Classification Automatique de messages : une approche hybride

O. Nouali

(1) Laboratoire des Logiciels de base, CE.R.I.S.T,
Rue des 3 frères Aïssiou, Ben Aknoun, Alger, Algérie
Fax : 213 21 912126 - Tél. : 213 21 916211
(2) LPL- Université de Provence
29, Av. Robert Schuman, F-13621 Aix-en-Provence, France.
Fax: +33 (0).42.59.50.96- Tél.: +33 (0) 42.95.36.23
E-mail : onouali@mail.cerist.dz

Mots-clefs – Keywords

Filtrage d'information, e-mail, classification de messages, propriétés linguistiques, réseaux de neurones, filtrage d'email.

Information filtering, e-mail, information classification, linguistic proprieties, neural network, Filters, e-mail filtering.

Résumé - Abstract

Les systèmes actuels de filtrage de l'information sont basés d'une façon directe ou indirecte sur les techniques traditionnelles de recherche d'information (Malone, Kenneth, 1987), (Kilander, Takkinen, 1996). Notre approche consiste à séparer le processus de classification du filtrage proprement dit. Il s'agit d'effectuer un traitement reposant sur une compréhension primitive du message permettant d'effectuer des opérations de classement. Cet article décrit une solution pour classer des messages en se basant sur les propriétés linguistiques véhiculées par ces messages. Les propriétés linguistiques sont modélisées par un réseau de neurone. A l'aide d'un module d'apprentissage, le réseau est amélioré progressivement au fur et à mesure de son utilisation. Nous présentons à la fin les résultats d'une expérience d'évaluation.

The current approaches in information filtering are based directly or indirectly on the traditional methods of information retrieval (Malone, Kenneth, 1987), (Kilander, Takkinen, 1996). Our approach to email filtering is to separate classification from filtering. Once the classification process is complete, the filtering takes place. This paper presents an approach to classify e-mail, based on linguistic features model. The main feature of the model is its representation by a neural network and its learning capability. At the end, to measure the approach performances, we illustrate and discuss the results obtained by experimental evaluations.

1 Introduction

Le courrier électronique est le mode de communication grandissant le plus rapidement aujourd'hui. Cependant, les utilisateurs d'Internet se retrouvent assez vite submergés de quantités astronomiques de messages dont le traitement nécessite un temps considérable.
Dans la pratique, la plupart des systèmes de filtrage du courrier électronique existants enregistrent des lacunes ou faiblesses sur l'efficacité du filtrage. Certains systèmes sont basés seulement sur le traitement de la partie structurée (un ensemble de règles sur l'entête du message), et d'autres sont basés sur un balayage superficiel de la partie texte du message (occurrence d'un ensemble de mots clés décrivant les intérêts de l'utilisateur).

Cet article décrit une approche pour filtrer le courrier électronique. Elle suppose que l'automatisation du processus de filtrage nécessite une phase importante de classification, qui constitue en quelque sorte un pré-filtrage.
Après une présentation de l'architecture et du fonctionnement général du système de classification, une description du réseau de neurones adopté pour notre application est donnée. Nous présentons à la fin les résultats d'une expérience d'évaluation.

2 Approche proposée

Une idée pour classer les messages, est de créer des espaces de messages (un espace pour chaque type). Et chaque nouveau message se trouvant être proche des messages de l'un des espaces définis, est alors considéré comme pertinent pour cet espace.
Après étude des différentes approches possibles pour le domaine du filtrage, ainsi que les différentes propriétés linguistiques des courriers électroniques (Kilander, Takkinen, 1996), (Nouali 2000), (Kosseim, Lapalme 2001), nous proposons une approche qui consiste à séparer la classification du filtrage. En effet, afin de mieux classer un message, notre système implémente un modèle linguistique qui représente et modélise une typologie de messages, dont la connaissance est construite initialement sur la base d'analyse de traits linguistiques associés à chaque espace de messages. Ce modèle sera enrichi progressivement au fur et à mesure de son utilisation. De plus, pour le processus de filtrage, les intérêts de l'utilisateur sont décrits par des profils et ils sont introduits dans le système sous forme de **mots clés** et/ou sous forme de **texte.**

