Filtrage cognitif de courriels

Omar Nouali[1], Bernard Toursel[2],
CEntre de Recherche sur l'Information Scientifique et Technique, CE.R.I.S.T

Abstract

In this paper, we present filtering process automation by taking of account the relative importance of information and the requirements in linguistic resources for its treatment. In this area we aim to evaluate if using linguistic information and analysis can improve the performance of a filtering system. Indeed, as well as using lexical characteristics, we use a set of criteria which are cues related to the email structure and content. The approach architecture is modelled using agents to offer a saving of time compared to a sequential algorithmic solution. At the end, to measure the approach performances, we illustrate and discuss the results obtained by experimental evaluations.

Keywords : Information filtering, machine learning, linguistic agents, filtering criteria.

1. Introduction

Dans le cadre de ce travail, nous nous intéressons à une application pratique du domaine de filtrage de l'information : le courriel. En effet, aujourd'hui, le courriel est devenu un moyen rapide et économique pour échanger des informations. Cependant, les utilisateurs se retrouvent assez vite submergés de messages.

Certains systèmes de filtrage de courriels existants sont basés seulement sur le traitement de la partie structurée (par exemple, dans le cas de messages non sollicités appelés *spam*, le filtrage opère généralement sur les adresses émettrices en se basant sur une liste noire des *spammeurs*), et d'autres sont basés sur un balayage superficiel de la partie texte du message en permettant aux utilisateurs d'écrire manuellement des règles logiques de filtrage à base de mots clés (Nouali *et al.* 2005 ; Gaussier et Stéfanini 2003 ; Kilander *et al.*1997 ; Belkin et Croft 1992).

Le problème, avec ces systèmes, est qu'ils ne sont pas précis car l'aspect sémantique est négligé et le processus de filtrage est une classification basée simplement sur une propriété lexicale, la présence ou l'absence de mots-clés que l'utilisateur doit indiquer au logiciel. Tout filtrage effectué sur ces bases ne peut que contenir une part d'imprécision.

Pour améliorer ces systèmes, une solution est que le processus de filtrage exploite le maximum d'informations et prendre en compte le besoin en ressources linguistiques pour son traitement.

Une solution qui tente de faire une analyse de toutes les informations présentes dans un courriel est complexe et difficile. En effet, l'interprétation d'un énoncé en texte libre nécessite des connaissances linguistiques et extralinguistiques (connaissances du monde, conventions,

[1] Laboratoire des Logiciels de base, CE.R.I.S.T, Ben Aknoun, Alger, Algérie, onouali@mail.cerist.dz

[2] LIFL, UPRESA CNRS 8022, Université des Sciences et Technologies de Lille, France, toursel@lifl.fr

etc.). Toutes ces connaissances sont difficiles à encoder dans les systèmes d'analyse automatique, du fait de leur complexité et de leur quantité (Nouali *et al.* 2005 ; Yang et Pedersen 1997).

Dans ce contexte, nous ne cherchons pas à faire une analyse complète et profonde du contenu des courriels, mais plutôt, une analyse partielle du contenu qui permet de dégager des informations linguistiques portant sur la structure et le contenu des courriels. Ce qui devrait permettre de distinguer les différents courriels, de situer un courriel par rapport aux autres et d'avoir un filtrage de courriels meilleur en qualité.

Nous visons principalement, par le recours à des informations linguistiques dans le domaine de filtrage, les objectifs suivants :

- Définir des profils correctement discriminants et non ambigus.
- Représentation étendue des courriels à filtrer.
- Mettre en rapport des formulations différentes mais sémantiquement proches afin d'augmenter les chances d'apparier un profil et un courriel à filtrer.

Pour la réalisation de notre approche de filtrage, nous proposons une implantation basée agents, entités indépendantes ayant chacune une tâche de filtrage bien précise à effectuer, interagissant selon des modes de coopération, de concurrence et de cœxistence. Ce qui offre un gain de temps par rapport à une solution algorithmique séquentielle, et permet d'avoir une approche ouverte et dynamique : en effet, de nouveaux traitements (critères de filtrage) peuvent être ajoutés au fil du temps, et le système de filtrage doit donc être capable de s'adapter pour intégrer ces nouveaux traitements de façon à augmenter l'efficacité globale, et ce, sans modifier l'existant.

