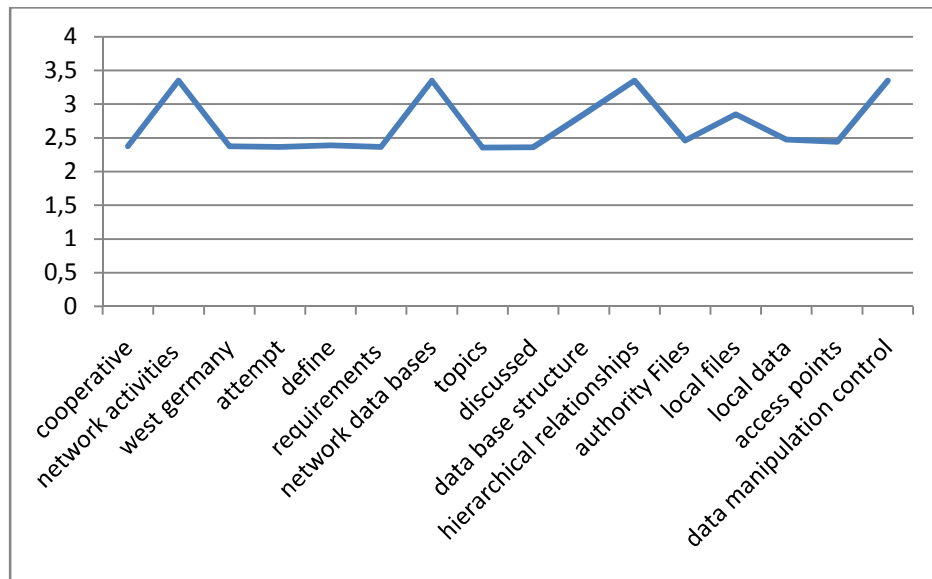


Figure 23. Signal thématique 'Library network' dans le texte 1

La représentation du flot thématique du thème « Library network » (figure 22) permet d’observer des variations des valeurs de pertinence moins importantes que celles observées dans le second document (figure 23). Ces variations indiquent une présence plus faible des termes du vocabulaire correspondant au thème, donc une présence relativement faible du thème ‘Library network’ dans ce document.

3.1.2.2.5 Niveaux de multi résolution dans un système de comparaison documentaire :

Afin de réaliser une comparaison documentaire et calculer la similitude entre documents, on doit calculer la pertinence entre leurs différentes représentations thématiques. Ce qui revient à comparer leurs représentations spectrales.

Pour calculer la similarité entre deux signaux représentant deux documents, on fera appel à la mesure de corrélation défini ainsi :

On suppose qu'on a les deux signaux suivants : $X(x_1, x_2, \dots, x_n)$ et $Y(y_1, y_2, \dots, y_n)$. Pour calculer le coefficient de corrélation entre les deux signaux on applique la formule suivante :

$$Corr(x, y) = \frac{\sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}} \quad (41)$$

\bar{x} et \bar{y} : représentent successivement les moyennes des valeurs de X et Y.

La plus forte valeur du coefficient de corrélation est +1, elle indique une parfaite corrélation entre les deux signaux donc les deux documents sont très proches par rapport à la thématique étudiée. Tandis que la valeur de -1 indique une corrélation inverse qui indique qu'un des deux signaux est plus haut que le second, qu'on interprète par la forte présence du thème dans un document et une présence beaucoup moins faible dans le second document.

Un coefficient de corrélation de 0,65 et plus, indique une forte corrélation entre les deux signaux X et Y. Un coefficient entre 0,3 et 0 indique une faible corrélation ou une corrélation inexistante.

Après expérimentation cette première comparaison reste approximative et imprécise, nous devons exécuter d'autres comparaisons plus fines donc des comparaisons à d'autres niveaux de résolutions plus petits. Pour réaliser ceci nous ferons appel à la propriété de multi résolution des Transformées en ondelettes.

i. Construction des signaux à différents niveaux de détails :

L'approche qu'on présente ici permet d'inclure tous les termes du segment dans la construction des signaux thématique. La moyenne des deux valeurs consécutives dans un signal, lisse le signal et la différence entre ces deux paires souligne le changement local de la valeur de la pertinence. C'est deux opérations génèrent un signal transformé.

Etant donnée un segment de taille de 8 mots, en moyennent les deux paires consécutives du signal correspondant, on obtient un signal correspondant à un segment de taille de 4 mots (avec certains détails), similairement, pour chaque segment de taille 2^{i-1} , on utilise le signal du segment de taille 2^i . La différence entre paires souligne le changement local d'un thème et pour reproduire le signal original, il faut additionner le signal des moyennes et celui de détails.

Résolution	Moyenne/Approximation	Détails/Différence
8	[2 4 8 12 14 0 2 1]	
4	[3 10 7 1.5]	[-1 -2 7 0.5]
2	[6.5 4.25]	[-3.5 2.75]
1	[5.375]	[1.125]

Tableau 10. Transformée de Haar d'un signal de dimension 1.

Cette transformation (Figure 24) est appelée la Transformée en ondelette *de Haar*, la plus simple des Transformées en ondelettes, elle est utilisée dans plusieurs applications de traitement d'images et de compression de signal [77], [78].

Les différents niveaux d'approximations présents dans le tableau 8 reflètent le degré de changement du flot thématique à différentes tailles de segment, donc on peut étudier différentes vues du texte à différentes tailles de segment. Nous appelons ces niveaux : niveaux de multi résolution. La Transformée en ondelette est une méthode efficace pour obtenir ces niveaux de multi résolution. Cependant, elle exige que la taille du signal d'origine soit une puissance de 2.

Dans le cas contraire, certaines méthodes de traitement de signal proposent d'étendre le signal en le complétant avec une certaine constante ou en répétant une partie du signal. Dans le contexte du flot thématique, ajouter une valeur quelconque au signal risque de changer et perturber la vraie représentation thématique.

Cette condition de la taille du signal (puissance de 2) provient du fait que l'approche consiste à faire à chaque niveau la moyenne des deux éléments consécutifs dans le signal thématique. Pour remédier à ce problème, on va procéder comme suit : diviser le texte en segments de taille égale (en deux sous segments à la fois), donc à chaque niveau de résolution, tous les segments auront la même taille ou qui se rapproche d'une puissance de 2.

Par exemple, un document qui se compose de 9 mots. Peut être divisé en deux segments de 4 mots le premier et 5 mots pour le second. Chacun de ces deux segments sera divisé à son tour en 2 segments de 2-2 mots pour le premier. Et 2- 3 mots pour le second et ainsi de suite. Au niveau le plus basique, nous aurons 7 segments de taille de 1 mot chacun et un segment de taille de 2 mots.

Cette solution vas nous permettre de traiter des segments dont la taille n'est pas nécessairement une puissance de 2 à la différence de la moyenne de la Transformée en ondelette. Cependant, nous devons connaitre a priori la taille du texte.

9

4				5			
2		2		2		3	
1	1	1	1	1	1	1	2

Tableau 11. Tailles de segments à différents niveaux de résolution pour un texte de 9 mots

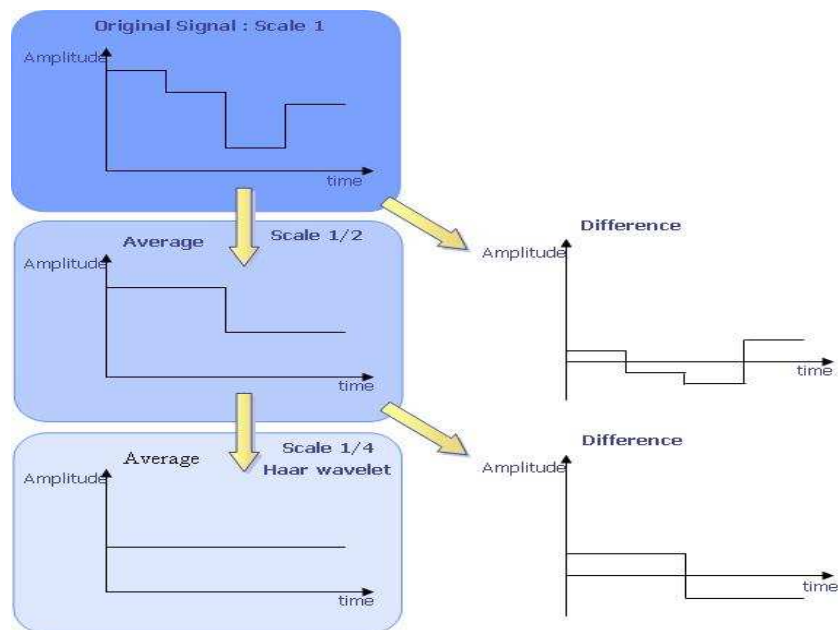


Figure 24 Propriété de multi-résolution de l'Ondelette de Haar

ii. Utilisation des niveaux de multi-résolution dans un système de comparaison documentaire

On pourrait comparer les différents niveaux de résolution générés par l'application de la Transformée en ondelette sur le signal d'origine (Figure 24) à un zoom qu'effectue un cameraman, en partant d'une vision globale d'une image (approximation grossière) pour se focaliser sur une zone d'intérêt en faisant apparaître tous les détails par rapport à la première image.

Les Ondelettes permettent de visualiser les données à multiples résolutions ou chaque résolution reflète une fréquence différente. Chaque étape de la Transformée en ondelettes produit deux ensembles de valeurs : un ensemble de moyennes et un second ensemble de différences (désignés généralement par les coefficients d'Ondelette).

L'idée de l'analyse multi résolution d'un signal consiste à le représenter par l'ensemble de ses approximations successives, ou chaque approximation est une version lissée de la précédente. Les approximations successives sont présentées à différentes résolutions d'où le nom de multi résolution.

Nous allons appliquer la Transformée en ondelette de Haar sur l'ensemble des signaux construits dans l'étape précédente afin d'exploiter sa priorité de multi résolution pour analyser les textes du corpus le plus en détails possibles et calculer le niveau de résolution où les documents seront le plus similaires possible, en procédons comme suivant :

Si on désigne le signal thématique S par le vecteur suivant :

$$S = \{s_0, s_1, \dots, s_n\} ,$$

La figure (25) se traduira ainsi :

$$S = (a, d)$$

'a' représente le vecteur des moyennes et 'd' le vecteur des détails de la décomposition sur la base d'Ondelette de Haar. Les moyennes et les détails sont reliés au signal par les formules suivantes :

$$a_k = \frac{S_{2k} + S_{2k+1}}{\sqrt{2}} \quad b_k = \frac{S_{2k} - S_{2k+1}}{\sqrt{2}} \quad \frac{1}{\sqrt{2}} : \text{Facteur de normalisation}$$

Une fois la première transformation réalisée, le processus sera répété sur les nouvelles valeurs des moyennes jusqu'à épuisement des termes.