3 Architecture et Fonctionnement global du Système

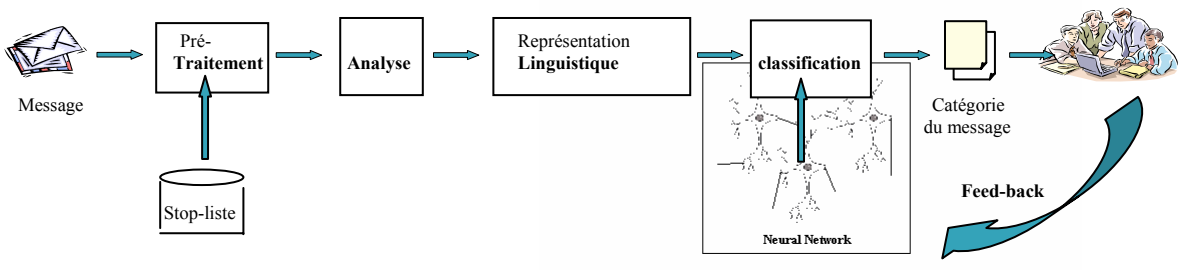

Figure 1 : une architecture globale du système

Après récupération de chaque message, un module de pré-traitement est lancé pour isoler les différents champs et le préparer aux différentes étapes ultérieures.

Après l'étape de pré-traitement, le message subit une analyse linguistique, qui est indépendante des intérêts de l'utilisateur. Il s'agit d'effectuer un traitement reposant sur une compréhension primitive du message. Le but de l'analyse est de déterminer les informations pertinentes à représenter, c'est à dire déterminer les informations véhiculées par ce message. Il s'agit d'analyser et d'extraire les traits linguistiques le caractérisant. La sortie de l'analyse passe par un arbre de classification qui affecte automatiquement une catégorie au message traité constituant en quelque sorte un pré-filtrage. Après l'étape de classification, le message passe par un module de filtrage qui permet d'évaluer son degré d'importance selon les spécificités de l'utilisateur et de prendre, en conséquence, des actions de filtrage, tel que supprimer, sauvegarder, signaler,…

Le facteur intelligent du système est sa faculté d'apprendre et d'améliorer l'efficacité du processus de classification. Le système dispose d'un apprentissage assisté appelé ***feed-back*** où l'utilisateur est invité à donner son avis, et par conséquent modifier le comportement du système.

4 Modèle linguistico-neuronal

Le modèle représentant la typologie des messages est construit à partir d'un ensemble de corpus de messages (messages personnels, professionnels,…). Après une analyse manuelle de chaque corpus, une liste de propriétés le caractérisant, est élaborée.
Une approche intéressante pour modéliser cette typologie est le modèle connexionniste qui a pour but d'imiter certaines fonctions du cerveau humain. Il consiste à représenter les informations sous forme d'un réseau et à permettre au système de faire évoluer ce réseau par la fonction d'apprentissage. Les nœuds du réseau représentent les concepts et les arcs représentent les associations entre les concepts (Denis, Gilleron, 2000), (Davalo, Naim, 1993).
Le réseau de neurones, adopté pour notre système, est représenté comme suit (figure 2):