Pour l'évaluation de cette approche et afin de statuer sur sa faisabilité et sur son apport en terme d'efficacité, nous avons mené un ensemble d'expériences sur quelques courriels bien particuliers : les messages indésirables (appelés *Spam*).

2. Architecture de base

Notre architecture de filtrage est composée principalement de 3 grands types d'agents (figure 1) :

- **Agent Document** non permanent qui se crée à chaque arrivée d'un nouveau courriel dans le système. Il se charge de piloter et de coordonner les opérations d'analyse et de filtrage. Pour préparer l'opération d'analyse, il effectue sur le courriel une phase d'étiquetage et une phase de normalisation. La phase d'étiquetage consiste à apposer une étiquette grammaticale à chaque segment de courriels (utile pour l'extraction de certains indices linguistiques). La phase de normalisation permet de réduire les variantes morphologiques des mots en une forme commune (rendre les verbes à l'infinitif, supprimer les formes plurielles, etc.).
- **Agent Critère** non permanent qui implémente un traitement spécifique sur le contenu du courriel. Il se charge d'analyser le courriel et d'extraire un ensemble de propriétés permettant de le caractériser.
- **Agent profil** permanent qui se charge de deux processus : préfiltrage et filtrage. Le processus de préfiltrage consiste à éliminer, en premier, les courriels qui ont des caractéristiques différentes de celles attendues par l'utilisateur. Le processus de filtrage consiste à décider si le courriel retenu après préfiltrage correspond ou pas au profil.

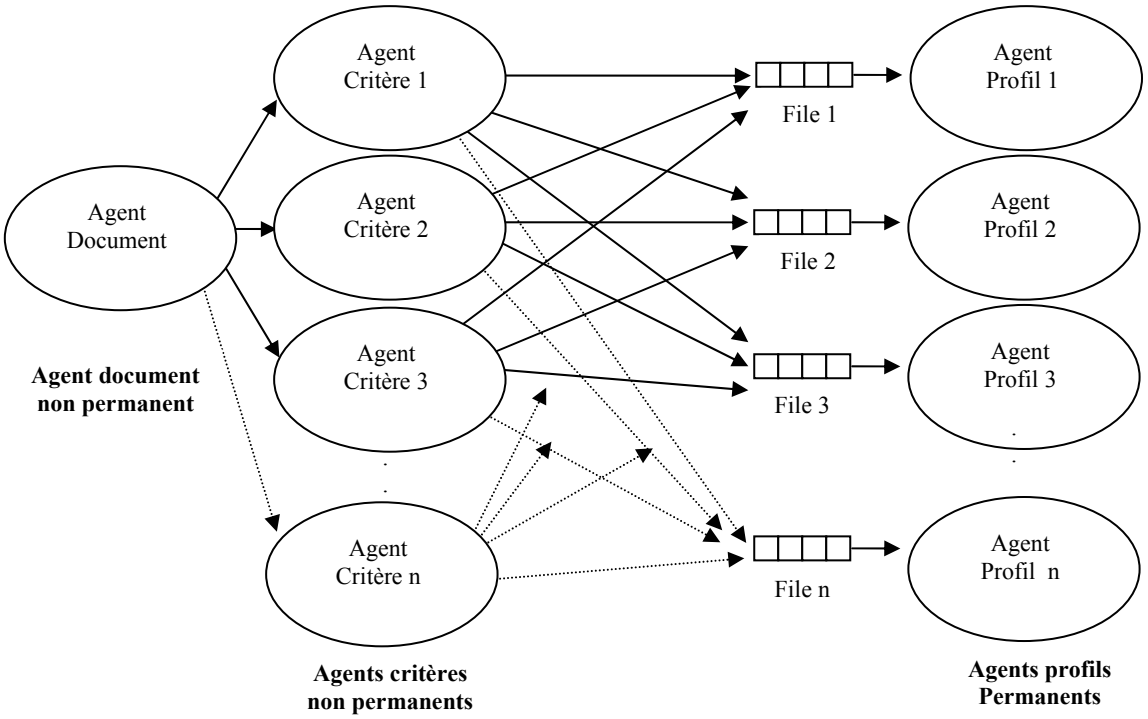

Figure 1. Architecture multi-agents de filtrage

Il existe plusieurs types d'agents critères (Nouali *et al.* 2005) :

