SETIT 2005
3rd International Conference: **S**ciences of **E**lectronic, **T**echnologies of **I**nformation and **T**elecommunications
March 27-31, 2005 – TUNISIA

Classifieur Neuronal pour la Catégorisation et le Filtrage de Documents Web

Amel BOUSTIL 1*, Zaidi SAHNOUN 2** et Smaine MAZOUZI 3*

* *Département d'informatique, Université de Skikda, BP 26 Route El-hadaik 21000, Skikda, Algérie.*
boustil1710@yahoo.fr
smazouzi@wissal.dz

** *Département d'informatique, Université Mentouri Constantine, labo LIRE, Constantine 2500, Algérie.*
SahnounZ@yahoo.fr

Résumé: Les moteurs de recherche classiques engendrent souvent des documents non pertinents, du fait que les occurrences de ces mots clés ne sont pas quantifiées. Dans cet article, nous proposons un système de catégorisation et de filtrage de documents, basé sur un réseau de neurones multicouches. Le système, installé sur un client navigateur Web, est configuré pour filtrer les documents issus d'une requête, adressée à un moteur de recherche. Le filtrage est effectué par un réseau de neurones, dont l'apprentissage était fait en Off line. L'utilisateur du classifieur doit définir un vecteur de termes et avoir à sa disposition un ensemble de documents du domaine, où chacun d'eux est affecté à un sujet choisi par l'utilisateur.
Mots clés: Classification de documents, Filtrage, Moteur de recherche, Réseaux de neurones.
.

1 Introduction

La pertinence de l'information accédée par un utilisateur devient, de plus en plus, importante, du fait du nombre colossal de documents électroniques publiés par les différents serveurs d'information à travers le monde. Le développement d'outils permettant le filtrage et la restriction de documents, selon le contexte ou selon un sujet bien spécifié, prend de l'importance au sein de la communauté des spécialistes en classification et filtrage de données [6]. Dans ce contexte, une compétition annuelle est organisée à la marge de la conférence « Text Retreival Conference », à laquelle participent les pionniers des éditeurs de logiciels spécialisés en documentation électronique. La catégorisation des textes, dont les résultats seront, d'une manière ou d'une autre, utilisés pour guider une recherche ultérieure, consiste à partitionner un ensemble de documents en classes, correspondant, chacune de ces dernières, à un sujet donné. Les sujets, plus ou moins distincts, sont caractérisés par un ensemble de descripteurs permettant leur catégorisation. Les méthodes de catégorisation se scindent en deux catégories distinctes : 1) méthodes classiques qui se basent sur les systèmes à base de connaissances utilisant les règles de production ; 2) les méthodes stochastiques basées sur les techniques d'analyse de données. Les classifieurs neuronaux, font une sous catégorie de ces dernières [7].

Dans cet article, nous exposons l'essentiel de notre méthode, qui consiste à l'utilisation d'un réseau de neurones multicouches. Basée sur le vecteur de fréquences relatives de termes associés à un sujet donné, un classifieur neuronal est entraîné en utilisant un ensemble de documents-types dont le sujet est bien défini à l'avance. Une fois l'apprentissage achevé, le classifieur pourra être utilisé pour classifier ou filtrer un ensemble de documents issu d'une recherche classique (basée sur les mots clés). En effet, tout document issu de cette dernière recherche est traité puis soumis au classifieur pour déterminer à quel sujet appartient-il. En cas de filtrage, le document mal classé par le classifieur, sera simplement écarté en considérant que ce document, bien qu'il est issu d'une recherche basée sur les mots clés, n'est pas pertinent vis-à-vis la spécification du sujet, définie et introduite par l'utilisateur à l'étape d'apprentissage du classifieur.

L'avantage de notre méthode est qu'elle permette la réalisation de classifieurs qui s'installent du coté client « Client-side ». Un même serveur, qui héberge un moteur de recherche classique, pourra servir un ensemble de clients où chacun de ces derniers dispose de son propre classifieur. Ceci permet une catégorisation d'un même fond documentaire, selon les points de vue de chaque utilisateur.

