DETECTION DE CONVERGENCE EN VUE DE L'OPTIMISATION D'UN SYSTEME DE FILTRAGE ADAPTATIF

TMAR Mohamed
tmar@irit.fr

TEBRI Hamid
tebri@irit.fr

BOUGHANEM Mohand
boughane@irit.fr

Adresse professionnelle
IRIT-SIG - Université Paul Sabatier de Toulouse
118, route de Narbonne, F-31062 Toulouse Cedex 4
Tel: (+33) (0)5.61.55.74.16

Résumé : Un système de filtrage adaptatif permet d'extraire, à partir d'une source dynamique de documents, les seuls documents pouvant intéresser un utilisateur ayant des centres d'intérêts relativement stables. Nous avons développé un système de filtrage adaptatif basé sur le principe de renforcement pour apprendre le profil, et la distribution de probabilités des scores des documents pertinents et non pertinents pour adapter la fonction de seuillage. Cet article décrit ce système, et les expérimentations effectuées pour mesurer l'efficacité de notre approche.

Mots-clés : filtrage adaptatif, Utilité, seuillage, convergence, distribution de probabilités.

Abstract : An adaptive filtering system allows to extract, from a stream of documents, the only documents being able to interest a user having relatively stable centers of interests. We developed a model of adaptive filtering system based on the principle of reinforcement to learn the profile. The threshold calibration is performed using the score probability distribution in two samples of documents: a sample of relevant documents and a sample of non-relevant documents. This paper describes our system, and the experimentations carried out to measure the effectiveness of our approach.

Keywords : adaptive filtering, Utility, threshold, convergence, probability distribution.

Détection de convergence en vue de l'optimisation d'un système de filtrage adaptatif

INTRODUCTION

L'avènement Web, contexte dans lequel la recherche d'information est une préoccupation centrale, a réactualisé la problématique de la recherche d'information, particulièrement dans la manière d'accéder aux informations. En effet, si avec un système de recherche d'information on accède volontairement à des informations via des requêtes ou par navigation, on assiste aujourd'hui de plus en plus à la prolifération de services qui ramènent des informations à l'utilisateur. Le processus qui permet de sélectionner l'information désirée dans ces flots d'informations s'appelle le filtrage d'information.

Un système de filtrage d'information permet à partir d'une source dynamique d'information (Internet, E-mail, News,…) de sélectionner et de présenter les seuls documents intéressant un utilisateur ayant un centre d'intérêt relativement stable appelé *profil.*

Le filtrage d'information est un processus dual à la recherche d'information [Belkin & al., 1992]. Ceci traduit qu'un processus de recherche d'information peut simuler un processus de filtrage d'information. Cependant, la plupart des systèmes de filtrage d'information sont basés sur des modèles de recherche d'information. Ainsi, les documents et les profils sont représentés par des listes de mots pondérés. L'appariement document-profil consiste à mesurer une similarité. La décision quant à l'acceptation ou le rejet d'un document est assurée par une fonction de décision souvent de type seuil. Si le score est supérieur au seuil le document est accepté sinon il est rejeté.

Or, en l'absence de base de référence, la détermination de ce seuil et les pondérations adéquates associées aux profils et aux documents sont les problèmes majeurs rencontrés dans ce domaine. En effet, dans un système de recherche d'information, les techniques de pondération et de reformulation automatique de requêtes, basées sur la collection de documents, se sont avérées efficaces, or, dans un système de filtrage d'information, à l'initialisation du processus de filtrage, on ne dispose d'aucune connaissance sur les documents à filtrer pour pouvoir construire une fonction de décision, ni pour bien pondérer les profils et les documents entrants. De plus, l'adaptation des profils aux différents flots pose un problème fondamental lié à l'incomplétude permanente des informations permettant de décrire exhaustivement les profils. Les solutions proposées aujourd'hui sont les suivantes : une solution synchrone ou adaptative, les différents facteurs du système de filtrage sont déduits à partir des documents filtrés cumulés dans le temps, et les solution asynchrone ou différée, les facteurs sont déduits à partir de collections de documents existantes.