- **Agent « critères avancés »** pour identifier la langue du courriel, l'auteur, l'adresse Internet, etc. Ces critères avancés sont nécessaires à l'opération de pré-filtrage permettant d'éliminer certains courriels non pertinents.
- **Agent lexical** pour identifier les propriétés lexicales du courriel (mots simples et mots composés).
- **Agent architectural** pour vérifier si le courriel présente des propriétés architecturales du genre : titres, introduction, conclusion, images, dessins, tableaux, caractères non alphanumériques (&, $, %, *, #, etc.), caractères numériques...
- **Agent structurel** pour identifier relations structurelles entre segments : addition, analogie, but, cause, énonciation, exemple, hypothèse, intensité, opposition, temps …).
- **Agent syntaxique** pour identifier les propriétés syntaxiques utilisées dans le courriel (formes passive/active, négation, interrogation…).
- **Agent énonciatif** pour identifier les propriétés énonciatives utilisées dans le courriel (pronoms personnels, discours rapporté direct, formes verbales, formes temporelles...).
- **Agent pseudo-sémantique** permettant d'étendre et d'améliorer la représentation du courriel (mots sémantiquement proches). En effet, après l'étape d'analyse linguistique effectuée par les agents critères (lexical, architectural, structurel, syntaxique et énonciatif), l'agent profil lance en premier l'agent pseudo sémantique pour compléter la représentation du courriel et par conséquent mesure la similarité pour décider si le courriel lui correspond ou pas.

Chaque agent critère diffuse ses résultats (caractéristiques linguistiques) aux agents profils et il se détruit.

3. Acquisition de connaissances

Pour le bon fonctionnement des différents agents, nous avons développé un ensemble d'outils d'apprentissage permettant l'acquisition automatique des connaissances.

3.1. Connaissances de l'agent lexical

Pour l'extraction automatique des connaissances de l'agent lexical, nous avons développé un outil permettant l'extraction des connaissances lexicales simples et la sélection des connaissances lexicales complexes.

La sélection des mots composés est basée sur une analyse syntaxique et grammaticale de surface (Nouali et Krinah 2006). Il s'agit de repérer des locutions (suites de mots) correspondant à une certaine syntaxe (généralement des groupes nominaux) et susceptibles de constituer des multi-termes porteurs de sens (figure 2).

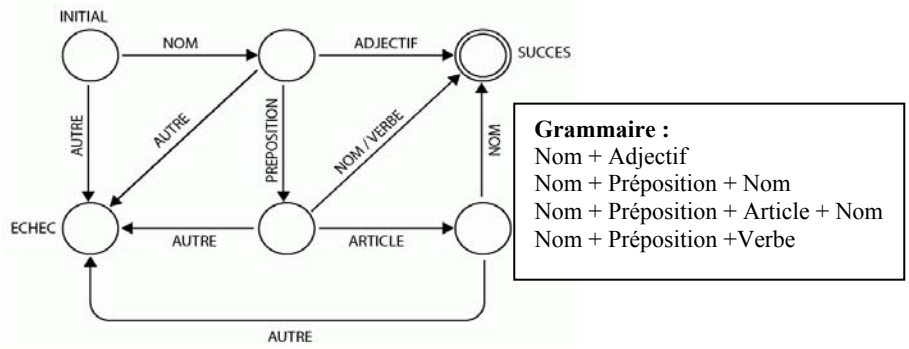

Figure 2. Grammaire d'extraction de mots composés

La validation des mots composés obtenus peut se faire de deux manières :

– la première est statistique : elle consiste d'abord à fixer un seuil de validité, ensuite à calculer pour chaque multi-termes obtenu sa fréquence d'apparition dans le courriel traité, l'objectif étant de ne retenir que les multi-termes dont le nombre d'occurrences atteint ou dépasse le seuil choisi ;

– la deuxième est manuelle : elle est basée sur une analyse des résultats obtenus. L'utilisateur pourrait intervenir, soit pour valider un multi-termes jugé pertinent et non retenu par le traitement automatique du fait de sa faible fréquence d'apparition, soit pour annuler une séquence retenue par l'outil mais qui pour l'analyseur humain ne porte pas de cohérence sémantique.