2 Prétraitement et codage de textes

Tous document soumis au système, que se soit en phase d'apprentissage, ou en phase de généralisation (classification et filtrage), est traité dans le but d'en extraire les fréquences des occurrences des termes définissant le domaine d'intérêt. En effet, soit un domaine d'intérêt incluant m sujets (S_i , i=1..m},, et soit un texte T , dont l'objectif de le faire associé à un des m sujets définissant le domaine. Un ensemble de N termes (mots) caractéristiques est utilisé pour caractériser le domaine d'intérêt. Cet ensemble de terme est choisi, généralement comme étant l'ensemble des mots clés définissant ce domaine. Le document T est traité afin d'en calculer le vecteur des fréquences d'occurrences des termes caractéristiques dans le texte, soit V^t. Ce vecteur d'occurrences est transformé en vecteur de fréquences relatives F^t..

$$F_i^t = \frac{V_i^t}{\sum_{i=1}^{N} V_i^t} \qquad (1)$$

Ainsi, tout texte T est représenté dans un espace de dimension N par un hyper point de vecteur coordonnées Ft.. Notons qu'il est tout à fait évident qu'un nuage d'hyper points, correspondant à un ensemble de textes, est loin d'être linéairement séparable [3]. Ceci nous a conduit à l'utilisation du perceptron multicouches qui permet la modélisation de fonctions non linéairement discriminantes.

3 Eléments de classification

Notons d'abord, qu'un mot contenu dans un texte et qui correspond à un des termes définissant le vocabulaire caractérisant le domaine, peut ne pas être, forcément, identique à ce dernier. Ceci impose que les termes du vocabulaire caractéristique, doivent être les plus simples et qu'une recherche basée sur les formes grammaticales, des mots doit être prise en compte. Pour notre système, nous avons développé une fonction modélisant la distance entre deux mots. Deux mots sont considérés correspondant l'un à l'autre si la distance qui les sépare est inférieure à un certain seuil.

Les vecteurs de fréquences relatives correspondant aux document-types sont utilisés pour entraîner le réseau de neurones. A ce stade, tout document-type doit être, préalablement bien classé. Cette dernière tâche est réalisée, en off line, par un client de recherche, ou au niveau d'un système proxy, s'il s'agit de l'utilisation du classifieur par un groupe de clients traitant du même domaine d'intérêt. Au moment de la recherche, le vecteur de fréquences relatives, associé à un texte issu d'une recherche classique, est utilisé comme entrée du classifieur pour en déterminer son sujet (sa classe). Notons ici, que les classes ne sont pas disjointes, et dans ce cas, un texte est caractérisé par sa distance au centre de la classe qu'elle lui est la plus représentative (Fig. 1).

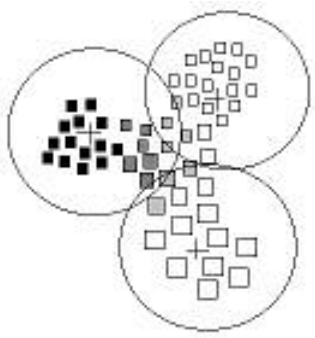

Figure 1 : Interférence de sujets

Le choix des N mots et des et des m sujets caractérisant le domaine d'intérêt est, dans notre cas, fait par un expert du domaine. En effet, certains travaux, publiés dans la littérature [5], font recours à une extraction automatique de ces mots à partir d'un échantillon de corpus, en se basant sur la sélection des mots dont les fréquences d'occurrence sont situées dans une plage donnée de valeurs, évitant ainsi les termes fréquents de la langue et les termes rares.

Cette dernière méthode est loin d'être une méthode de caractérisation sémantique du domaine d'intérêt, du fait qu'elle se base exclusivement sur une approche statistique (fréquences moyennes d'occurrence). En fait, pour notre méthode, qui consiste au partitionnement d'un ensemble de k documents formant le jeux d'apprentissage, en m sujets, en se basant sur le même vocabulaire, composé de m mots caractéristiques est motivée par deux raisons différentes :

1. Permet à un utilisateur de bien spécifier, selon ses convictions, les termes définissant les différents sujets partageant le même vocabulaire et appartenant à un même domaine.
2. Le choix du vocabulaire est indépendant de l'ensemble d'apprentissage. Ce vocabulaire provient exclusivement d'une caractérisation sémantique du domaine, faite par un expert du même domaine.