La solution que nous proposons est synchrone, les profils et la fonction de décision sont appris d'une façon incrémentale pendant le filtrage. La plupart des systèmes de filtrage d'information actuels respectent très peu la notion d'incrémentalité, car :

- ils utilisent souvent des bases d'entraînement [Zhai & al., 1998][Robertson & al., 1999]. Ces systèmes se basent sur les statistiques des termes et des documents dans ces bases pour estimer les valeurs de plusieurs paramètres de filtrage, or ces statistiques sont très variables au cours du filtrage et inconnues à l'initialisation du processus. De plus, les documents issus des bases d'entraînement peuvent avoir des caractéristiques des documents provenant de la source, les statistiques utilisées dans l'entraînement peuvent alors être peu crédibles,
- ils effectuent l'apprentissage du profil et le seuillage d'une manière quasi-différée [Kwok & al., 1999][Hoashi & al., 1999]. En effet, plusieurs effectuent l'apprentissage du profil et/ou le seuillage à chaque réception de n (10, 100, 1000) documents. Avec cette manière périodique de fonctionnement, le système perd beaucoup de qualités incrémentales et tend vers les systèmes de filtrage différé [Malone & al., 1987],
- plusieurs informations sont parfois supposées connues avant le démarrage du filtrage (proportion de documents pertinents, quelques documents pertinents d'entraînement, …) [Zhai & al., 1998], ceci n'est pas vrai dans le cas du filtrage, car le système démarre avec une information sauf le profil initial de l'utilisateur.

Le modèle que nous proposons pour notre part est purement adaptatif et incrémental. Aucune information autre que le profil initial n'est connue au démarrage du processus de filtrage. Les statistiques des documents et des

profils sont actualisées au fur et à mesure que le système reçoit des documents. L'adaptation des profils et le seuillage se font d'une manière adaptative et incrémentale à chaque réception d'un document pertinent.

Nous proposons une méthode d'adaptation du profil basée sur le principe de renforcement. Chaque fois qu'un document est sélectionné et jugé pertinent pour un profil donné, le système doit adapter le profil de sorte à modifier sa représentation. De plus, nous proposons une méthode d'adaptation du seuil basée sur la distribution des scores d'un échantillon de documents pertinents et non pertinents sélectionnés. Dans la première section, nous présentons notre modèle de filtrage. La seconde section décrit notre approche d'adaptation du profil et du seuil. Enfin, nous présentons dans la dernière section les expérimentations et les résultats obtenus.

1 - MODELE DE BASE

Le modèle de filtrage que nous proposons est basé sur une approche vectorielle. Les documents et les profils sont représentés sous forme d'une liste de termes pondérés. Le processus de filtrage consiste à comparer le score résultant de la similarité document-profil à un seuil, si le score est supérieur au seuil le document est sélectionné sinon il est rejeté. Ce seuil, les profils et les caractéristiques liées à la pondération des termes des documents évoluent à chaque arrivée d'un document.

1.1. Représentation des profils

Un profil est un ensemble de termes sans les mots vides. Il est représenté sous une forme vectorielle :

$$p^{(0)} = ((tp_1, \omega_1^{(0)})...(tp_n, \omega_n^{(0)})) \qquad (1)$$

Avec $\omega_i^{(0)}$ le poids du $ième$ terme tp_i dans le profil à l'instant $t = 0$. t est incrémenté à chaque arrivée d'un document et représente l'instant où le système reçoit un document. Initialement, le poids du terme dans le profil est calculé comme suit :

$$\omega_i^{(0)} = \frac{tfp_i}{\max_j(tfp_j)} \qquad (2)$$

Où tpf_i est la fréquence du terme tp_i (noté aussi t_i) dans le profil. La formule est à priori simple, car au début du processus de filtrage on ne dispose d'aucune information autre que le profil initial. Cependant, ce poids sera ajusté par apprentissage au fur et à mesure les documents arrivent.

1.2. Représentation des documents

A chaque arrivée d'un document, celui-ci est indexé par un module d'indexation de **Mercure** [Boughanem, 2000]. Le résultat de cette opération est une liste de termes. Chaque terme du document est pondéré en utilisant la formule suivante :

$$d_i^{(t)} = \frac{tf_i^{(t)}}{h_3 + h_4 . \frac{dl^{(t)}}{\Delta l^t} + tf_i^{(t)}} \log(\frac{N^{(t)}}{n_i^{(t)}} + 1) \qquad (3)$$

$d^{(t)}$: le document reçu à l'instant t ,
$tf_i^{(t)}$: fréquence d'apparition du $ième$ terme dans le document $d^{(t)}$,
h_3, h_4 : paramètres constant,
$dl^{(t)}$: longueur ou nombre de termes du document $d^{(t)}$,
$\Delta l^{(t)}$: longueur moyenne d'un document,
$N^{(t)}$: nombre de documents déjà examinés,
$n_i^{(t)}$: nombre de documents contenant le terme tp_i parmi les documents déjà examinés.
Les paramètres $\Delta l^{(t)}, N^{(t)}$ et $n_i^{(t)}$ sont mis à jour à chaque arrivée d'un document.