3.1.3 Résultats de l'analyse multi résolution

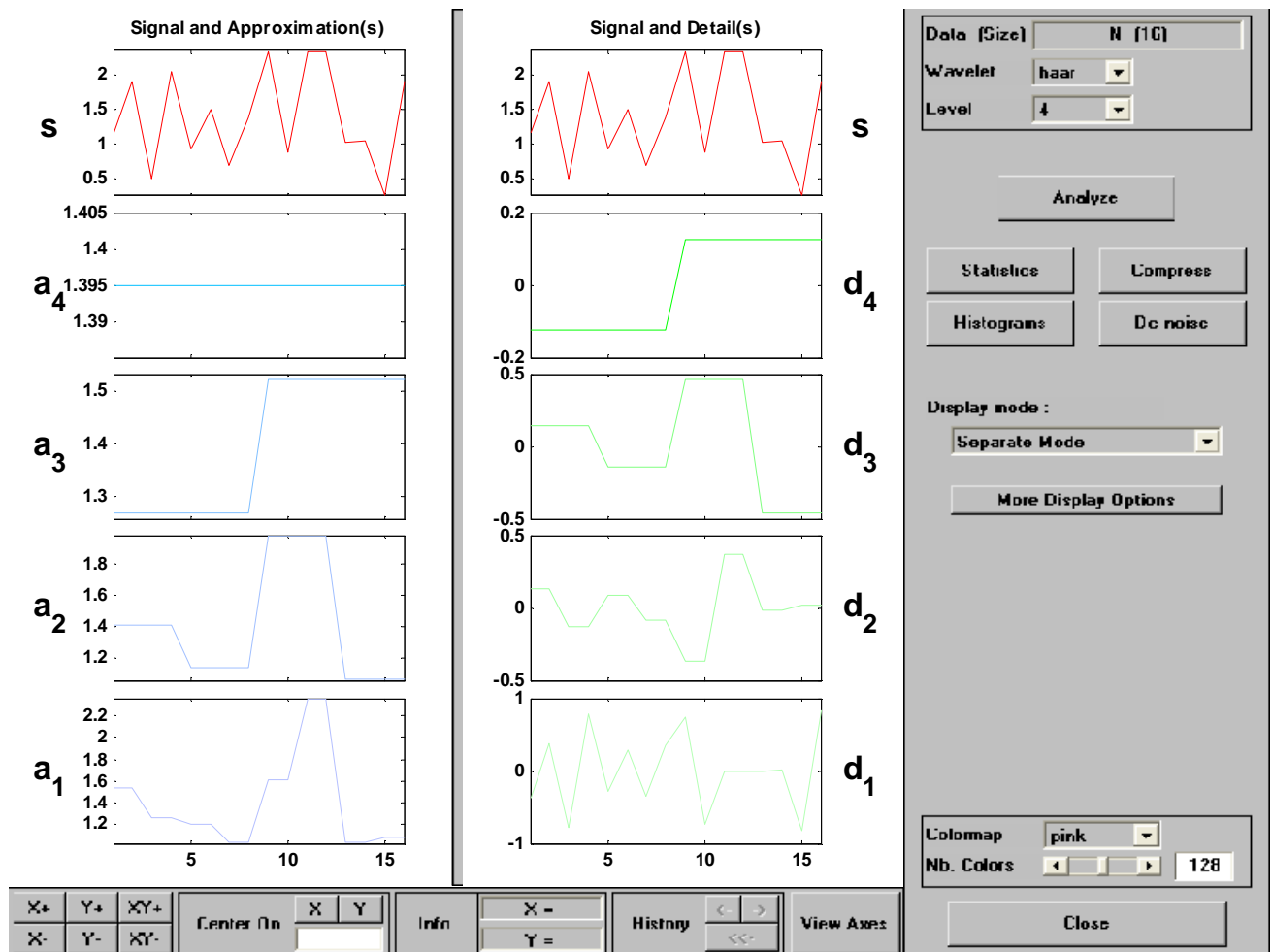


Figure 25. Transformée de Haar du signal 'Library network' correspondant au texte 1 (coefficients d'approximations et de détails)

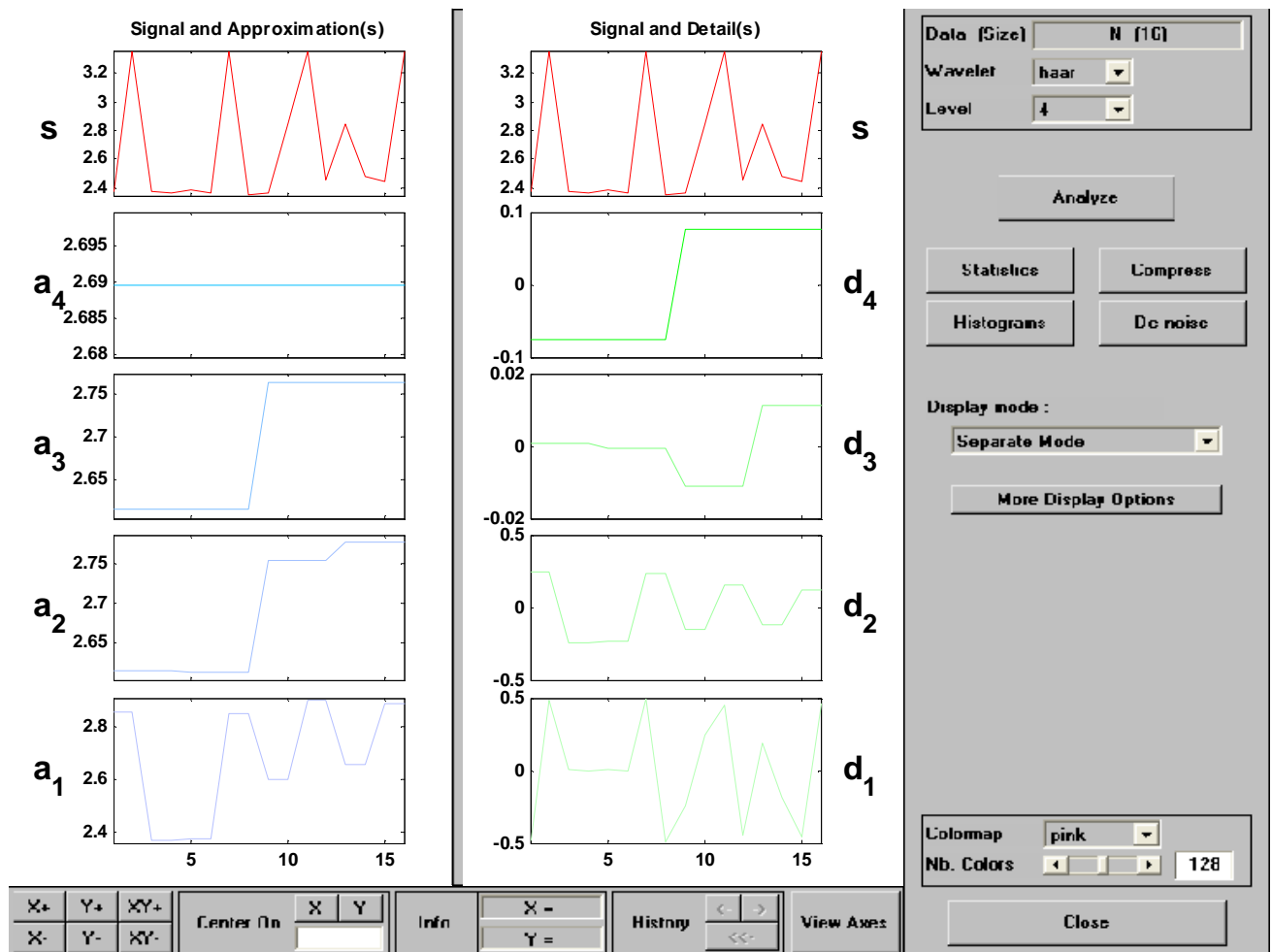


Figure 26. Transformée de Haar du signal 'Library network' correspondant au texte 2

La figure 27 présente la faible similitude entre les deux signaux représentant les deux textes au premier niveau de résolution. La corrélation peut être ainsi mesurée sur les différents niveaux de résolution ; sur les niveaux les plus bas, les coefficients de détails en valeur absolue sont grands, par conséquent, la valeur de corrélation est petite, progressivement ces coefficients de détails diminuent et la valeur de la corrélation augmente pour indiquer la similarité entre les deux signaux.

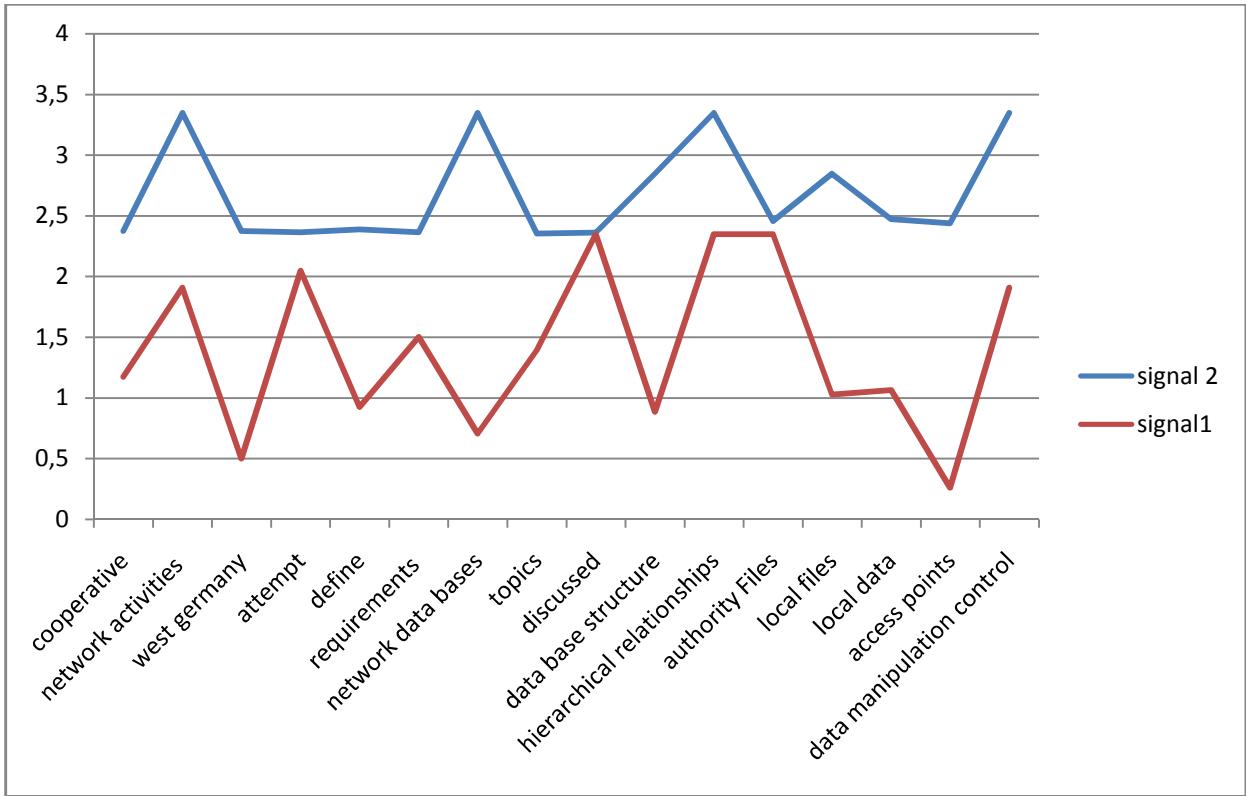


Figure 27. Faible similarité entre les deux signaux thématique au niveau de similarité 0.