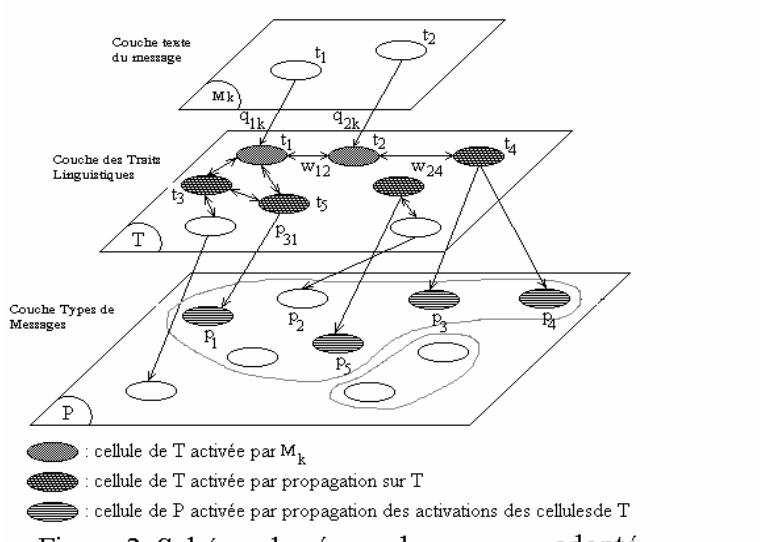

Figure 2: Schéma du réseau de neurones adopté

Ce réseau est composé de trois couches:

- La couche M_k représentant l'entrée du réseau, est créée dynamiquement à chaque récupération d'un nouveau courrier. L'existence du neurone *i* sur cette couche implique la présence du trait linguistique t_i dans le texte du message reçu. Il existe un lien synaptique q_{ik} reliant chaque neurone du courrier à un terme de la couche *T* reflétant le poids de ce terme dans le message M_k.

- La couche *T* représente l'ensemble des propriétés linguistiques existantes dans la base des messages. Ces propriétés sont reliées entre elles par des liens w_{ij} représentant la cooccurrence de deux termes t_i et t_j. Chaque terme de cette couche est relié à la couche *P* par des liens p_{im} valués, tel que *m* appartient à l'ensemble des traits linguistiques représentant le type du message.

- La couche *P* ou sortie du réseau, représente les différents types de messages regroupés en domaines génériques (type personnel, type professionnel, …).

Le réseau reçoit en entrée un seuil de similarité *S* (calculé à partir du corpus initial d'exemples) et un vecteur de termes *V* (représentant le message à analyser). Il tente de trouver au moins un type dont la valeur de similarité avec le vecteur *V* dépasse le seuil *S*. En effet, le vecteur V (couche du message *Mk*) active les termes du message sur la couche *T*. Ces termes activés sur *T* vont propager ensuite leurs activations à leurs proches voisins. Enfin les termes activés directement à partir de la couche *Mk* et ceux activés par propagation, vont envoyer leurs signaux d'activation vers la couche *P*. Les termes de la couche *P* recevant des signaux suffisants, s'activent à leur tour et constituent les types correspondants au message reçu. Le type le plus représentatif du message constitue la réponse du réseau.
La fonction d'activation choisie pour toutes les couches est la fonction sigmoïde définie par (Davalo, Naim, 1993), (Nouali, 2000):

$$f(x) = \frac{e^x - 1}{e^x + 1} \qquad (1)$$

5 Types de classification

5.1 Classification sans propagation (C.S.P.)

Le vecteur du message M_k représentant le signal d'entrée sera propagé à la couche *T*, où chaque neurone de cette couche calcule son entrée selon la formule suivante :

$$E^T(t_i) = q_i \qquad (2)$$

Où t_i représente un trait linguistique et q_i ={1 si t_i est trouvé dans le texte du message et 0 sinon}
Chacun des termes de *T* sera ensuite activé selon l'équation :

$$S^T(t_i) = f(E^T(t_i)) \qquad (3)$$

Où f est la fonction Sigmoïde.

Chaque nœud activé de la couche *T* propage sa sortie à travers les liens p_{ij} vers la couche *P*. Les termes de cette couche vont à leur tour évaluer leurs entrées pondérées selon la formule suivante (Davalo, Naim, 1993), (Nouali, 2000):

$$E^P(P_j) = \sum_{i=1}^{T} S^T(t_i).p_{ij} \qquad (4)$$

Puis s'activeront selon la formule:

$$S^P(P_j) = f(E^P(P_j)) \qquad (5)$$

Les termes activés dont la sortie $S^P(P_j)$ est supérieure au seuil *S* sont considérés comme représentatifs du message et sont triés par ordre décroissant. La valeur la plus grande est renvoyée en sortie. Dans le cas où aucun terme n'a atteint le seuil *S* la valeur 0 est renvoyée.