1.3. Processus de filtrage de documents

Le processus de filtrage consiste à calculer un score, noté $rsv(d^{(t)}, p^{(t)})$ entre le document $d^{(t)}$ et le profil $p^{(t)}$. Ce score est défini par le produit scalaire entre le document et le profil :

$$rsv(d^{(t)}, p^{(t)}) = \sum_{\substack{t_i \in d^{(t)} \\ tp_j \in p^{(t)} \\ t_i = tp_j}} d_i^{(t)} * \omega_j^{(t)} \qquad (4)$$

Le score calculé est ensuite comparé à un seuil de filtrage, pour décider si le document est accepté ou non : si $rsv(d^{(t)}, p^{(t)}) \geq seuil^{(t)}$ alors le document $d^{(t)}$ est sélectionné, sinon il est rejeté. Le seuil est appris à chaque arrivée d'un document pertinent. Le processus d'adaptation du seuil sera détaillé ci-dessous.

2 - ADAPTATION DU PROFIL ET DU SEUIL

2.1. Apprentissage du profil

Le processus d'apprentissage du profil est adaptatif et incrémental. Il est effectué à chaque fois un document est jugé comme étant pertinent par l'utilisateur. Il permet de modifier la représentation du profil de l'utilisateur, en ajustant les poids des termes, en ajoutant ou en éliminant les termes du profil. Nous décrivons dans les sections suivantes, les étapes nécessaires pour effectuer cet apprentissage.

2.1.1. Conception du processus d'apprentissage

L'idée de base de notre processus d'apprentissage est basée sur le principe de renforcement [Sutton & al., 1998]. Quand un document est jugé pertinent, il faut pouvoir trouver une nouvelle représentation du profil qui permet de retrouver le document avec un score fort. Autrement dit, on sera amené à améliorer le profil tel que $rsv(d^{(t)}, p^{(t)}) = \beta$ ou β est le score désiré. Le problème à résoudre revient alors à chercher les $\omega_j^{(t)}$ qui satisfont l'équation :

$$\sum_{\substack{t_i \in d^{(t)} \\ tp_j \in p^{(t)} \\ t_i = tp_j}} d_i^{(t)} * \omega_j^{(t)} = \beta \qquad (5)$$

Toutefois cette équation admet une infinité de solutions, alors nous proposons d'ajouter une contrainte pour avoir une solution unique. L'intérêt dans tous ça, est de faire tendre le poids de chaque terme vers son poids idéal. Le poids idéal correspond au poids du terme qui permet de discriminer l'ensemble des documents pertinents et celui des non pertinents. Ainsi, si le poids idéal d'un terme t_i est donné par une fonction f donnée par la formule (9) , et si le poids dans le profil est $\omega_i^{(t)}$, alors $\frac{\omega_i^{(t)}}{f(d_i^{(t)}, r_i^{(t)}, s_i^{(t)})}$ est une constante, où $r_i^{(t)}$ (resp. $s_i^{(t)}$) représente le nombre de documents pertinents (resp. non pertinents) contenant le terme t_i. Le système à résoudre devient alors :

$$\begin{cases} \displaystyle\sum_{\substack{t_i \in d^{(t)} \\ tp_j \in p^{(t)} \\ t_i = tp_j}} d_i^{(t)} * \omega_j^{(t)} = \beta \\ \forall (t_i, t_j) \in d^{(t)^2}, \ \dfrac{\omega_i^{(t)}}{f(d_i, r_i^{(t)}, s_i^{(t)})} = \dfrac{\omega_j^{(t)}}{f(d_j, r_j^{(t)}, s_j^{(t)})} \end{cases} \qquad (6)$$

La solution du système (6) est l'ensemble des poids du profil qui permet de retrouver le document $d^{(t)}$. Elle correspond à des poids provisoires qui vont intervenir dans le calcul du poids global du profil.