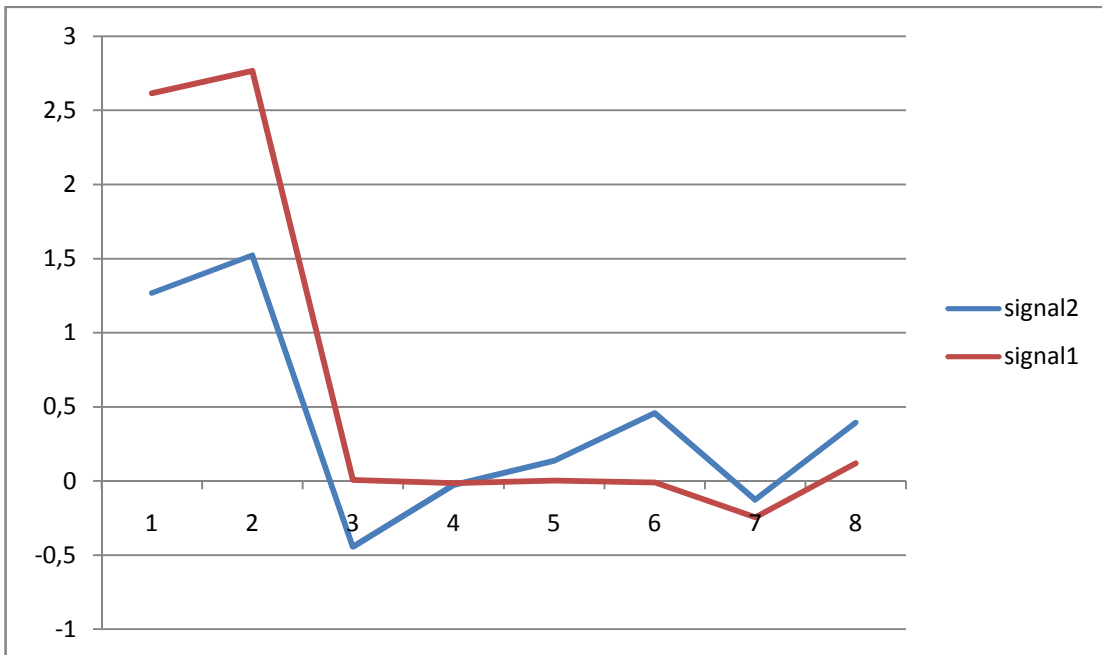


Figure 28. Similarité modérée entre les deux signaux thématique au niveau de similarité 1.

Dans le tableau 12 on présente les différentes valeurs de similarité entre les deux signaux sur les différents niveaux de détails de l'analyse de multi résolution de l'Ondelette de Haar.

Niveaux de résolution (Haar)	corrélation
0	0,18
1	0,25
2	0,28
3	1
4	1

Tableau 12.valeurs de corrélation entre les deux signaux S1 et S2 à différents niveaux de détails.

Deux signaux sont corrélés si leurs valeurs diminuent et augmentent au même moment. Selon le tableau 12, les deux signaux se corrélient à partir du niveau 3 avec une faible corrélation sur les niveaux inférieurs. Les deux documents traitent le même thème mais avec une importante variation du vocabulaire d'où la différence de la similarité aux premiers niveaux de résolutions.

3.2 Représentation spectrale des requêtes

3.2.1 Introduction

Dans les modèles classiques de la recherche d'information, le critère de sélection d'un document est fondé sur l'appartenance (resp. la fréquence) d'un mot de la requête pour le modèle booléen (resp. le modèle vectoriel) dans ce document. Ces modèles procèdent avec une **approche globale** sur l'influence des occurrences d'un terme t sur la pertinence d'un document par rapport à une requête. Ce qui revient à dire que la distribution des termes de la requête n'intervient pas dans le calcul de la pertinence d'un document.

Cependant, le sens général d'un texte ne dépend pas seulement du vocabulaire employé mais aussi de la position relative à chaque terme utilisé et donc de la distribution de ces termes. C'est pourquoi les méthodes de proximité se basent sur une **approche locale** dans le sens où elles

modélisent la distribution et la proximité des termes de la requête dans le document pour le calcul de la pertinence de ce dernier. Dans ces méthodes, le score de pertinence d'un document dépend des intervalles contenant les termes de la requête présents dans le document.

3.2.2 Modélisation Spectrale des requêtes

Dans notre système de recherche nous allons faire correspondre à chaque mot de la requête une modélisation spectrale qui va prendre en considération à la fois ces positions ainsi que sa fréquence d'apparition dans le texte. La pertinence d'un document dépendra alors de la combinaison des représentations spectrales de tous les mots de la requête. Cette modélisation prendra la forme d'un signal : *signal de mot*.

Chaque signal de mot possède deux propriétés : l'amplitude, qui est reliée à la fréquence du mot dans le document et la phase, qui est reliée à la position du mot dans le document.

Pour une recherche d'information nous allons combiner l'amplitude et la phase des signaux des mots de la requête afin de calculer la pertinence d'un document donnée. De cette manière nous pouvons observer quels documents ont une occurrence élevée des mots de la requête et quels documents ont les mots de la requête en proximité étroite.

3.2.3 Processus de comparaison spectrale document /requête

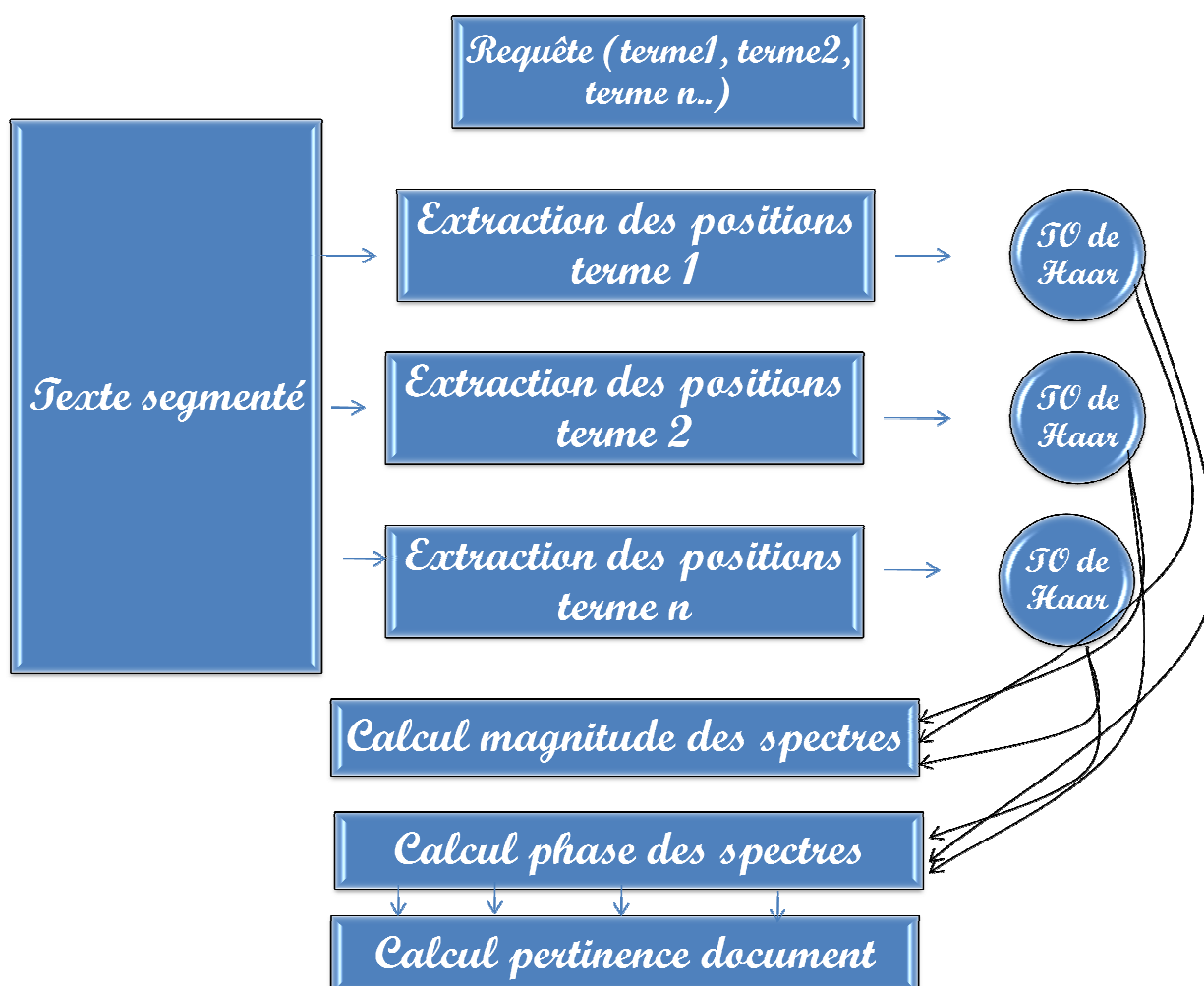


Figure 29. Processus de comparaison spectrale document /requête

3.2.3.1 Construction des signaux des mots d'une requête

Le signal d'un mot appartenant à la requête, correspond à une séquence ordonnée des occurrences du mot selon ses apparitions consécutives dans le document, autrement dit, le signal du mot représente le flot d'un mot spécifique dans un document.

Afin de normaliser la taille des signaux et réduire par conséquent les calculs nécessaires pour évaluer la pertinence des documents, nous allons utiliser une fenêtre de segmentation du texte dont le principe est le suivant :

Nous allons supposer que la taille du signal du mot est de ' B composantes' une puissance de 2. On procède à la segmentation du texte en plusieurs segments de taille égale. Un texte dont la taille est de W mots, sera segmenté en X segments de taille (W/B) .