Cette dernière raison confère a tout outil, ainsi conçu, un aspect sémantique, dont nous jugeons primordial pour tout système intelligent de catégorisation ou de filtrage d'informations.

4 Le classifieur multicouches

Le perceptron multicouches utilisé modélise m fonctions non linéaires faisant chacune la classification des documents par rapport à un des m sujets. On démontre, que toute fonction non linéaire poura être modélisée par un perceptron multicouches [2] et est parcimonieuse [4]. Le réseau est composé de N entrées (Fig. 2), correspondant chacune à une fréquence relative Fit issue d'un document T, d'une couche de l neurones cachés et d'une couche de m neurones représentant les sorties S1,S2, … Sm du classifieur. La sortie Si est associé au sujet i du domaine traité. Le Codage des sorties qui se base sur le choix de 1-parmis-N [1] ne conviendra pas à notre système. En effet, les sujets définissant un même domaine sont généralement en interférence mutuelle. Un document traitant d'un sujet donné, soit Si, est caractérisé par sa distance minimale au centre de la classe représentant le sujet. Toute distance aux centres des autres classes, n'est pas forcément infinie, et dans certaines situation un document peut se situer aux frontières de deux (ou plus de) classes.

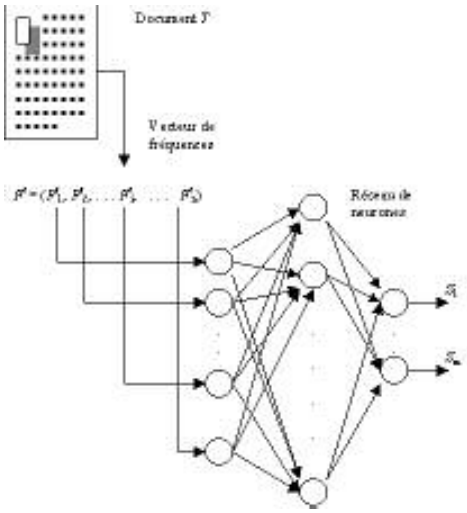

Figure 2 : Principe du classifieur

L'apprentissage du réseau de neurones est basé sur la classification manuelle (par un expert du domaine) d'un ensemble, d'apprentissage A, de taille k assez consistante, de documents. Ces derniers doivent être des document-types des sujets correspondants. Ceci n'empêche pas qu'un document se retrouve représentant de deux (ou plus de) sujets. L'expert procédera, pour chacun de ses documents, à la définition d'un vecteur d'appartenances relatives aux différents sujets. Soit un document da , de l'ensemble d'apprentissage A, son vecteur d'appartenances relatives $P_a = (P_{a1}, P_{a2}, \ldots P_{am})$, est composé de m éléments ou chacun, (soit P_{ai}), représente le degré d'appartenance du document da à un des sujets du domaine (soit S_i).

Les valeurs du vecteur P_a sont normalisées de telle sorte que :

$$\sum_{i=1}^{m} P_i^a = 1 \quad , \text{équivalent à } 100\% \tag{2}$$

Ainsi, tout élément du vecteur Pa, représente le pourcentage d'appartenance du document à un des sujets du domaine. La composante Paj, de valeur maximale, correspond dans ce cas, au sujet principal traité par le document.