Soit n le nombre de termes distincts dans le document à l'instant t, et $f_i^{(t)} = f(d_i^{(t)}, r_i^{(t)}, s_i^{(t)})$. Le système (6) peut être réécrit comme suit : $\forall i \in \{1 \ldots n\}$

$$\begin{cases} \dfrac{\omega_1^{(t)}}{f_1^t} = \dfrac{\omega_i^{(t)}}{f_i^t} \quad \Leftrightarrow \quad \omega_1^{(t)} d_{j1}^{(t)} = f_1^{(t)} d_{j1}^{(t)} \dfrac{\omega_i^{(t)}}{f_i^{(t)}} \\ \dfrac{\omega_2^{(t)}}{f_2^t} = \dfrac{\omega_i^{(t)}}{f_i^t} \quad \Leftrightarrow \quad \omega_2^{(t)} d_{j2}^{(t)} = f_2^{(t)} d_{j2}^{(t)} \dfrac{\omega_i^{(t)}}{f_i^{(t)}} \\ \vdots \\ \dfrac{\omega_n^{(t)}}{f_n^{(t)}} = \dfrac{\omega_i^{(t)}}{f_i^{(t)}} \quad \Leftrightarrow \quad \omega_n^{(t)} d_{jn}^{(t)} = f_n^{(t)} d_{jn}^{(t)} \dfrac{\omega_i^{(t)}}{f_i^{(t)}} \end{cases} \qquad (7)$$

Où j_k correspond à l'index dans le document du terme indexé par k dans le profil ($t_k = tp_{jk}$). En additionnant le premier opérande de chaque équation et après quelques transformations, on obtient pour chaque terme son poids provisoire $p\omega_i^{(t)}$ qui est donné par :

$$\forall i,\ p\omega_i^{(t)} = \frac{\beta . f(d_i^{(t)}, r_i^{(t)}, s_i^{(t)})}{\sum_j f(d_i^{(t)}, r_i^{(t)}, s_i^{(t)}).d_j^{(t)}} \qquad (8)$$

Cette valeur permet d'affecter à chaque terme un poids idéal en tenant compte de son importance dans le document $d^{(t)}$, l'ensemble des documents pertinents et non pertinents ($r_i^{(t)}$ et $s_i^{(t)}$) et le score désiré du document (β).

Le choix de la fonction f dépend de plusieurs paramètres, la fréquence d'apparition du terme dans le document, le nombre de document pertinents et non pertinents contenant ce terme, le nombre total de documents pertinents sélectionnés, etc. Nous avons expérimenté certaines fonctions et avons opté sur une fonction dérivée de la formule de Robertson-Sparck Jones [Robertson & al., 1976] :

$$f(d_i^{(t)}, r_i^{(t)}, s_i^{(t)}) = d_i^{(t)}.\log(1 + \frac{r_i^{(t)}(S^{(t)} - s_i^{(t)})}{(s_i^{(t)} + 1)(R^{(t)} - r_i^{(t)} + 1)}) \qquad (9)$$

Où $R^{(t)}$ (resp. $S^{(t)}$) représente le nombre de documents pertinents (resp. non pertinents) sélectionnés par le système à l'instant t .

2.1.2. Modification du profil

L'adaptation du profil consiste à améliorer sa représentation à chaque fois un document pertinent est sélectionné par le système. Dans notre cas, l'adaptation consiste à utiliser les poids provisoires $p\omega_i^{(t)}$ pour contribuer à l'apprentissage des termes dans le profil. Nous utilisons la formule de distribution de gradient suivante :

$$\omega_i^{(t)} = \omega_i^{(t)} + \log(1 + p\omega_i^{(t)}) \qquad (10)$$

Pour valider ces différentes formules, des expérimentations ont été effectuées sur une base **Reuters** issue de TREC[1]-10. Cette technique d'adaptation permet d'améliorer la représentation du profil et de séparer les documents pertinents des documents non pertinents [Boughanem & al., 2002].

2.2. Adaptation du seuil

Dans un contexte incrémental, l'adaptation du profil entraîne automatiquement une variation à la hausse des documents des scores. Par conséquent, l'adaptation du seuil devient nécessaire pour arriver à sélectionner le maximum de documents pertinents et à rejeter le maximum de documents non pertinents.

Plusieurs techniques ont été envisagées et expérimentées pour le seuillage [Tmar, 2002]. L'approche que nous proposons dans cet article est basée sur la distribution des scores des documents. Nous supposons que la distribution des scores des documents suit une certaine loi de probabilité. En se basant sur cette loi, nous pouvons décider si un document est pertinent ou non, selon sa probabilité de pertinence.