Le signal S d'un mot w dans un document d sera représenté par :

$$S_{w,d} = [f_{d,w,0}, f_{d,w,1}, f_{d,w,2}, \dots, f_{d,w,n-1}]$$

Soit l'exemple suivant :

« *La politique de civilisation, la France veut la faire en/engageant une nouvelle politique de l'immigration qui permette la maîtrise/en commun des flux migratoires par le pays d'accueil et /les pays de départ. Elle veut la faire avec une/politique d'immigration décidée ensemble et gérée ensemble. Avec une politique /fondée sur les quotas fixés en fonction des capacités d'accueil/et d'intégration qui évite tous les drames humains, tous les /rejets, toutes les exclusions liés à une immigration non maîtrisée. /Nicolas Sarkozy, 2007.*

La taille du texte de cet exemple est de 80 mots, on le partitionne en 8 segments de taille de 10 mots chacun, étant donnée qu'on a choisit que la taille du signal de chaque mot de la requête sera de 2^3 composantes.

On choisit de construire le signal du mot '*Politique*' appartenant à une requête exemple, dans le premier segment du texte le mot apparaît **1 fois**, dans le deuxième segment il apparaît **1 fois**, en suivant le même processus, le signal du mot *politique* sera égale à :

$$S_{politique, d} = [1 \ 1 \ 0 \ 0 \ 2 \ 0 \ 0 \ 0]$$

Remarque : on remarque que si on considère le texte en entier comme un seul segment, on obtient un signal avec une seule composante qui correspondra à la fréquence du mot politique (fréquence égale à 4), dans le tout le document. (Modèle vectoriel classique).

3.2.3.2 Pondération des valeurs des composantes du signal

Le modèle vectoriel utilise lors de l'indexation des documents différentes mesures de pondération afin de réduire l'impact de certaines propriétés des documents et/ou des mots. Ces propriétés sont également présentes dans la représentation spectrale, donc on doit faire appel à des mesures de pondérations.

Dans le cadre de notre approche, on fera appel à la mesure « $Tf*Idf$ », (se reporter à la section 1.1.1 du chapitre 2). La mesure du nombre d'occurrence d'un terme dans une collection ne permet pas de capturer sa spécificité. Or un terme commun à de nombreux documents est moins utile qu'un terme commun à peu d'entre eux. C'est que motive la combinaison des mesures tf et idf . En bref, tf mesure l'importance d'un mot dans un document, tandis qu' idf mesure sa spécificité dans une collection.

1.6.5.7 Application de la Transformée en ondelette :

L'étape suivante du processus de la recherche d'information spectrale consiste à appliquer une Transformée en ondelette sur les signaux des mots de la requête construits dans l'étape précédente, afin de pouvoir comparer les spectres des mots de la requête et calculer la pertinence de chaque document par rapport à la requête.

Le signal transformé $Tf = [tf_0, tf_1, tf_3, \dots, tf_n]$

1.6.5.8 Calcul de pertinence document/requête:

Afin de calculer la pertinence d'un document par rapport à une requête, nous allons combiner les composantes des spectres des mots de la requête.

Pour l'utilisateur, un document est considéré comme pertinent par rapport à une requête s'il contient une occurrence élevée des termes de la requête, dans le cas de notre modèle cela implique une amplitude élevée du spectre et une proximité des termes de la requête, donc une phase similaire des spectres des mots de la requête.

Une fois qu'on a obtenu les composantes d'ondelette de chaque terme de la requête, nous devons les combiner pour pouvoir calculer la pertinence de chaque document par rapport à la requête. Sachant que l'amplitude du signal est liée à l'occurrence du terme de la requête et sa phase est

liée à la position du terme dans le document. Nous allons traiter ces deux notions séparément. Le spectre des valeurs combiné est défini comme suit :

$$S_d = [S_{d,0}, S_{d,1}, \dots, S_{d,B-1}]$$

Chaque composante d'ondelette est égale à :

$$S_{d,b} = \Phi_{d,b} \sum_{t \in T} H_{d,t,b} \quad (42)$$

Avec $H_{d,t,b}$ représente l'amplitude de la $b^{\text{ième}}$ composante du terme t dans le document d ,

$\Phi_{d,b}$ est la phase.

3.2.3.5 Algorithme de recherche d'information spectrale

1)- Pour chaque terme t de la requête :

A- Construire le signal correspondant dans le document D

B- Pondération des composantes du signal $I_f * I_{df}$

C- Appliquer la transformée en ondelette sur le signal

2)- Pour chaque document D

Pour chaque composante du signal de terme :

- Calculer les composantes amplitudes $H_{d,t,b}$.
- Calculer les composantes phases $\Phi_{d,b}$
- Combiner les phases obtenues
- Combiner les scores obtenus afin de calculer le score de pertinence du document

3.2.4 Expérimentation

Nous allons expliquer le processus de l'approche spectrale qu'on a présenté dans la partie précédente à travers l'exemple suivant :

Le corpus des données :

Nous allons appliquer notre approche spectrale sur un ensemble de deux documents représentant deux discours de deux candidats à la présidentielle en France en 2007. L'ensemble des données est présenté dans le tableau 13. La taille du document 1 est de 88 mots et la taille du second document est de 80 mots.

Le tableau 14 contient 3 termes sélectionnés dans les deux documents de notre corpus, pour chaque terme et dans chaque document correspondra un signal contenant 8 composantes (taille du signal puissance de 2). Chaque composante représente l'occurrence d'apparition de ce terme dans un segment donné du texte.

Par exemple, la taille du second document est de 80 mots, pour pouvoir construire des signaux de taille de 2^3 , le texte sera segmenter en 8 segments de taille de 10 mots chacun. Le premier élément du signal du terme *politique* dans le second document indique que le terme *politique* apparaît 1 fois dans le premier huitième du document. Le cinquième élément du signal a pour valeur 2, ce qui implique que le terme *politique* apparaît 2 fois dans le cinquième huitième du document.

Document 1	<i>« La première des politiques d'immigration, c'est le rétablissement de nos frontières et l'objectif d'immigration zéro. Pour préserver notre identité et notre sécurité, il faut en effet limiter l'accès à notre territoire, ce qui suppose d'en avoir le contrôle et donc d'entrer en négociation avec nos partenaires européens pour récupérer, au plus vite, les moyens de maîtriser notre destin. Pour limiter l'accès au sol français : nous pourrons ainsi lutter contre l'immigration clandestine, faire la chasse aux faux touristes, contrôler l'arrivée des demandeurs d'asile. Pour lutter contre l'immigration chaotique. » JM. Le Pen, 2007</i>
Document 2	<i>« La politique de civilisation, la France veut la faire en engageant une nouvelle politique de l'immigration qui permette la maîtrise en commun des flux migratoires par le pays d'accueil et les pays de départ. Elle veut la faire avec une politique d'immigration décidée ensemble et gérée ensemble. Avec une politique fondée sur les quotas fixés en fonction des capacités d'accueil et d'intégration qui évite tous les drames humains, tous les rejets, toutes les exclusions liés à une immigration non maîtrisée. » N. Sarkozy, 2007</i>

Tableau 13. Corpus des données pour la recherche d'information spectrale

Mots	Document 1	Document 2
immigration	[1 1 0 0 0 0 1 1]	[0 1 0 0 1 0 0 1]
politique	[1 0 0 0 0 0 0 0]	[1 1 0 0 2 0 0 0]
maîtrise	[0 0 0 0 0 1 0 0]	[0 1 0 0 0 0 0 1]

Tableau 14. Un ensemble de termes d'indexation avec leurs signaux dans les deux documents.

Nous ignorons l'étape de la pondération des valeurs des signaux dans cette expérimentation, pour nous concentrer sur le procédé de l'application de la Transformée en ondelette et le calcul de la pertinence des documents.

Considérons qu'on interroge notre base de données avec la requête de recherche suivante : « *immigration maîtrisée* ». Le système de recherche extrait les données concernant les deux termes *immigration* et *maîtrisée* pour le calcul de la pertinence de chaque document, mais il ignore ceux concernant le terme *politique*.

Nous allons examiner le calcul de la pertinence du premier document par rapport à la requête en détails.

La première étape du calcul de la pertinence consiste à appliquer la Transformée en ondelette sur les deux signaux, nous avons fait le choix d'utiliser l'ondelette de Haar qui possède un support compact de petite taille, cela implique que sa transformée du signal nécessitera le minimum d'espace de stockage. Les coefficients d'Ondelette obtenus sont les suivants :

Mots	Document 1
immigration	[1/2 0 1/2 -1/2 0 0 0 0]
maîtrise	[1/8 -1/8 0 1/4 0 0 -1/2 0]

Nous allons additionner l'amplitude des deux signaux

Mots	Document1 amplitude
immigration	[1/2 0 1/2 +1/2 0 0 0 0]
maîtrise	[1/8 +1/8 0 1/4 0 0 +1/2 0]
Total	[5/8 1/8 1/2 3/4 0 0 1/2 0]

Après le calcul de l'amplitude, nous allons combiner les phases sur les deux signaux :

Mots	Document1 phase
immigration	[1 0 1 -1 0 0 0 0]
maîtrise	[1 -1 0 1 0 0 -1 0]
Combinaison des phases	[1 1/2 1/2 0 0 0 1/2 0]

Combinons l'amplitude et la phase obtenues précédemment pour calculer le score de pertinence du document :

Document 1 amplitude	[5/8 1/8 1/2 3/4 0 0 1/2 0]
Document 1 phase	[1 1/2 1/2 0 0 0 1/2 0]
Document 1 vecteur de pertinence	[5/8 1/16 1/4 0 0 0 1/4 0]
Document 1 Pertinence	19/16= 1,1875

Pour le calcul de la pertinence du document 1, on additionne les valeurs du vecteur correspondant, on obtient le score 1,1875. Et en suivant le même processus de calcul de pertinence pour le second document nous obtenons le score de 2,5625.

Si on examine les positions des mots de la requête dans leurs représentations signal, on peut observer que les scores obtenus pour les documents reflètent la proximité et les fréquences d'apparition des termes de la requête dans les documents.

En effet, concernant le document 2, nous pouvons observer que les termes de la requête apparaissent dans la même composante du signal donc ils apparaissent étroitement dans le même segment de ce document. Il a le score le plus élevé.

Le document 1 est rangé en seconde position en réponse à la requête et nous observons que les termes de la requête apparaissent dans des composantes éloignées dans la représentation spectrale correspondante.

3.2.5 Comparaison des résultats

Afin de juger l'efficacité de notre approche de recherche d'information spectrale, nous avons comparé ses résultats à ceux obtenus avec les méthodes de recherche d'information traditionnelles et spécialement la représentation vectorielle classique.