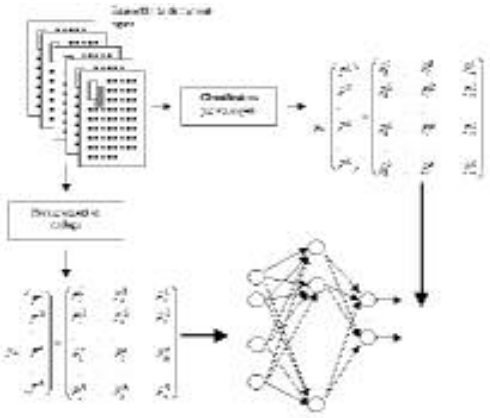

Figure 3 : Apprentissage du classifieur

La rétro-propagation du gradient était adoptée comme méthode pour l'apprentissage du perceptron multicouches. L'ensemble des exemples A est représenté par la matrice F, où chaque ligne représente le vecteur de fréquences relatives des mots caractéristiques (pour cause de longueur) mettre 1/2 ligne de blanc entre les lignes de l'expression.

$$F = \begin{pmatrix} F^1 \\ F^2 \\ . \\ F^t \\ . \\ F^k \end{pmatrix} = \begin{pmatrix} F_1^1 & . & F_i^1 & . & F_n^1 \\ F_1^2 & . & F_i^2 & . & F_n^2 \\ . & . & . & . & . \\ F_1^t & . & F_i^t & . & F_n^t \\ . & . & . & . & . \\ F_1^k & . & F_i^k & . & F_n^k \end{pmatrix} \tag{3}$$

Les vecteurs de sorties qui représentent les catégorisation (pourcentages) des documents par rapport aux m sujets du domaine d'intérêt, sont représentés par une matrice P où chaque ligne Pa contient les m pourcentages d'appartenance du document da aux m sujets.

$$P=\begin{pmatrix} P^1 \\ P^2 \\ . \\ P^a \\ . \\ P^k \end{pmatrix} = \begin{pmatrix} P_1^1 & . & P_i^1 & . & P_m^1 \\ P_1^2 & . & P_i^2 & . & P_m^2 \\ . & . & . & . & . \\ P_1^a & . & P_i^a & . & P_m^a \\ . & . & . & . & . \\ P_1^k & . & P_i^k & . & P_m^k \end{pmatrix} \quad (4)$$

Afin d'agir sur l'efficacité du perceptron à reproduire, lors de la généralisation, les formes utilisées par l'apprentissage, ainsi qu'un ensemble de tests, nous avons opté à la variation du nombre de neurones dans la couche cachée. Lors des tests, effectué sur un ensemble important de document-types concernant l'informatique, extrait d'un CD de formation , et en variant la taille du vocabulaire des mots caractéristiques et le nombre de sujets relevant du domaine, nous avons constaté que le nombre de neurones de la couche cachée dépend, beaucoup plus, de la taille de l'ensemble d'apprentissage (k), comparé à la taille du vocabulaire caractéristique {F} et au nombre de sujets {S}. Dans l'objectif de quantifier la performance du classifieur à la phase de généralisation, nous avons utilisé, pour chaque ensemble d'apprentissage, un ensemble de test. Ce dernier est composé de k/4 documents bien classés (par l'expert), mais n'ayant pas été utilisés pour l'apprentissage. Le réseau de neurones, ainsi entraîné et testé, est retenu comme classifieur s'il arrive à classer correctement 80% des documents de l'ensemble de test. Afin de pouvoir réaliser cet étude, nous étions obligé de développer un outil pour la génération automatique de perceptrons multicouches. Cet outil permet, inter activement via une interface utilisateur, ou via un script, la génération automatique du perceptron, en lui spécifiant les différents paramètres nécessaires. Cet outil est utilisé principalement pour déterminer le nombre de neurones dans la couche cachée, puis évaluer la performance du classifieur obtenu.