3.2. Réseau de co-occurrences lexicales

Nous avons développé un outil permettant la construction d'un réseau de co-occurrences lexicales implicite propre à l'utilisateur (profil) (Nouali et Blache, 2006). Ce réseau lexical permet, lors d'une opération de filtrage, d'améliorer la représentation du courriel en prenant en considération les termes qui existent dans le courriel et qui n'existent pas dans le profil. Il s'agit de les remplacer par des termes du profil sémantiquement proches. En effet, certains courriels pertinents sont constitués de termes statistiquement insignifiants ce qui oblige l'opération de filtrage à les rejeter.

Les nœuds du réseau représentent l'ensemble des termes pertinents existants dans le profil. Les liens entre termes représentent la cooccurrence (w_{ij} représente la cooccurrence de deux termes t_i et t_j) (figure 3).

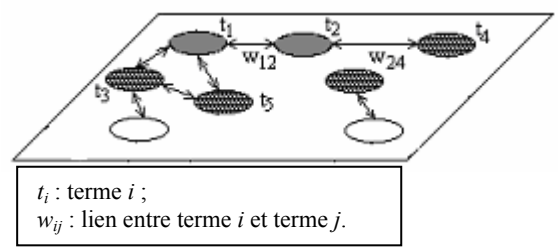

Figure 3. Réseau de co-occurrences

Notre processus de construction et de mise à jour du réseau est comme suit (Nouali et Blache 2006) :

- Constituer un ensemble de courriels pour le profil considéré (le profil *Spam* par exemple).

- Construire la matrice (courriels, termes).

- Calculer la similarités des termes deux à deux. En effet, nous choisissons que les termes n'appartenant pas au profil considéré. Ensuite, nous calculons la similarité entre ces nouveaux termes et ceux qui existent dans le profil.

- Regrouper ensemble les termes les plus proches.

Le calcul de la similarité entre termes se mesure à l'aide de la formule *Cosine* qui calcule le cosinus de l'angle entre leurs vecteurs respectifs. Plus le cosinus de l'angle entre les deux vecteurs est proche de 1, plus les vecteurs sont proches ce qui implique une plus grande ressemblance entre les deux termes. La figure 4 donne un aperçu du réseau généré automatiquement par apprentissage automatique (cas du profil *SPAM*) :

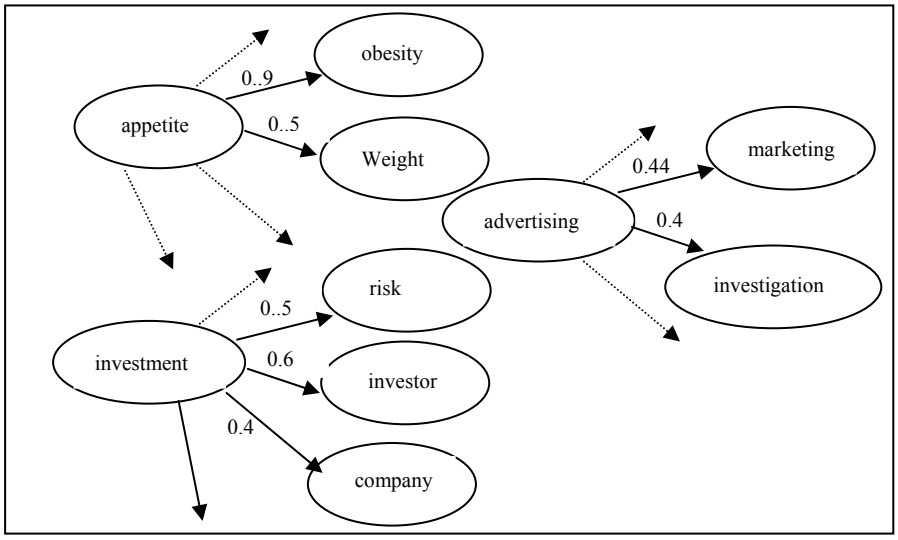

Figure 4. Aperçu du Réseau de co-occurrences SPAM

4. Évaluation

Nous avons mené des tests pour mesurer les performances du système de filtrage du point de vue *précision* et *rappel*, mesurer l'importance et le rôle des caractéristiques linguistiques dans la représentation des courriels, et montrer comment l'opération d'apprentissage agit sur l'efficacité du filtrage.