Le modèle vectoriel représente un document ou une requête par un vecteur dans l'espace des termes. C'est-à-dire l'espace de représentation construit à partir des termes de l'index. Cet espace comporte 46 termes.

Chaque document sera représenté par un vecteur de dimension 46, les éléments valent pour la plupart zéro : seuls ceux qui correspondent aux termes présents dans le document se sont pas nuls. De la même manière on représente la requête « immigration maîtrisée » par un vecteur dans le même espace vectoriel que les vecteurs représentant les documents.

Pour évaluer la correspondance entre chaque document et la requête nous allons utiliser la mesure de Cosinus définie ainsi :

$$Cos = \frac{\sum_{k=1}^t (w_{ki} * w_{qk})}{\sqrt{\left[\sum_{k=1}^t (w_i^k * w_i^k) * \sum_{k=1}^t (w_q^k * w_q^k) \right]}}$$

La mesure de similarité donne un angle de 88° séparant le vecteur représentant le document 1 et le vecteur de la requête. Et un angle de 62° degré entre le vecteur représentant le document 2 et celui de la requête.

Plus l'angle qui sépare les deux vecteurs est petit, plus le document est proche de la requête. D'après ces résultats, le document 2 est considéré comme plus pertinent par rapport à la requête que le document 1. Ces résultats confirment ceux obtenus avec le modèle spectral qu'on propose.

3.2.6 Discussion

La complexité des calculs dans le système de recherche d'information qu'on propose est évaluée en nombre d'opérations nécessaires pour exécuter une tâche de comparaison entre un document et une requête.

- Sois N le nombre de documents dans le corpus,
- sois B la taille du signal utilisée pour construire les spectres des mots de la requête dans les documents,
- sois T le nombre des mots de la requête.

Organisation du calcul :

Le calcul de la pertinence d'un document par rapport à une requête se déroule ainsi :

- La pondération du spectre de chaque mot de la requête nécessite $O(NTB)$ opérations,

- l'application de la transformée en ondelette nécessite $O(NTB)$ opérations,
- le calcul de l'amplitude pour chaque spectre nécessite $O(NTB)$ opérations,
- le calcul de la phase de chaque spectre nécessite $O(NTB)$ opérations,
- le calcul des composantes du score de chaque document en multipliant les composantes amplitudes et phases nécessite $O(NB)$ opérations,
- le calcul du score du document en additionnant les composantes du score nécessite $O(NB)$ opérations.

On note que le modèle vectoriel classique est un cas particulier de la modélisation spectrale qu'on propose, pour $B=1$. La modélisation spectrale permet de choisir et de varier la taille B en faveur de la vitesse des calculs ou l'exactitude des résultats.

Des travaux en cours s'attachent à valider le modèle en utilisant d'autres types de transformées en ondelettes sur des corpus de type et de taille différente, afin de pouvoir juger la pertinence de notre modèle spectrale en terme de précision et de rappel.

CONCLUSION

Nos travaux s'intéressaient à la modélisation de l'information textuelle pour l'analyse et la recherche d'information. Nos recherches ont porté essentiellement sur les points suivants :

- Le mode de représentation des documents dans un corpus
- le mode de représentation des requêtes exprimées par un utilisateur, pour la recherche d'information,
- la comparaison, à l'aide de la modélisation que nous avons défini, entre un document et une requête ou entre plusieurs documents.

Les principaux modèles de représentation de l'information : modèle booléen, booléen pondéré, modèle vectoriel, modèle probabiliste, etc...ont été conçus, pour la plupart, il y'a une trentaine d'années. La grande majorité des recherches actuelles se fonde sur ces modèles pour améliorer les résultats des SRI.

Dans cette thèse nous avons proposé un nouveau système de recherche d'information, désigné « *système de recherche d'information spectral* », se fondant sur un nouveau modèle de représentation de l'information, qui suppose que les données textuelles soient représentées par des signaux. Cette modélisation nous a permis de faire appel à des outils mathématiques connus en traitement du signal. Notamment pour effectuer les comparaisons documentaires. Parmi ces outils, nous avons présenté les Transformées en ondelettes.

Pour formaliser notre système, nous avons développée dans un premier temps, la modélisation des documents textuels en signaux thématiques. Dans cette représentation les termes du texte sont remplacés par leurs valeurs de pertinence par rapport à une thématique donnée. Le flot de cette thématique peut être tracé dans le texte sous format de signal qui modélisera la présence et la variation de cette thématique dans le texte.

Cette modélisation du contenu des textes peut être utilisée dans plusieurs applications : la comparaison documentaire, la visualisation, la segmentation, la détection des thèmes, ...

Nos expérimentations ont démontré que la qualité du signal dépend des mesures utilisées dans le choix des thèmes représentatifs dans un corpus ainsi que les mesures de calcul de pertinence du vocabulaire lié à chaque thème.

La Transformée en ondelettes, apparue au début des années 1990, est un outil utilisé dans de nombreux domaines notamment le traitement des signaux, traitement des images, dans la visualisation, etc. La Transformée en ondelettes représente un signal quelconque par une superposition des signaux élémentaires (les Ondelettes) oscillants mais localisées dans le temps, à la différence de la Transformée de Fourier.

L'application de la Transformée en ondelette de Haar sur les signaux représentant un texte, nous a permis de tirer avantage de sa propriété de multi résolution. Cette propriété permet d'avoir des approximations du signal d'origine (des versions lisses) qu'on a pu utiliser dans la comparaison des documents à différents niveaux de détails.

Cette proposition de modélisation peut également être utilisée dans la visualisation de grands corpus de données textuelles en temps réel. Une application similaire présentant un prototype de visualisation qui permet à l'utilisateur d'observer les changements et naviguer dans un texte en utilisant un seul signal thématique et à différents niveaux de résolution est présentée par Miller [2]. La construction des 'signaux thématiques' permettent d'enrichir ce prototype.

Nous avons ensuite défini les modalités de calcul de la pertinence entre les requêtes des utilisateurs et un corpus de données. Notre hypothèse est qu'un document est considéré comme pertinent par rapport à une requête s'il contient une occurrence élevée des termes de la requête ainsi qu'une certaine proximité de ces termes. Dans le système de recherche d'information spectral, cela revient à représenter chaque terme de requête par un spectre. La fréquence d'apparition élevée correspond à une amplitude élevée du spectre, la proximité des termes correspond à une phase similaire de tous les spectres des termes de la requête.

Nous avons utilisé l'Ondelette de Haar pour valider ce modèle et nous avons pu comparer et confirmer l'exactitude de nos résultats avec la représentation vectorielle classique. Néanmoins, le choix de l'ondelette de Haar est ici totalement arbitraire et liée à son utilisation importante en théorie du signal. Des travaux en cours s'attachent à valider le modèle en utilisant d'autres types de Transformées en ondelettes sur des corpus de type et de taille différente, afin de pouvoir juger la pertinence de notre modèle spectrale en terme de précision et de rappel.

Cette thèse propose un changement dans la modélisation de l'information textuelle dans le domaine de la recherche d'information. La modélisation des textes en signaux doit permettre de sortir de la typologie traditionnelle des systèmes de recherche d'information actuels. Elle permettra de mieux considérer les occurrences et les positions des termes, pour une analyse plus exacte du contenu des documents dans des tâches de recherche d'information.

La modélisation spectrale dans la recherche d'information textuelle est un concept récent, pour aller plus loin dans le développement de ce modèle, plusieurs pistes de recherche restent à explorer :

- L'analyse et l'utilisation de différentes mesures et méthodes de calcul de pertinence et de sélection des thèmes représentatifs dans un corpus,
- l'observation de la variation de la similarité entre documents en utilisant d'autres types de Transformées en ondelettes et sur différents niveaux de résolution,
- le choix de la taille du spectre de la requête par rapport à la taille du texte, dans une tâche de comparaison entre document et requête.

ANNEXE 1

Outils d'extraction terminologique

Introduction :

Le but de l'extraction terminologique est d'extraire des documents, les descripteurs linguistiques représentatifs de leur contenu. Dans la dernière décennie de nombreux projets de recherche ont été menés dans le domaine de l'extraction automatique de termes à partir de corpus et ils ont abouti à la réalisation de plusieurs logiciels robustes.

Dans cette partie, nous donnons la description d'un certain nombre de ces systèmes, en insistant sur les caractéristiques des méthodes plutôt que sur le type de terminologie acquise. Il existe trois types de méthodes :

1. méthodes linguistiques : LEXTER, ANA, FASTER, INTEX, TERMINO
2. méthodes statistiques : XTRACT
3. méthodes mixtes : ACABIT, EXIT

1. méthodes d'extractions linguistiques :

1.1. Description de l'outil « LEXTER » :

LEXTER a été développé par Didier Bourigault [84] au sein de la Direction des Études et Recherches d'EDF dans le cadre d'un projet de gestion de la documentation technique de l'entreprise. Le système LEXTER fonctionne en plusieurs étapes :

L'outil utilise un étiqueteur externe pour identifier les frontières potentielles qui ne peuvent pas être des constituants d'un terme (verbes, conjonctions, pronoms). Ensuite, il repère les groupes nominaux bornés par ces frontières : ces groupes nominaux sont dits *maximaux*. Ces groupes nominaux maximaux sont ensuite, récursivement, décomposés en tête et en expansion.

Enfin, un réseau terminologique est créé en reliant chaque terme candidat à sa tête, à son expansion et aux termes candidats dont il est lui même tête ou expansion.

1.1.1. Caractéristiques de LEXTER :

- LEXTER extrait des termes candidats à partir d'un corpus préalablement étiqueté et désambiguïsé.
- Le système résout des associations ambiguës d'adjectifs et de prépositions au sein des groupes nominaux complexes.
- LEXTER est sûrement un des meilleurs extracteurs de syntagmes nominaux. L'outil ACABIT (présenté ci-dessous) ne repère que les syntagmes associés à certains schémas syntaxiques, tandis que les méthodes statistiques ne repèrent que les termes très fréquents ou encore MANTEX qui ne repère que les termes répétés.
- Le rappel (le nombre de documents pertinents retrouvés par rapport au nombre total de documents pertinents dans la base) très élevé de LEXTER se paye par une précision (le nombre de documents pertinents retrouvés rapporté au nombre de documents total retrouvés.) assez faible ; si on utilise LEXTER seul, ce coût supplémentaire n'est justifié que si le temps de création de la base de termes finale reste raisonnable. Sinon, les systèmes plus silencieux (c'est-à-dire ne détectant pas tous les termes) mais donnant moins de travail à l'expert seront plus intéressants.

1.1.2. Spécifications Supplémentaires :

- Systèmes d'exploitation sous lesquels tourne l'outil : Unix.
- Type d'interface d'utilisation : Interface ligne de commande.
- LEXTER est la propriété d'EDF (Division de la Recherche et du Développement). Pour une utilisation commerciale, ou pour la recherche et l'enseignement, une convention préalable près d'EDF est nécessaire.