5.2 Classification avec propagation (C.A.P.)

La propagation se fait grâce à l'extension des signaux d'entrée à travers les liens de cooccurrence entre les traits linguistiques. Dans ce mode de classification, après qu'un nœud évalue son activation, il va la transmettre aux proches voisins, à travers les liens de cooccurrence, ce qui implique la reformulation du vecteur d'entrée V.
Les nœuds termes activés par propagation calculent leurs entrées par (Davalo, Naim, 1993), (Nouali, 2000):

$$E^T(t_i) = \sum_{i=1}^{T} S^T(t_i).w_{ij} \qquad (6)$$

Et seront activés selon la formule (3).

6 Evaluation

Pour réaliser notre expérience, nous avons débuté par l'étude d'un corpus de 150 messages. Il regroupe une variété de types de messages : Personnels (55), professionnels (70), indésirables (15),... Après une analyse manuelle du corpus, nous avons élaboré une liste de propriétés linguistiques caractérisant chaque type de messages. Nous nous sommes limités à quatre types génériques de messages pour construire notre modèle (personnel, appels aux communications, Spam et autres).

L'expérience consiste à présenter au système en deux cas différents, un ensemble de courriers à filtrer en plusieurs sessions. Puis mesurer à chaque fois la précision et le rappel et effectuer un apprentissage assisté pour mesurer son efficacité et son influence sur les deux facteurs.

