

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA
RECHERCHE SCIENTIFIQUE

**Université des Sciences et de la Technologie
Houari Boumediène (U.S.T.H.B.) Alger**

Faculté d'Electronique et d'Informatique
Département Informatique

THÈSE

Présentée pour l'obtention du diplôme de

DOCTORAT D'ÉTAT en INFORMATIQUE
Spécialité : **Informatique**

par

Omar NOUALI

Thème

***Filtrage d'Information Textuelle sur les réseaux
une Approche Hybride***

soutenue le 20 Novembre 2004, devant la commission d'examen :

Mr. N. BADACHE,	Professeur,	USTHB.	Président
Mr. Ph. BLACHE,	Dteur de Recherche,	CNRS/FRANCE	Dteur de thèse
Mme. A. AISSANI,	Professeur,	USTHB.	Co-Dteur de thèse
Mr. M. AHMED-NACER,	Professeur,	USTHB.	Examineur
Mr. M. BOUFAIDA,	Professeur,	Univ. CONSTANTINE	Examineur
Mr. M. BOUALEM ,	C. S,	FRANCE/TELECOM	Invité

Résumé

Le sujet de la thèse se situe dans la problématique globale du traitement de l'information dynamique et de l'analyse de contenu. Elle est motivée par le souci de faciliter à l'utilisateur, submergé d'informations diverses, l'accès à l'information pertinente. Plus précisément, l'objet des travaux de recherche présentés, concerne l'automatisation du processus de filtrage de l'information pertinente et personnalisée. Il s'agit d'offrir une assistance à l'utilisateur, visant à optimiser le temps consacré à la recherche et à la consultation de l'information, en prenant en compte l'importance relative de l'information et les besoins en ressources pour son traitement.

Les premières investigations dans ce travail ont été d'explorer le potentiel des techniques de plusieurs domaines de recherche liés au traitement de l'information textuelle. L'un de ces domaines concerne l'apprentissage automatique, qui constitue une phase incontournable dans la conception d'un système de filtrage automatique de l'information. Nous proposons une solution évolutive qui offre au système de filtrage la possibilité d'apprendre à partir de données ciblées (profils des utilisateurs), d'exploiter ces connaissances apprises (pour filtrer l'information) et de s'adapter à la nature de l'application (textes traités) dans le temps.

Un autre domaine concerne le traitement automatique du langage naturel. Il intervient par la nécessité d'utiliser des ressources et des traitements linguistiques dans le processus de filtrage. Sur ce volet, notre objectif est de (dé)montrer que l'intervention de connaissances et de traitements linguistiques peut considérablement améliorer les performances d'un système de filtrage de l'information. En effet, le couplage entre méthodes statistiques et symboliques (quantitatives et linguistiques) donne plus d'efficacité au filtrage. Ce constat est d'ailleurs souvent évoqué pour un grand nombre d'applications liées au traitement de l'information textuelle. Ainsi, l'apport du domaine linguistique dans notre travail se concrétise sous plusieurs aspects. D'une part, nous proposons un ensemble de connaissances linguistiques sous forme de modèles réduits (issues de modèles linguistiques de textes). Il s'agit d'un ensemble d'indicateurs sur le texte, portant sur la structure et sur le contenu. Un texte est soumis à un processus d'analyse automatique qui permet de lui associer un ensemble de termes et de propriétés linguistiques, qui servent à le caractériser et permettent de le situer par rapport à d'autres textes. Ces connaissances, classées sous plusieurs niveaux (matériel, énonciatif, structurel et syntaxique), sont indépendantes du domaine d'application. Par ailleurs, la fiabilité des traitements repose sur l'opération d'apprentissage. Dans le cadre de ce travail, l'objectif n'est pas d'effectuer une analyse complète et profonde du contenu des textes. Il s'agit d'effectuer une analyse dite partielle, s'échelonnant sur plusieurs niveaux, pour identifier certaines propriétés linguistiques. Celles-ci permettent de distinguer les différents types de textes et de classer ensuite les nouveaux textes. D'autre part, pour l'aspect sémantique, nous proposons d'utiliser un ensemble de connaissances linguistiques (réseau lexical et cooccurrence de critères) permettant d'améliorer la représentation du texte. Des termes complémentaires sont ainsi impliqués dans le processus de décision, même s'ils n'apparaissent pas explicitement dans le texte (par exemple, la substitution de certains termes par d'autres termes proches sémantiquement).