1.2. Description de l'outil ANA « Apprentissage Naturel Automatique »

ANA [85] est le premier système qui s'inspire de l'apprentissage humain de la langue maternelle, il se divise en deux modules : module familiarisation et module découverte.

1. Le module Familiarisation : il extrait, automatiquement quelques éléments de connaissance sur la langue utilisée et le domaine abordé, sous la forme de quatre listes :

1^{ère} liste : elle contient les mots fonctionnels les plus fréquents de la langue utilisée : articles, pronoms, adverbes. Le système les sélectionne automatiquement grâce à une procédure entièrement statistique. Exemple {"a", "alors", "après", "au", "auraient", "aussi", "autre", "avait", "avant"...

2^{ème} liste : sont des chaînes de caractères comprenant des caractères blancs mais pourtant considérés comme des mots. Ces mots sont le résultat de la variation morphologique de certains mots fonctionnels. Ainsi, "de" devient "des" au pluriel, "de la" au féminin et "du" au singulier. Exemple {"de l", "de la", "est en", "et la" "est le", "la on", "on a" }

3^{ème} liste : elle contient les mots de schémas : mots fonctionnels structurants les groupes de mots : « de, des, en ... ».

4^{ème} liste : contient les 'bootstrap' : quelques termes du domaine dont il est question dans le corpus de textes.

2. Le module Découverte : il utilise les quatre listes créées au niveau du module familiarisation ainsi que le document à étudier pour sélectionner la terminologie du domaine abordé.

1.2.1. Caractéristiques de 'ANA':

- ANA fonctionne sur des textes même de mauvaise qualité, on ne peut donc s'attendre à n'y trouver que des structures syntaxiques correctes.
- ANA est multilingue, il a été testé sur le français, l'anglais et l'italien.
- Le système est indépendant vis-à-vis de la langue utilisée dans les textes à traiter.

1.2.2. Spécifications Supplémentaires :

- Système d'exploitation sous lequel tourne l'outil : Unix.
- Type d'interface d'utilisation : Interface ligne de commande.
- Le système ANA n'a jamais été industrialisé et il n'est pas à vendre.

1.3. Description de l'outil FASTR « Filtrage et Acquisition Syntaxique et Termes »

FASTR [86], [87] est un outil de reconnaissance de termes, il sert principalement à indexer des documents à partir d'un thésaurus ou d'une liste de termes contrôlés.

FASTR prend en entrée un ensemble de termes simples ; leur analyse morphologique est recyclée sous forme de règles comprenant un squelette hors contexte et des éléments lexicaux. Un ensemble de métarègles permet de décrire des variations des termes trouvés dans les textes. L'auteur de l'outil, Christian Jacquemin répertorie trois familles de variations : les variantes syntaxiques (expansion nominale remplacée par une conjonction), morphosyntaxiques (la tête ou l'expansion changent de partie du discours) et sémantico-syntaxiques (la tête ou l'expansion sont remplacés par un élément sémantiquement proche). Ces règles et métarègles sont implémentées dans un analyseur robuste et rapide pouvant traiter efficacement de gros corpus.

FASTR permet donc de repérer des variantes de termes. Et peut aussi servir à acquérir de nouveaux termes simples par un processus inverse.

1.3.1. Spécifications Supplémentaires :

- Systèmes d'exploitation sous lesquels tourne l'outil : Unix, Mac, Windows.
- Type d'interface d'utilisation : Interface ligne de commande.
- L'outil a besoin des ressources suivantes : un étiqueteur, une base de morphologie, un thésaurus avec des liens de synonymie.
- L'outil est disponible pour des fins de recherche, il est téléchargeable depuis le lien électronique : <http://www.limsi.fr/Individu/jacquemi/FASTR/>
- L'outil téléchargé fonctionne pour l'anglais et le français.
- Des versions pour le Japonais, l'Espagnole et l'Allemand sont en cours de développement.

1.4. Description de l'outil « INTEX » :

INTEX [88] est un environnement linguistique permettant l'analyse des corpus en se basant sur l'utilisation de ressources lexicales à très large couverture. Il comprend plusieurs dictionnaires électroniques et des grammaires représentées par des graphes à états finis. Les utilisateurs peuvent ajouter leurs propres ressources au système.

Ces outils sont utilisés sur des textes pour localiser des structures lexicales et syntaxiques, lever les ambiguïtés et étiqueter des mots simples ou composés.

INTEX est utilisé par les linguistes pour analyser des corpus et développer des ressources linguistiques, mais peut être également vu comme un système de Recherche d'Information.

1.4.1. Spécifications supplémentaires :

- Système d'exploitation sous lequel tourne l'outil : Windows.
- Type d'interface d'utilisation : Interface graphique, Interface ligne de commande, Interface de programmation (API).
- Disponibilité de l'outil pour la recherche : INTEX est téléchargeable depuis le lien électronique : www.nyu.edu/pages/linguistics/intex , ainsi que l'ensemble des dictionnaires et grammaires. Un numéro de licence est fourni par l'auteur pour chaque utilisation.

La nouvelle mouture du logiciel INTEX, appelée NooJ (89) a été réécrite en particulier pour répondre aux besoins des utilisations pédagogiques.

NooJ⁷ intègre des outils de traitement automatique du langage qui offrent à l'enseignant des possibilités de traiter un corpus et des procédures de recherche, de test et d'entraînement pour l'étudiant.

Les quatre principales caractéristiques de NooJ :

- **Traitement de corpus** : NooJ permet de traiter des ensembles importants de documents, qui peuvent être dans n'importe quel format : pages Internet, documents XML, Microsoft WORD, etc.

- **Les travaux pratiques** : NooJ propose des mini-applications pédagogiques qui peuvent être utilisées en séance de travaux pratiques, autour des problèmes comme : les différents codages informatiques des textes, l'ordre alphabétique dans chaque langue, les expressions rationnelles, la morphologie flexionnelle, etc.

⁷ Site du logiciel NooJ, disponible et gratuitement téléchargeable. <http://www.nooj4nlp.net>

- **Les projets :** À tout moment, on peut sauvegarder la configuration de l'environnement NooJ sous la forme d'un fichier "projet". Ca permet de faire appel à une configuration avec un corpus et des préférences de traitement spécifiques.

- **Construction, édition et gestion sophistiquées de concordances :** NooJ permet de combiner plusieurs requêtes, filtrer manuellement les résultats incorrects, puis sauvegarder la concordance résultante.

2. Méthodes d'extraction statistiques:

2.1. Description de l'outil « Terminology Extractor» :

Terminology Extractor 'TE' est un outil qui extrait des listes de mots avec des fréquences, à partir des documents Microsoft Word (97/2000/XP), des documents HTML et des feuilles de styles.

L'outil 'TE' emploie un certain nombre de dispositifs et d'algorithmes pour fournir le meilleur rendement, Le logiciel emploie la forme canonique de chaque mot et Il utilise également comme filtres des listes de mots de commandes (pronoms, articles, prépositions...). En outre les noms propres sont maintenus dans leurs formes originales.

2.1.1. Spécifications supplémentaires :

- Systèmes d'exploitation sous lesquels tourne l'outil : Windows.
- Type d'interface d'utilisation : Interface graphique.
- Disponibilité de l'outil pour la recherche : l'outil est vendu au Prix de : \$CAN 445 (à peu près \$US 300)
- Une version de démonstration est disponible sur le lien suivant :

<http://www.chamblon.com/download30.htm>.

3. Méthodes d'extraction mixtes :

3.1. Description de l'outil 'ACABIT' "Automatic corpus-Based Acquisition of Binary terms" :

ACABIT [90] a pour objectif de préparer la tâche du terminologue en lui proposant une liste ordonnée de termes candidats pour un corpus préalablement étiqueté et lemmatisé et il effectue une analyse syntaxique suivie d'un traitement statistique.

L'auteur collecte des schémas syntaxiques de termes simples et des mécanismes de variation permettant d'obtenir des termes plus complexes, le système utilise un ensemble de transducteurs pour extraire les termes composés formés selon les schémas précédents et ne conserve que les formes lemmatisées. La simple fréquence étant un critère insuffisant pour détecter les termes du domaine. ACABIT repose sur l'utilisation de diverses mesures statistiques qui retiennent le mieux les termes candidats sans être sensible aux fréquences.

3.1.1. Spécifications Supplémentaires :

- Systèmes d'exploitation sous lesquels tourne l'outil : Unix.
- Type d'interface d'utilisation : Interface ligne de commande.
- ACABIT nécessite pour le français un corpus étiqueté à l'aide de l'étiqueteur de Brill adapté à l'INALF, puis le lemmatiseur FLEMM développé par F. Namer.
- ACABIT peut être utilisé librement à des fins de recherche, il est téléchargeable depuis la page perso de Béatrice Daille :

<http://www.sciences.univnantes.fr/info/perso/permanents/daille/>

ANNEXE 2

Outils d'analyse par Ondelettes

1. WAVELAB :

WAVELAB est une bibliothèque de fonctions MATLAB portant sur les ondelettes et les transformées temps-fréquence associées. Elle est maintenue et améliorée à l'université de Stanford par David Donoho [91]. Son utilisation nécessite l'achat de MATLAB, un produit de la société The MathWorks, qui offre un environnement interactif de calcul numérique et de visualisation.

WAVELAB comprend plus de 800 fichiers, dont des programmes, des données, de la documentation et des scripts. Des versions pour Unix, Linux, Macintosh et PC sont disponibles.

Voici quelques exemples de fonctions utilisées dans WAVELAB :

- ReadSignal : lit un signal dans un jeu de données constituées de signaux en dimension 1.
- ReadImage : lit une image dans une collection d'images.
- DWT : Transformée en ondelettes réelles.
- FWT_YM : Transformée en ondelettes d'Yves Meyer.

2. LAST WAVE :

LAST WAVE est un environnement de traitement de signal et de l'image, écrit en Langage C. Ce logiciel est gratuit et autonome est ne nécessite pas de logiciel commercial additionnel et peut être obtenu sur Internet à l'adresse suivante :

<http://www.cmap.polytechnique.fr/~bacry/LastWave/>

LAST WAVE a été créé et maintenu par Emmanuel Bacry, à l'école Polytechnique en France. Ce logiciel comprend un langage en ligne de commande et un langage graphique pour afficher des objets simples (boutons, chaînes,...) et plus complexe (signaux, images, Transformée en ondelettes).

3. Outils d'analyse freeware :

Nous allons décrire quelques logiciels en freeware qui sont disponibles sur Internet :

- 3.1. EMBEDDED IMAGE COMPRESSION : c'est un logiciel en C++ pour la compression d'image en ondelettes :

<http://www.cipr.rpi.edu/research/SPIHT>

- 3.2. MEGAWAVE : une collection de commandes en C sous Unix pour le traitement par ondelettes, avec des applications de traitement du son et des images.