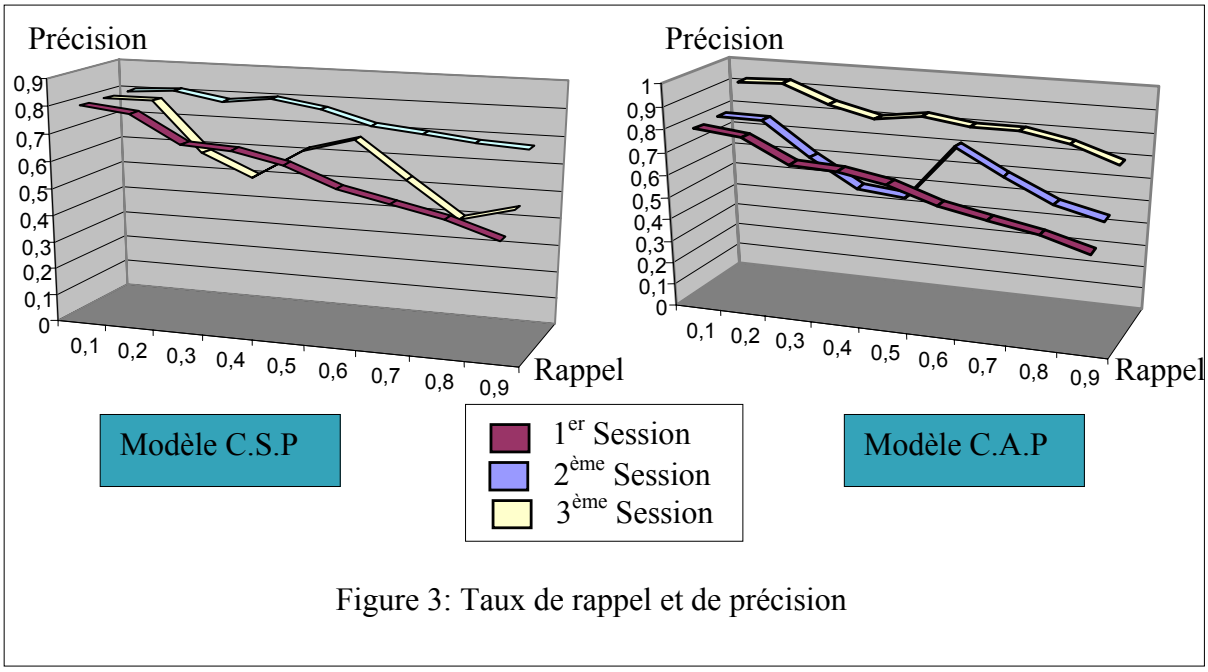

Figure 3: Taux de rappel et de précision

Après plusieurs sessions d'apprentissage assistés, nous constatons que la convergence du modèle de C.A.P. vers un modèle de classification satisfaisant est plus rapide que celle du modèle de C.S.P (figure 3). En effet, dans le modèle C.A.P., la cooccurrence des traits linguistiques est prise en considération, ce qui permet d'augmenter le taux de *rappel* tout en gardant une bonne précision. Par exemple, en ce qui concerne les messages de type personnel, on obtient un taux de filtrage avec 92% de précision pour un rappel de 70%.

7 Conclusion

Le système développé permet d'analyser et d'attribuer une catégorie aux messages en se basant sur les propriétés linguistiques véhiculées par ces messages. Ces propriétés sont modélisées par un réseau de neurone. Chaque nœud du réseau correspond à un trait linguistique. Chaque trait linguistique est pondéré par un poids qui représente son degré d'importance. Le réseau est construit initialement à partir d'un petit corpus de messages et qui est enrichi progressivement au fur et à mesure de l'utilisation du modèle. Et ceci, à l'aide d'un module d'apprentissage qui permet de modifier la pondération des propriétés linguistiques selon les jugements de l'utilisateur. L'apprentissage assisté (feed-back) permet au système d'approcher la pertinence de l'utilisateur et de s'adapter ainsi à ses besoins en lui permettant la modification, l'ajout et la suppression de propriétés.
Les résultats de tests obtenus sur notre petit corpus semblent satisfaisants. Néanmoins, pour tester l'adaptabilité de l'approche, il serait intéressant d'étendre les tests sur un échantillon de messages plus important (par exemple, les forums de discussion qui génèrent des flux considérables de messages et qui sont relativement typés du point de vue thématique).
D'autres part, l'évaluation a été faite par un calcul des mesures Précision/Rappel. Nous comptons utiliser la nouvelle métrique, appelée fonction d'utilité, fonction d'efficacité ou fonction de gain (Hull, 1998), pour positionner notre système par rapport aux autres systèmes de l'étude TREC-7 en comparant les scores.

Références

Copeck T., Barker K., Delisle S., Szpakowicz S. (2000), Automating the Measurement of Linguistic Features to Help Classify Texts as Technical, Conference TALN2000, Lausanne.

Davalo E., Naim P. (1993), *Des Réseaux de Neurones*, Edition Eyrolles.

Denis F., Gilleron R. (2000), *Apprentissage à partir d'exemples*, Notes de cours, Université Charles de Gaulle, Lille3.

Hull D. (1998), "The TREC-7 Filtering Track: Description and Analysis", dans les actes de Text Retrieval Conference (TREC), University of Maryland.

Kilander F., Takkinen J. (1996), *Information ritrieval and information filtering (IRIF), spring 96: Classification & Filtering of Internet messages.*
http://www.ida.liu.se/labs/iislab/courses/irif/irif_filtering_sw.html

Kosseim L., Lapalme G. (2001), *Critères de selection d'une approche pour le suivi automatique du courriel*, RALI, DIRO, Université de Montréal.

Michel C. (1999), *Evaluation de systèmes de recherche d'information, comportant une fonctionnalité de filtrage, par des mesures endogènes*, thèse de doctorat en sciences de l'information et de la communication, Université Lumière LyonII.

Malone T. W., Kenneth R. (1987), *"Intelligent information sharing systems",* Computing practices, Communications at ACM, volume 30, N°5.

Nouali O., Belghoul F., Abdelaziz Y.R. (2000), *Système de filtrage du courrier électronique: E-FILTER*, Mémoire d'ingénieurs, INI, Alger.