Pour effectuer nos tests nous avons travaillé avec un corpus de 1800 messages construit à partir d'un ensemble de messages que nous avons collectés durant une bonne période de temps, contenant 1100 mails de classe *spam* et 700 non *spam*.

Nous avons divisé le corpus en une base d'apprentissage et une base de tests selon le découpage suivant :

- Base d'apprentissage : 750 *spam* et 500 *non spam*.
- Base de test : 350 *spam* et 200 *non spam*.

Nous présentons et nous discutons, dans ce qui suit, les résultats des performances, au cours d'une évaluation quantitative de notre approche de filtrage, dans plusieurs cas de configuration :

- performances en fonction des caractéristiques lexicales composées (tableau 1) ;
- performances en fonction des caractéristiques linguistiques (tableau 2) ;
- mesurer l'importance et le rôle de l'information sémantique (tableau 3) ;
- évaluation du temps de réponse (figure 5).

Caractéristiques des profils	**Performances**	
	Précision	Rappel
Mots simples	79%	95%
Mots simples + Mots composés	80,2%	97,5%
Mots simples + Mots composés + Pondération	85%	98,4%

Tableau 1 : Performances en fonction des caractéristiques lexicales composées

Nous constatons que les mots composés corrèlent avec les courriels du profil Spam considéré, mais statistiquement sont insignifiants (valeur faible). Les résultats des tests étaient meilleurs lorsque nous avons modifié l'importance de ces différents mots composés, en leur attribuant une forte valeur du poids.

Nous avons ajouté des caractéristiques supplémentaires (un ensemble d'indicateurs linguistiques automatisables sur le courriel), que nous avons défini et identifié, aux caractéristiques lexicales du profil Spam considéré. Par exemple, le domaine (.com, .gov, .edu, .fr, etc.) , la longueur du message, le type du contenu (html/txt), la langue du message, les mots en majuscule, les abréviations, les caractères non alphanumérique ($, !, #, %, *, &, etc.), les caractères numériques, la taille des phrases, l'horaire d'envoi (nuit/jour), etc.

Performances	**Caractéristiques lexicales seulement**	**Caractéristiques lexicales + Caractéristiques supplémentaires**
Précision	85%	89.6%

Tableau 2 : Performances en fonction des caractéristiques

Nous constatons que les performances de filtrage sont légèrement améliorées. Ceci s'explique par le fait que les courriels rejetés par le processus de filtrage (le cas de caractéristiques lexicales seulement) par absence de mots clés ou valeur très faible, sont acceptés cette fois ci, et ceci à cause de la présence de certaines caractéristiques supplémentaires.

En effet, ces différentes caractéristiques supplémentaires ajoutées aux caractéristiques lexicales permettent d'augmenter les chances d'apparier un courriel et un profil.

Caractéristiques	**caractéristiques + réseau lexical**	
sans feedback	**sans feedback**	**avec feedback**

48%	51%	62%

Tableau 3 : Performances en fonction du réseau lexical

Nous constatons que le réseau lexical améliore légèrement les résultats de filtrage. En effet, les mots inconnus sont remplacés par d'autres mots plus proches du réseau lexical, ce qui permet d'augmenter le taux de précision. Mais, l'amélioration de la qualité des résultats nécessite le lancement de l'apprentissage du réseau après chaque session de filtrage.

Pour mesurer le temps d'attente par l'utilisateur, nous simulons deux machines (mono-processeur) sur lesquelles les agents fonctionnent : *machine1* permet une exécution séquentielle et *machine2* permet une exécution simultanée de plusieurs agents.

