N° d'ordre : 01/2004-E/IN

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.	Examinateur
Mr. M. BOUFAIDA,	Professeur,	Univ. CONSTANTINE	Examinateur
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 "boite 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

Partie 2 : Implémentation & Evaluation

Chapitre IV
Architecture de filtrage et Générateur d'interfaces…… 133

Chapitre V
Filtrage, Typologie et Caractéristiques des messages électroniques 171