5 Filtrage de documents WEB

Le classfieur est utilisé pour le filtrage de documents Web. L'utilisateur d'un navigateur Web formule une requête classique, basée sur un ensemble de mots clés. Le serveur, hébergeant un moteur de recherche conventionnel (public sur le Web), répond au navigateur par l'envoi des URLs de tous les documents satisfaisant la requête. En effet, ce n'est qu'une partie de ces documents qui est pertinente pour l'utilisateur et qui traite effectivement de son sujet d'intérêt. Tout document extrait, suite à cette dernière recherche, est soumis au classifieur avec le sujet d'intérêt introduit, implicitement ou explicitement, avec la requête de l'utilisateur. Le code de sujet calculé par le classifieur est ensuite comparé au code de sujet introduit, et ainsi le document traité, est soit sélectionné et confirmé pour l'affichage, soit rejeté. Notons ici, que chaque utilisateur d'un navigateur Web peut disposer de son propre classifieur. Ainsi, les résultats d'une même requête, exprimée par deux utilisateurs ayant deux classifieurs différents, peuvent ne pas être les mêmes. En effet, les deux requêtes, engendrent le même ensemble de documents au niveau du moteur de recherche, par contre la classification restreindra différemment chacun des deux ensembles selon les deux classifieurs (Fig. 4).

Pour un groupe d'utilisateurs ayant le même domaine d'intérêt, il est possible d'installer le classifieur sur un serveur proxy en procédant à son apprentissage par un expert du domaine. Cependant, un utilisateur s'intéressant à un seul sujet peut, à lui seul, entraîné son classifieur, en lui soumettrant un ensemble de documents traitant du sujet en question. Cette dernière façon relève de la discrimination automatique d'un ensemble de documents.

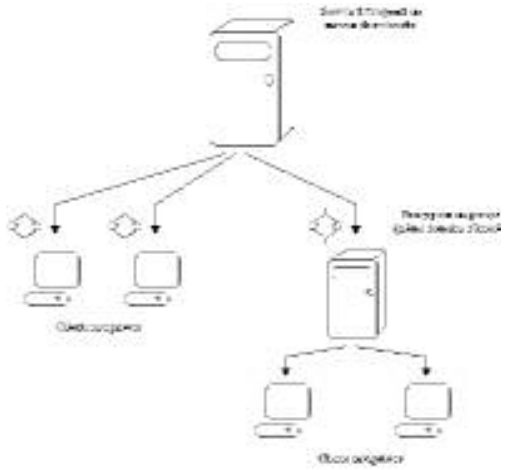

Figure 4 : Installation du classifieur

Conclusion

Dans le présent travail, nous avons conçu un système permettant la classification et le filtrage de documents en utilisant un perceptron multicouches. Le codage des entrées était basé sur les fréquences relatives d'occurrences de mots caractéristiques dans le texte. Cet ensemble de mots, dit vocabulaire, est spécifié par l'utilisateur du classifieur, qui sous entend expert du domaine. L'avantage de ce système et qu'il permet une spécialisation, coté client, d'un moteur de

recherche conventionnel qui sert tous ses clients en répondant, d'une manière monotone, à des requêtes classiques basées sur un ensemble mots clés. Dans ce cas, la spécialisation consiste à la sélection d'un sous ensemble de documents qui sont bien catégorisés par le classifieur.

Dans les perspectives de ce travail, nous envisageons le traitement de l'information non factuelle, telle que l'image et le son contenue dans les documents multimédia, en investiguant dans le codage de ce genre d'information en dans la mise en œuvre d'outils permettant l'expression de requêtes multimédia. Sur un second axe, il est possible de tester la classification automatique par les réseaux de neurones avancés tels que les cartes auto-organisatrices de Kohonen.

Références

[1] : G. Dreyfus et al. , *Réseaux de neurones : méthodologie et applications*, Edition Eyrolles 2002.

[2] : K. Hornik, *Approximation capabilities of multilayer feedforward network*s, Neural Networks, 3, pp. 551-560.

[3] : M. Milgram , *Reconnaissance des formes : méthodes numériques et connexionnistes,* Edition Armand Colin, 1993.

[4] : J.P. Nadal, *réseaux de neurones : de la physique à la psychologie* , Edition Armand Colin, 1995.

[5] : A. Singhal , *Pivoted length normalization*, Proceding of the 19th annual international conference of research and development in information retrieval (SIGIR'96), pp. 21-29.

[6] : M. Striker, *Training context-sensitive neural networks with few relevant examples*, proceding of TREC-9 proceding, 2001.

[7] : F. Wolenski and al. , *Using learning based filters to detect rule-based filtring absolescence*, conference RIAO, Paris, 2000.