<http://megawave.cmla.ens-cachan.fr/index.fr.php>

- 3.3. RICE WAVELET TOOLBOX : une boîte à outils MATLAB d'ondelettes avec des transformées orthogonales et bi-orthogonales et des applications au débruitage.

<http://www-dsp.rice.edu/software/RWT>

- 3.4. TIME FREQUENCY : une boîte à outils MATLAB pour l'analyse de signaux non- stationnaires

<http://tfd.sourceforge.net/>

BIBLIOGRAPHIE

- [1] **SMAIL, N.** "*Modélisations pour l'analyse et la recherche documentaire en texte intégral: état de l'art et perspectives*". Université de Paris Est-Marne La Vallée. 2004. mémoire de DEA.
- [1'] **EPPSTEIN, R. SMAIL, N**" *La modélisation du texte en signal : vers un nouveau modèle de représentation de l'information*", VSST'2007, Marrakech, Octobre 2007.
- [2] **MILLER, N.M, PAK, C.W et BREWSTER, M** "*Topic Islands A Wavelet-Based Text Visualization System*". 1998. IEEE Visualization. pp. 189-196.
- [3] **Al-Halimi, Reem khalil** "*Mining Topic Signals from Text*". University of Waterloo. Ontario, CANADA : s.n., 2003. p. 176, Thèse de doctorat .
- [4] **SALTON, G.** "*The SMART Retrieval System- Experiments in Automatic Document Processing*". [éd.] N.J. Englewood Cliffs. s.l. : Prentice-Hall, 1971.
- [5] **DEERWESTER, S.** "*Indexing by latent semantic analysis*". Journal of the American Society for Information Science, pp. 391-407, 1990.
- [6] **HEARST, M.A.** "*Multi-Paragraph segmentation of Expository Texts*". [éd.] Association for Computational Linguistics. Las Cruces, New Mexico : Proceedings of the 32nd annual meeting on Association for Computational Linguistics. pp. 9-16. 1994.
- [7] **MOOERS, C.N.** "*Application of Random Codes to the Gathering of Statistical Information*", MIT, Thèse de Master, 1948.
- [8] **CLEVERDON, C.W,** "*The Cranfield tests on index language devices*", Aslib Processing 19, pp. 173-193, 1967.
- [9] **CHIRAMELLA, Y.** "*A prototype of an intelligent system for information retrieval:IOTA*", Information Processing and Management, pp. 285-303, 1987.
- [10] **AFNOR, Associations française de normalisation.** "*Vocabulaire de la documentation*". Paris-La Défense : Collection Les Dossiers de la normalisation, ISNNS 0297-4827, p 159. ISBN 2-12-484221-8, 1987.
- [11] **ABITEBOUL, S, et al** "*The Lorel Query Language for Semi Structured Data*", Journal of Digital Libraries , Vol. 1, pp. 68-88, 1997.
- [12] **ABITEBOUL, S.** "*Querying Semi-structured Data*". Delphi, Greece : International Conference On Database. pp. 1-18, 1997.
- [13] **MIZZARO, S.** "*How many relevance's in information retrieval?*", Italie : Departement of Mathematics and Computer Science, University of Udine, 1998.

- [14] **FRIEDER, O et GROSSMAN, D.** "*Information Retrieval, Algorithms and Heuristics*". 2e édition. s.l. : The Springer International Series in Engineering and Computer Science, 1998.
- [15] **SALTON, G.** "A vector space model for automatic indexing". *ACM*. 1975, Vol. 18, 11, pp. 613-620.
- [16] **BESANCON, R.** "*Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de textes: application au calcul de similarités sémantiques dans le cadre du modèle DSIR*". Ecole Polytechnique Fédérale de Lausanne. 2001. Thèse de doctorat.
- [17] **ROBERTSON, E.** "*The probability ranking principle in IR*". San Francisco, CA, USA : Morgan Kaufmann Publishers Inc, 1997, *Journal of Documentation*, pp. 281-286. 1-55860-454-5.
- [18] **LANCASTER, F.** "*Information retrieval systems: characteristics, testing, and evaluation*", 2 édition. NEW YORK : John Wiley & Sons, 1979. 0-471-04673-6.
- [19] **CHAUDIRON, S.** "*L'évaluation des systèmes de traitements de l'information textuelle: vers un changement de paradigme*". Université Paris X. 2001. p. 300, mémoire pour HDR.
- [20] **HARTER, S et HERT, C.** "*Evaluation of information retrieval systems: approaches, issues and methods*". [éd.] Washington, DC, ETATS-UNIS American Society for Information Science. 1997, *Annual review of information science and technology* , Vol. 32, pp. 3-94. 0066-4200.
- [21] **SPARCK JONES, K.** "*Information retrieval experiment*". [éd.] Butterworth-Heinemann. Londres : s.n., 1981. p. 360. 978-0408106481.
- [22] **CORET, A.** "*Accès à l'information textuelle en français: le cycle exploratoire Amaryllis*". Avignon : s.n., 1992. Premières Journées FRANCIL. pp. 5-8.
- [23] **CHAUDIRON, S et SCHMITT, L.** "*AMARYLLIS : an evaluation-based program for Text Retrieval*". Athens, ELRA : s.n., 2000. Workshop Proceedings of LREC 2000. pp. 65-68.
- [24] **CLEVERDON, C.** "*Report of the first step of an investigation into the comparative efficiency of indexing systems*". college of Aeronautique. Cranfield : s.n., 1960.
- [25] **BENZECRI, J.P.** "*L'analyse des données Tome 2: l'analyse des correspondances*". Paris : Dunod, 1973. 2-04-007225-X.
- [26] **SALTON, G.** "*Automatic Text Processing: The Transformation Analysis and Retrieval of Information by Computer*". s.l. : Addison-Wesley series in computer science, 1988. p. 543. 978-0201122275.

- [27] **HEARST, M.A.** "*TileBars: visualization of term distribution information in full text information access*". Denver, Colorado, United States : ACM PRESS/Addison-Wesley Publishing Co, 1995. Conference on Human Factors in Computing Systems, Proceedings of the SIGCHI, pp. 59-66. 0-201-84705-1.
- [28] **ROBERTSON, G et MACKINLAY, J.** "*Cone Trees: animated 3D visualizations of hierarchical information*". New Orleans, Louisiana, United States : ACM Press, 1991. Conference on Human Factors in Computing Systems, Proceedings of the SIGCHI, pp. 189-194. 89791-383-3.
- [29] **LAMPING, J, RAMANA, R et PIROLI, P.** "*A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies*". Denver, Colorado, United States : ACM Press/ Addison-Wesley Publishing , 1995. Conference on Human Factors in Computing Systems, Reaching through technology. pp. 401-408. 0-201-84705-1.
- [30] **MOKRANE, A et PONCELET, P.** "*Cartographie automatique du contenu d'un corpus de documents textuels*". Belgique : s.n., 2004. 7 Journées Internationales de L'Analyse Statistiques de Données Textuelles. pp. 816-823.
- [31] **JOUVE, O et CORONINI, R.** "*Analyse de données et analyse des mots associés, comparaison d'algorithmes sur un corpus concernant la prise en compte du risque dans le développement des OGM dans le domaine des végétaux*". Toulouse : s.n., 1998. veille stratégique scientifique & technologique VSST. pp. 241-256.
- [32] **JOUVE, O.** "*Les outils d'analyse et de filtrage d'information :L'exemple du projet Sampler*". Paris-La Défense : s.n., 1999. CISI.
- [33] **DOUSSET, B et DKAKI, T.** "*TETRALOGIE : A new method for Competitive Intelligence*". Marrakech, Maroc : s.n., 1995. International Conference on Industrial Engineering and Management.
- [34] **DOUSSET, B et DKAKI.** "*TETRALOGIE: a platform for scientific and technological survey*". [éd.] LORIA. Nancy : s.n., 2006. International Workshop on Webometrics, Infometrics and Scientometrics & Seventh COLLNET Meeting.
- [35] **BAEZA-YATES, R et RIBEIRO-NETO, B.** "*Modern Information Retrieval*". s.l. : ACM/Press, Addison Wesley Longman, 1999.
- [36] **SALTON, G et BUCKLEY, C.** "*Approaches to passage retrieval in full text information systems*". ACM . New York : Annual ACM Conference on Research and Development in Information Retrieval, Proceedings of the 16th annual international, 1993. pp. 49-58. 0-89791-605-0.
- [37] **DUCROT, O et SCHAEFFER, J.M.** "*Nouveau dictionnaire encyclopédique des sciences du langage*". s.l. : éditions du SEUIL, 1995.
- [38] **RASTIER, F.** "*L'Analyse thématique des données textuelles : l'exemple des sentiments*". Paris : Collection Etudes de sémantique lexicale, 1995. 2-86460-244-X.

- [39] **MILLER, N et BREWSTER, M.** "*Information retrieval system utilizing wavelet transform*". *United States Patent 6070133* [éd.] Battelle Memorial Institute. 30 MAY 2000.
- [40] **BOOKSTEIN, A, KLEIN, S et RAITA, T.** "*Clumping properties of content-bearing words*". 2, s.l. : John Wiley & Sons, 1998, *Journal Of the American Society of Information*, Vol. 49, pp. 102-114. 0002-8231.
- [41] **FOUREIR, J.B.** "*Théorie analytique de la chaleur*". Paris : s.n., 1922.
- [42] **MEYER, Y.** "*Ondelettes et opérateurs*", *Tome 1*. [éd.] Hermann. Paris : Actualités mathématiques, 1990.
- [43] **MALLAT, S.** "A theory for multiresolution signal decomposition: the wavelet representation". *Pattern Analysis and Machine Intelligence, IEEE trans.* 1989, Vol. 11, 7, pp. 674-693.
- [44] **Burke Hubbard, B.** "*Ondes et ondelettes. La saga d'un outil mathématique*". [éd.] Pour la Science. Paris : s.n., 1995. 2-9029-1890-9.
- [45] **MEYER, Y et NOWAL, M.** "*La surprenante ascension des Ondelettes*". [éd.] Société d'éditions scientifiques. Paris, France : s.n., Février 2005, Recherche, pp. 56-59. 0029-5671.
- [46] **MALLAT, S.** "*Une exploration des signaux en Ondelettes*". Palaiseau : Editions de l'école Polytechnique, 2000.
- [47] **TANG, Y, WICKERHAUSER, V et YUEN, P.C.** "*Wavelet analysis and Its Applications*". Hong Kong, china : Springer, 2001. Seconde International Conference, WAA 2001. p. 462.
- [48] **JEONG, Myong-Kee.** "Wavelet-Based Methodology in Data Mining for Complicated Functional Data ". rapport deThèse: Industrial and Systems Engineering, Mai 2004.
- [49] **SHAHABI, C, CHUNG, S et SAFAR, M.** "*2D TSA-tree: A Wavelet-Based Approach to Improve the Efficiency of Multi-Level Spatial Data Mining*". Thirteenth International Conference on Scientific and Statistical Database Management. pp. 59-68. 2001.
- [50] **SHAHABI, C, TIAN, X et ZHAO, W.** "*TSA-tree: a wavelet-based approach to improve the efficiency of multi-level surprise and trend queries on time-series data*". 2000. SSDBM, 12th International Conference on Scientific and Statistical Database Management. pp. 55-68. 0-7695-0686-0.
- [51] **VENKATESAN, R, et al.** "*Robust image hashing*". CANADA : s.n., 2000. 2000 International Conference on Image Processing, Proceedings. Vol. 3, pp. 646-666. 0-7803-6297-7.