Pour la validation de notre approche, un outil d'aide à la génération d'interfaces de filtrage (baptisé GIFI) a été développé. Il est destiné à faciliter la tâche des utilisateurs développeurs dans l'élaboration de systèmes de filtrage de l'information. Il permet d'assister l'utilisateur dans le processus d'acquisition de l'application (corpus de textes) et de génération de ressources (vocabulaire lexical, propriétés linguistiques, modèle de filtrage). Il repose sur une

conception modulaire, lui permettant de s'adapter à des extensions ou à des mises à jour éventuelles. Cet outil est basé sur une architecture ouverte permettant l'ajout de composants et offrant à l'utilisateur la possibilité de choisir, à chaque étape du processus de génération, les outils à utiliser. Ainsi, cette "boîte à outils" matérialise l'implémentation d'une approche hybride de filtrage de l'information. Elle repose sur le principe d'une analyse partielle utilisant un ensemble de connaissances, où le repérage de propriétés linguistiques permet, d'une part, d'améliorer la représentation des textes, et d'autre part un filtrage de meilleure qualité.

Pour l'évaluation de notre approche et afin de statuer sur sa faisabilité et sur son apport en terme d'efficacité, nous l'avons expérimentée sur une application pratique de filtrage de l'information : *filtrage du courrier électronique*. La période actuelle voit une prolifération colossale et démesurée des courriers électroniques non sollicités et indésirables (appelés *Spams*). Paradoxalement, au moment où le courrier électronique s'impose comme le moyen de communication incontournable pour les entreprises, les institutions académiques et même pour les particuliers, le problème des courriers indésirables atteint des proportions intolérables. Ce problème devient très sérieux pour les utilisateurs du courrier électroniques et engendre des pertes considérables, en temps et en argent, pour les entreprises. A travers les différentes expériences réalisées, nous avons montré l'applicabilité et l'adaptabilité d'une approche hybride au processus de filtrage de l'information. En effet, les résultats obtenus sur le corpus de messages utilisé, nous ont permis de valider l'intérêt des connaissances linguistiques et de l'apprentissage automatique pour l'amélioration des performances d'un système de filtrage de l'information.

Mots clés :

Filtrage de l'information, apprentissage automatique, propriétés linguistiques, modèles linguistiques réduits, spam.

Table des Matières

Remerciements
Résumé
Table des matières

Introduction Générale..... 1

Partie 1 : Etat de l'art

Chapitre I

Linguistique informatique et Analyse automatique de textes 7

1 La linguistique formelle et le langage..... 7

1.1 Etude du langage..... 7

1.1.1 Approche structuraliste de Saussure..... 8

1.1.2 Approche distributionnaliste de Harris..... 8

1.1.3 Approche Générative de Chomsky..... 9

1.1.4 Approches basées contraintes..... 10

1.2 La linguistique formelle..... 14

1.2.1 Analyse morphologique..... 14

1.2.2 Analyse lexicale..... 15

1.2.3 Analyse syntaxique..... 16

1.2.4 Analyse Sémantique..... 19

1.2.5 Analyse pragmatique..... 20

1.2.6 Représentation de sens..... 21

2 Linguistique de corpus.....	26
2.1 Corpus.....	26
2.2 Extraction de termes.....	27
2.3 Etiquetage morphosyntaxique.....	28
2.3.1 Désambiguïsation morphosyntaxique.....	29
2.3.2 Quelques étiqueteurs catégoriels.....	29
2.3.3 Autres types d'étiquetage.....	30
2.4 Analyse syntaxique.....	30
2.4.1 Ambiguïté syntaxique.....	31
2.4.2 Analyseurs syntaxiques.....	31
2.5 Analyse sémantique.....	34
2.5.1 Ressources lexicales.....	34
2.5.2 Relations sémantiques.....	36
2.6 Analyse pragmatique.....	36
3 Analyse automatique de textes.....	38
3.1 Exemples d'applications.....	38
3.2 Approches d'analyse.....	39
3.2.1 Analyse globale.....	39
3.2.2 Analyse locale.....	39
3.2.3 Analyse de surface ou partielle.....	40
3.3 Typologie des textes.....	42
3.3.1 Approche Biber.....	43
3.3.2 Approche Bronckart.....	45
3.3.3 Approche Bergounioux et d'Abert.....	45
4 Filtrage d'information et langage naturel.....	46
4.1 Problèmes de la langue.....	46
4.2 Approche non linguistique.....	48
4.3 Approche linguistique.....	48
4.4 Outils TAL.....	49
5 Conclusion.....	49