2.2.1. Principe de l'approche de l'adaptation du seuil

L'approche que nous proposons consiste à estimer, pour un échantillon de documents, sa distribution de probabilités discrète. Grâce à une technique de régression linéaire, nous transformons cette distribution de probabilités discrète en une densité de probabilités continue. Ceci va nous permettre de choisir une valeur du seuil dans un intervalle continu de scores.

---

[1] Text REtrieval Conference : *un programme d'évaluation des systèmes de recherche et de filtrage d'information*

2.2.2. Modélisation des distributions des scores

L'appariement document-profil permet de fournir un score donné pour ce document. La probabilité qu'un document tiré aléatoirement ait un score donné est par définition égale au nombre de documents ayant eu ce score divisé par le nombre de documents total :

$$p(X = score) = \frac{|\{d \mid rsv(d, p) = score\}|}{|\{d\}|} \qquad (11)$$

Comme les valeurs des scores sont très variées (voir figure. 1), elles ont tendance à être équiprobables ($|\{d \mid rsv(d, p) = score\}| = 1 \; ou \; 0$). La distribution des scores tend alors à être uniforme (voir figure. 2). En effet, dans un échantillon, il est très peu probable de trouver deux ou plusieurs documents ayant exactement le même score.

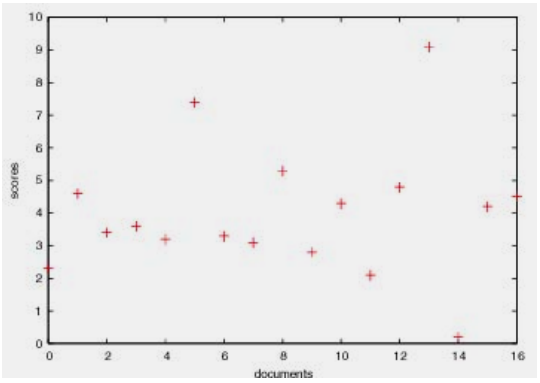

Figure.1 : La distribution des scores est très éparpillée

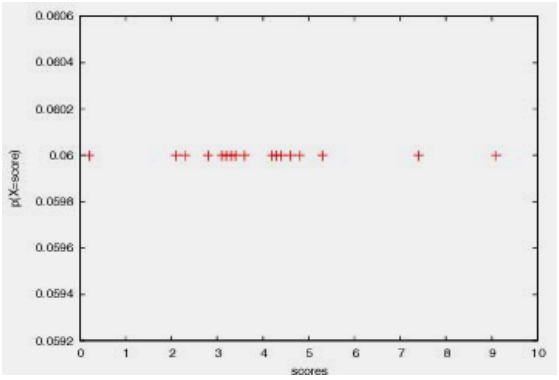

Figure.2 : La distribution uniforme n'est pas significative

Pour donner une distribution de probabilités plus significative, nous proposons qu'au lieu de calculer la probabilité qu'un document ait un score, nous calculons la probabilité que le score d'un document appartienne à un intervalle. Nous choisissons des intervalles réduits pour que les scores des documents appartenant au même intervalle soient réellement presque égaux. Soit n le nombre des ces intervalles, $I_1, I_2, ..., I_n$ de même rayon où :

$$I_i = [score_{i-1}, score_i]$$
$$score_{i-1} = \min_d rsv(d, p) \qquad (12)$$
$$score_i = \max_d rsv(d, p)$$

Le nombre d'intervalles est proportionnel à la taille de l'échantillon, car plus la taille de l'échantillon augmente, plus le domaine de définition des scores des documents s'élargit. Nous choisissons n comme la moitié de la taille de l'échantillon :

$$n = \frac{|\{d\}|}{2} \qquad (13)$$

La probabilité d'appartenance du score d'un document à un intervalle est définie par :

$$p(score_i < X < score_{i+1}) = \frac{|\{d \mid score(d, p) \in]score_i, score_{i+1}[\}|}{|\{d\}|} \qquad (14)$$

La distribution des probabilités par intervalle de scores est plus réaliste. La figure 3 illustre la distribution des probabilités basée sur la formule (14). Elle montre que la distribution des scores des documents admet une allure poisssonnienne.