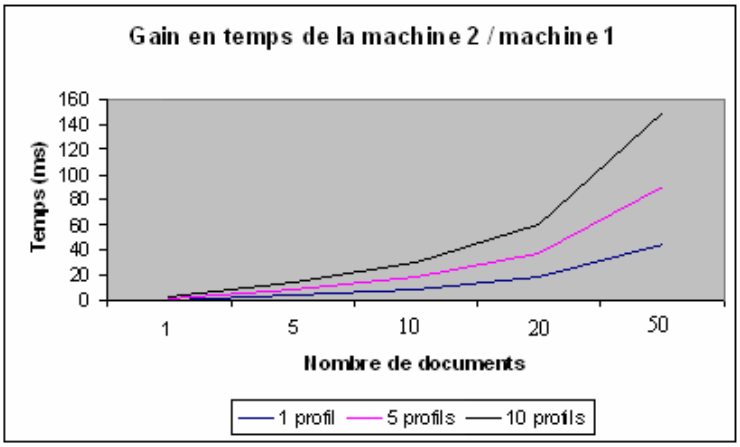

Figure 5. Gain en temps en fonction du nombre de profils et courriels

La machine2 présente un temps de réponse meilleur que la machine1 en variant le nombre de profils et le nombre de courriels. Le filtrage basé agents s'adapte mieux dans un environnement parallèle.

5. Conclusion

Cet article propose une approche évolutive qui s'adapte à la nature des courriels au cours du temps et qui exploite le maximum d'informations pour filtrer le courriel que nous avons modélisé à l'aide d'agents pour offrir un gain de temps par rapport à une solution algorithmique séquentielle. Elle fait essentiellement appel à deux ressources et traitements linguistiques.

Un traitement statistique d'un corpus Spam nous a permet de proposer un certain nombre de critères spécifiques aidant à améliorer les résultats de filtrage. Les résultats semblent satisfaisants, mais nous ne nous pouvons pas affirmer qu'ils constituent une connaissance suffisante de discrimination de courriels de type *spam*. Néanmoins, la probabilité d'avoir un courriel de type *spam* est plus forte quand ces critères sont vérifiés.

Un réseau lexical permet au système de relier un message à un profil même s'ils n'ont pas de mots clés en commun. En effet, les mots inconnus sont remplacés par d'autres mots plus proches, ce qui permet d'augmenter le taux de *rappel* tout en gardant une bonne *précision*.

A travers les différentes expériences réalisées, la conclusion que l'on peut évoquer est que l'apprentissage automatique est un passage obligé dans la conception et l'amélioration des performances d'un système de filtrage d'information dynamique et les méthodes linguistiques combinées aux méthodes statistiques semblent prometteuses pour avoir un filtrage cognitif efficace.

Références

BELKIN N.J. et CROFT W.B. (1992), « Information filtering and information retrieval : two sides of the same coin? », in *Communications of the ACM* 35/12 : 29-38.

BOUZGHOUB M. et KOSTADINOV D. (2005), « Personnalisation de l'information : Aperçu de l'état de l'art et définition d'un modèle flexible de profils », in *Actes de la 2ème Conférence en Recherche d'Information et Applications (CORIA'05)*, France : 201-218.

BURKE R. (2002), « Hybrid recommender systems : Survey and experiments », in *User Modeling and User Adapted Interaction* 12/4 : 331-370.

CROFT W.B. (1993), « Knowledge-based and Statistical approaches to Text Retrieval », in *IEEE EXPERT* 8/2 : 8-12.

GAUSSIER E. et STEFANINI M.H. (2003), « Assistance intelligente à la recherche d'information », in *Traité des sciences et techniques de l'information, Hermes Lavoisier* : 255-282.

KILANDER F., FAHRAEUS E. et PALME J. (1997), « Intelligent Information Filtering », *Technical report 97-002*, Dpt of Computer and Systems Sciences, Stockholm University.

NGUYEN A.T., DENOS N. et BERRUT C. (2006), « Modèle d'espaces de communautés basé sur la théorie des ensembles d'approximation dans un système de filtrage hybride », in *Actes de la 3e Conférence en Recherche Information et Applications, CORIA'06*, Lyon.

NOUALI O. et BLACHE P. (2006), « Generation Tool of Information filtering interface », in *1st International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006*, Merida`s Conference Hall, Spain.

NOUALI O. et KRINAH A. (2006), « Improvement of a retrieval and filtering systems by an automatic multi words extraction tool », in *CSIT2006 4th International conference on computer science and information technology*, Applied Science University, Amman, Jordan.

NOUALI O., REGNIER A. et BLACHE P. (2005), « Classification de courriers électroniques », in *Revue d'intelligence Artificielle, RIA* 19/6, *Hermes Lavoisier*.

YANG Y. et PEDERSEN J. O. (1997), « A comparative Study on Feature Selection in Text Categorization », in *International Conference on Machine Learning, ICML*, Nashville, TN, USA.