Chapitre II

Filtrage d'Information Textuelle..... 52

1 Généralités.....	53
1.1 Histoire.....	53
1.1.1 Systèmes de veille économique.....	53
1.1.2 Diffusion sélective d'information.....	54
1.1.3 Naissance de la notion de filtrage d'information.....	54
1.2 Définitions.....	55
1.3 Domaines d'application.....	59
1.4 Quelques travaux précédents.....	60

2 Un Système de filtrage.....	63
2.1 Architecture de base.....	63
2.2 Caractéristiques.....	64
2.3 Evaluation.....	65
2.3.1 Notion de pertinence (Relevance).....	65
2.3.2 Critères d'évaluation.....	66
2.3.3 Programmes d'évaluation.....	67
3 Approches pour le filtrage d'information.....	69
3.1 Méthodes classiques.....	69
3.1.1 Filtrage par chaînes de caractères (Fulltext).....	69
3.1.2 Filtrage par langage restreint.....	69
3.1.3 Filtrage par regroupements (Clustering).....	69
3.1.4 Méthodes booléennes.....	70
3.1.5 Méthodes utilisant la logique floue.....	74
3.1.6 Méthodes vectorielles.....	74
3.1.7 Méthodes probabilistes.....	77
3.2 Méthodes symboliques.....	78
3.2.1 Filtrage par règles.....	78
3.2.2 Filtrage textuel linguistique.....	79
3.3 Réseaux de neurones.....	83
3.3.1 Définition.....	83
3.3.2 Modélisation formelle.....	83
3.3.3 Modèles.....	85
3.4 Méthodes collaboratives.....	88
3.4.1 Filtrage collaboratif.....	88
3.4.2 Filtrage par agents.....	89
3.5 Modélisation des intérêts de l'utilisateur.....	90
3.5.1 Etude d'observation.....	90
3.5.2 Modélisation par mots clés et par document.....	91
3.5.3 Relevance feedback.....	91
3.5.4 Annotations collaboratives.....	92
3.5.5 Anti-profil.....	92
4 Conclusion.....	93

Chapitre III

Apprentissage et Classification Automatique de textes..... 95

1 Système d'apprentissage.....	95
2 Classification automatique.....	96
2.1 Classification supervisée.....	96
2.2 Classification non supervisée.....	97

3. Analyse et représentation du contenu des textes.....	97
3.1 Les modèles de représentation de textes.....	97
3.1.1 Représentation non linguistique ou « sac de mots ».....	98
3.1.2 Représentation linguistique.....	99
3.2 Techniques de sélection et de réduction du vocabulaire de représentation.....	99
3.2.1 Les mots les plus fréquents.....	100
3.2.2 Fréquence de documents (DF).....	100
3.2.3 Information Gain (IG).....	101
3.2.4 Correlation Coefficient CHI (χ^2).....	101
3.2.5 Mutual Information (MI).....	102
3.2.6 Analyse en composantes principales (ACP).....	103
3.2.7 Latent Semantic Analysis (LSA).....	103
3.3 Les techniques de pondération de poids ou codage.....	103
3.4 Mesure de similarité entre textes.....	105
4 Les approches probabilistes.....	107
4.1 Approche Naïve Bayes.....	107
4.2 Modèles de Markov Cachés (MMC).....	108
4.3 Machines à Vecteurs Supports (MVS).....	109
5 Méthode des plus proches voisins.....	110
6 Méthode de Rocchio.....	112
7 Apprentissage symbolique.....	113
7.1 Les arbres de décision.....	113
7.2 Les règles de décision.....	116
8 Apprentissage adaptatif.....	117
8.1 Réseaux de neurones.....	117
8.2 Algorithmes génétiques.....	122
8.2.1 Concepts de base.....	123
8.2.2 Fonctionnement général d'un algorithme génétique.....	125
8.2.3 Algorithmes génétiques et filtrage d'information.....	126
9 Classification par recherche directe.....	126
10 Classification ascendante.....	128
11 Classification descendante.....	130
12 Conclusion.....	130