- [52] **DONOHO, L.D et JOHNSTONE, L.M.** "Minimax estimation via wavelet shrinkage. *Annals of Statistics*". 1998, pp. 879-921.
- [53] **GHOLAM HOSEIN, S et SUROJIT, C.** "*WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases*". 1998. Proc. 24th Int, Very Large Data Bases VLDB. pp. 428-439.
- [54] **CHAN, P et FU, C.** "*Efficient Time Series Matching by Wavelets*". Sydney, NSW, Australia : s.n., 1999. Data Engineering, Proceedings., 15th International Conference on. pp. 126-133. 0-7695-0071-4.
- [55] **JACOBS, C, FINKESTEIN, A et SALESIND, H.** "*Fast multiresolution image querying*". [éd.] ACM. 1995. Proceedings of the 22nd annual conference on Computer graphics and interactive techniques. pp. 277-286. 0-89791-701-4.
- [56] **NATSEV, A et RASTOGI, R.** "WALRUS: a similarity retrieval algorithm for image databases". [éd.] ACM Press. *Knowledge and Data Engineering, IEEE Transactions on*. pp. 395-406.
- [57] **ARDIZZONI, S et BARTOLINI, I.** "*Windsurf: Region-Based Image Retrieval Using Wavelets*". 1999. Tenth International Workshop on Database and Expert Systems Applications. pp. 167-173.
- [58] **WANG, O et WEIDERHOLD, G.** "*Wavelet-based image indexing techniques with partial sketch retrieval capability*". éd. IEEE Computer Society. 1997. Proceedings of the IEEE international forum on Research and technology advances in digital libraries. pp. 13-24. 0-8186-8010-5.
- [59] **CHAKRABARTI, K et MINOS, G.** "*Approximate query processing using wavelets*". [éd.] Springer Berlin / Heidelberg. Septembre 2001, The VLDB Journal The International Journal on Very Large Data Bases, pp. 199-223. 1066-8888 .
- [60] **ZOBEL, J et MOFFAT, A.** "Exploring the similarity space". *ACM SIGIR Forum*. 1998, pp. 18-34.
- [61] **BUCKLEY, C et WALZ, J.** "*SMART in TREC 8*". 1999. The Eight Text REtrieval Conference TREC 8. pp. 577-582.
- [62] **AREF, W.G et BARBARA, D.** "*Efficient processing of proximity queries for large databases*". Taipei, Taiwan : s.n., 1995. In proceedings of the Eleventh International Conference on Data Engineering. pp. 147-154. 0-8186-6910-1.
- [63] **HAWKING, D.** "*Relevance Weighting Using Distance Between Term Occurrences*" . TR-CS-96-08, The Australian National University . 1996.
- [64] **CLARKE, C.L.A et CORMACK, G.V.** "Shortest-substring retrieval and ranking". *ACM Transactions on Information Systems*. ACM, 2000, Vol. 18, 1, pp. 44 - 78 .

- [65] **MERCIER, A et BEIGBEDER, M.** "*Calcul de pertinence basée sur la proximité pour la recherche d'information*". s.l. : Lavoisier, 2006. pp. 43-60. 2-7462-1401-6.
- [66] **DAMERAU, F.** "*Evaluating computer-generated domain-oriented vocabularies*". 6, 1990, Information Processing and Management, Vol. 26, pp. 791-801. 0306-4573.
- [67] **MITCHELL, T.** "*Machine Learning*". s.l. : McGraw Hill, 1997. p. 414. 0070428077.
- [68] **CHURCH, K.** "*Empirical estimates of adaptation: the chance of two noriegas is closer to $p/2$ than p^2* ". [éd.] Association for Computational Linguistics. 2000. Proceedings of the 18th conference on Computational linguistics - Volume 1. pp. 180 - 186 . 1-55860-717-X.
- [69] **LITMAND, D.J et PASSONNEAU, R.J.**"*Combining Multiple Knowledge Sources for Discourse Segmentation*". [éd.] Association for Computational Linguistics. Cambridge, Massachusetts : s.n., 1995. Proceedings of the 33rd annual meeting on Association for Computational Linguistics. pp. 108 - 115 .
- [70] **BEN-HAZEZ, S et DESCLES, J.P.**"*Modèles d'exploration contextuelle pour l'analyse sémantique de textes*". Tours, France : s.n., 2001. Traitement Automatique des Langues Naturelles TALN. pp. 73-82.
- [71] **SALTON, G et BUCKLEY, C.** "*Automatic Text Decomposition Using Text Segments and Test Themes*". [éd.] HYPERTEXT '96. ACM. Washington, USA : s.n., 1996. In Proceedings of the the Seventh ACM Conference on Hypertext . pp. 53-65.
- [72] **FERRET, O et GRAU, B.**"*Utiliser des coprus pour amorcer une analyse thématique*". [éd.] Hermès. 2001. Traitement automatique des Langues. Vol. 42.
- [73] **SALTON, G.**" *Developments in automatic text retrieval*. 5023, Science, Vol. 253, pp. 974-980. 0036-8075. 1991.
- [74] **MCDONOUGH, J et JEANRENAUD, P.**"*Approaches to topic identification on the switchboard corpus*". 1994. IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 385-388.
- [75] **VAPNIK, V.** "*The nature of Statistical Learning Theory*". NewYork : Springer, 1995.
- [76] **STRANG, G et TRUONG, N.** "*Wavelets and Filter Banks*" . s.l. : Wellesley-Cambridge Press, 1996.
- [77] **JENSEN, A et LA.COUR- HARBO, A.** "*Ripples in Mathematics: the Discrete Wavelet Transform*". s.l. : Springer, 2001. p. 246 . ISBN:3540416625.
- [78] **CLARKE, C.L.A et CORMACK, G.V.**"*Relevance ranking for one to tree term queries*". 2000. Proceedings of RIAO-97, 5th International Conference ``Recherche d'Information Assistee par Ordinateur". Vol. 36, pp. 291-311.
- [79] **HAWKING, D et THISTLEWAITE, P.**"*Proximity Operators- So Near And Yet So Far*". [éd.] Gaithersburg, MD, ETATS-UNIS (National Institute of Standards and

- Technology. Gaithersburg, MD , ETATS-UNIS : s.n., 1995. the Fourth Text RETrieval Conference TREC4 no 500236 (792 p). pp. 131-143.
- [80] **RASOLOFO, Y et SAVOY, J.** "*Term Proximity Scoring for Keyword-based Retrieval Systems*". s.l. : Springer, 2003. 25th European Conference on IR Research ECIR. Vol. 2633, pp. 207-218. 978-3-540-01274-0.
- [81] **CALLON, M et TURNER, W.** "From translation to problematic networks an introduction to Co-word analysis". *Social Science Information*. 1983, Vol. 22.
- [82] **DELECROIX, B et EPPSTEIN, R.** "*Co-word analysis for the non-scientific information example of Reuters Business Briefings*". 2004, Data Science Journal, Vol. 3.
- [83] **Callon, M, Bastide, F, Turner, W.** "*Les mécanismes d'intéressement dans les textes scientifiques*", Cahiers S.T.S, N°4, 1984.
- [84] **BOURIGAULT, D et ASSADI, H.** "*Acquisition et modélisation de connaissances à partir de textes: outils informatiques et éléments méthodologiques*". Rennes, France : s.n., 1996. 10 ème congrés Reconnaissances des Formes et Intelligence Artificielle RFIA. pp. 505-514.
- [85] **ENGUEHARD, C.** "*Acquisition de terminologie à partir de gros corpus*". Nantes : s.n., 1993. Informatique et Langu Naturelle ILN. pp. 373-384.
- [86] **JACQUEMIN, C.** "*Approche mixte pour l'extraction automatique de terminologie: statistique lexicales et filtres linguistiques*". Université PARIS 7. Paris : s.n., 1994. Thèse de Doctorat en Informatique Fondamentale.
- [87] **JACQUEMIN, C.** "*Variations terminologiques: Reconnaissance et acquisition automatique de termes et de leurs variations en corpus*". Université de Nantes . 1997. Habilitation à diriger des recherches.
- [88] **SILBEZTEIN, M.** "*INTEX: an FST toolbox*". [éd.] UK Elsevier Science Publishers. 1, 2000, Theoretical Computer Science, Vol. 231, pp. 33-46. 0304-3975.
- [89] **Silberztein, M.** "NooJ: an oriented object approach", *Proceedings of the 4th and 5th INTEX workshop*, Besançon: Presses universitaires de Franche-Comté, 2004.
- [90] **DAILLE, B.** "*Approche mixte pour l'extraction automatique de terminologie statistique lexicale et filtres linguistiques*". Université Paris 7. 1994. Thèse de Doctorat en Informatique Fondamentale.
- [91] **DONOHO, D.L et BUCKHEIT, J.B.** "*Wavelab and reproducible research*". [éd.] A.Antoniadis and G.Oppenheim. Berlin : Springer-Verlag, 1995, Wavelet ans Statistics, pp. 53-81.
- [92] **YANG, Y et PEDERSEN, J.O.** "*A Comparative Study on Feature Selection in text Categorization*". San Francisco, US : s.n., 1997. Proceedings of the Fourteenth International Conference on Machine Learning. pp. 412-420. 1-55860-486-3.

- [93] **Cox T.F. et Cox M.A.A.**, "*Multidimensional Scaling*", *Second Edition*, Londres, Chapman & Hall, 2000.
- [94] **KOHONEN. T.**, "Self-organization and associative memory". *Springer series in information sciences* 8, Springer Verlag, 1988.
- [95] **YANCHUN. Z, KATSUMI. Z.**, " Web Technologies Research and Development-APWeb 2005", 7th Asia-Pacific Web Conference, Shanghai, China, Mars 2005, Proceedings.

ⁱ <http://www.adbs.fr/>