Partie 2 : Implémentation & Evaluation

Chapitre IV

Architecture de filtrage et Générateur d'interfaces.....	133
1 Architecture de base du système de filtrage.....	133
2 Identification de la langue.....	135
3 Etiquetage.....	135
3.1 Catégoriseur Brill.....	136
3.1.1 Apprentissage.....	136
3.1.2 Codage.....	137
3.2. Base de connaissance INALF.....	138
3.2.1 Jeu d'étiquettes.....	139
3.2.2 Corpus-échantillon.....	139
3.2.3 Base de connaissances.....	140
4 Normalisation.....	142
5 Analyseur linguistique.....	143
5.1 Motivation.....	144
5.2 Travaux antérieurs.....	145
5.3 Les modèles linguistiques réduits.....	146
5.3.1 Modèle lexical.....	146
5.3.2 Modèle concernant la mise en forme matérielle (l'architecture du texte).....	147
5.3.3 Modèle énonciatif.....	146
5.3.4 Modèle structurel.....	148
5.3.5 Modèle syntaxique.....	149
6 Expansion de la représentation.....	151
6.1 Motivation.....	151
6.2 Approche pseudo sémantique proposée.....	151
6.3 Réseau lexical.....	152
6.3.1 Construction du réseau.....	153
6.3.2 Apprentissage.....	155
6.4 Cooccurrence de critères.....	155
6.5 Processus de filtrage sémantique.....	155
7 Processus de filtrage.....	156
8 GIFI, un assistant à la génération d'interface de filtrage.....	157
8.1 Motivation.....	157
8.2 Acquisition de textes.....	158
8.3 La sélection des critères de filtrage.....	158
8.3.1 Vocabulaire Lexical.....	158
8.3.2 Caractéristiques supplémentaires.....	158
8.4 Modèle de filtrage.....	160

8.4.1 Modèle de base adopté.....	160
8.4.2 Apprentissage	161
8.5 Description de l'interface graphique.....	164
9 Conclusion.....	170

Chapitre V

Filtrage, Typologie et Caractéristiques des messages électroniques 171

1 Messagerie électronique.....	171
1.1 Anatomie ou format d'un message électronique.....	171
1.1.1 Partie structurée.....	171
1.1.2 Partie non structurée.....	172
2 Filtrage d'emails.....	173
2.1 Définition de filtrage de messages.....	173
2.2 Quelques systèmes de filtrage du courrier électronique.....	174
2.3 Stratégies de filtrage.....	176
2.4 Approches pour traiter le courrier électronique.....	176
2.4.1 Classification de textes.....	176
2.4.2 Extraction d'information.....	177
2.4.3 Raisonnement par cas.....	177
2.4.4 Recherche d'information.....	177
2.4.5 Question/réponses.....	178
3 Typologie et choix du corpus.....	178
4 Caractéristiques des mails.....	179
4.1 Caractéristiques lexicales.....	179
4.2 Caractéristiques supplémentaires.....	180
5 Filtrage automatique d'emails, une approche adaptative et multi niveaux.....	182
5.1 Architecture générale du système.....	183
5.1.1 Pré-traitement.....	183
5.1.2 Analyseur automatique.....	184
5.1.3 Processus de filtrage.....	184
5.2 Niveaux de filtrage.....	185
5.3 Connaissances utilisées.....	186
5.4 Correction.....	187
6 Evaluation.....	187
6.1 Le corpus.....	188
6.2 Critères d'évaluation.....	188
6.3 Expériences.....	190
6.3.1 Performances en fonction des caractéristiques lexicales seulement	190

6.3.2 Performances en fonction des mots composés.....	190
6.3.3 Performances en fonction du nombre de caractéristiques linguistiques.....	191
6.3.4 Mesurer l'importance et le rôle de l'apprentissage assisté.....	192
6.3.5 Mesurer l'importance et le rôle du filtrage avec propagation (horizontal).....	193
6.4 Discussion.....	193
7 Conclusion.....	194
Conclusion.....	197
Bibliographie Personnelle.....	201
Bibliographie.....	203
Annexes.....	224
Annexe A : Jeu d'étiquettes.....	224
Annexe B : Liste des mots vides.....	226
Annexe C : Modèles linguistiques réduits.....	230
Annexe D : Système expert.....	242
Liste des Figures.....	245
Liste des Tables.....	247
Liste des Algorithmes.....	248
Glossaire.....	249