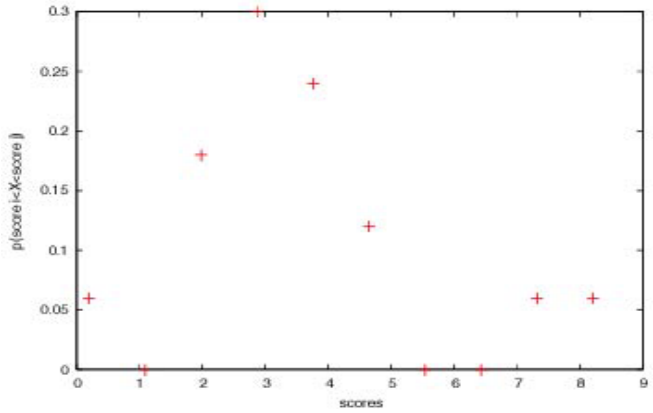

Figure.3 : La distribution est poissonnienne plutôt qu'uniforme

Plusieurs méthodes sont envisageables pour estimer une loi de probabilité suivie par les scores des documents : la régression paramétrique et l'estimation par le maximum de vraisemblance [Saporta, 1996]. L'application de telles méthodes permet d'estimer les paramètres qui permettent de fournir une représentation la plus fiable possible (celle qui passe le plus proche possible par tous les scores).

Dans un contexte expérimental, nous soulignons quelques inconvénients dans l'utilisation de ces méthodes :

- Si on admet la forme de la fonction à priori, elle peut ne pas être valable pour des conditions expérimentales particulières, car les scores des documents sont aléatoires,
- On doit disposer d'un nombre minimum d'échantillons (de documents) pour avoir des estimations non biaisées,
- La distribution des scores dépend de la fonction de pondération des documents.

Pour pallier ces problèmes, nous adoptons une méthode permettant de diviser l'espace probabilisé en plusieurs intervalles, tels que la distribution des scores dans chaque intervalle est linéaire. On applique la linéarisation sur l'ensemble des probabilités des documents pertinents et non pertinents, afin de déterminer le seuil de filtrage.

2.2.3. Linéarisation de la distribution de probabilités

La linéarisation consiste à parcourir le domaine de définition d'une fonction donnée, et de diviser ce domaine en intervalles tels que la courbe représentative de restriction de la fonction sur chaque intervalle peut être assimilée à une courbe linéaire. Nous utilisons cette technique pour linéariser respectivement, la courbe représentative de la densité de probabilités des scores des documents pertinents et non pertinents.

Le processus de détection des intervalles linéaires consiste à chercher le maximum de points adjacents tels que la courbe reliant ces points est linéaire. Nous appliquons la méthode des moindres carrés utilisée pour la régression linéaire [Saporta, 1996], pour déterminer la linéarisation d'un ensemble de points. Le principe est de calculer à chaque fois l'écart quadratique entre les points considérés et la courbe linéaire. Si l'erreur est inférieure à un seuil donné, on ajoute à la courbe le point suivant et on vérifie la linéarité de la nouvelle courbe, sinon on élimine ce point de cet ensemble et on recherche un nouvel ensemble de points à linéariser. La figure.4 illustre un exemple de linéarisation possible. Le processus de linéarisation est le suivant :

1. *c=1 ('c' est indice d'une classe linéaire),*
2. *P=∅,*
3. *seuil_erreur=0.001,*
4. *pour i∈ {0…n-1} (n est le nombre de points (score, probabilité),*
   a. *$P \leftarrow P \cup \{i\}$,*
   b. *Déterminer l'équation de la droite $D_c : y(x) = a + b.x$ par la régression linéaire sur tous les points de coordonnées ($j, p_j = p(score_j < score < score_{j+1})$) $\forall j \in P$,*
   c. *Calculer l'erreur représentée par l'écart quadratique des points ($j, p_j = p(score_j < score < score_{j+1})$) $\forall j \in P$ et la droite D_c :*

$$E = \sum_{j \in P} d^2((j, p_j), D_c)$$

$$d^2((j, p_j), D_c) = \frac{a + b.j - p_j}{\sqrt{a^2 + 1}}$$

d. *Si E > seuil_erreur,*

i. $C_c = (d_c = \min(j \in P), f_c = \max(j \in P), a_c, b_c)$ *avec* a_c *et* b_c *sont les coefficients de l'équation de la droite* $y = a_c + b_c.x$ *par la régression linéaire sur tous les points de coordonnées* $(j, p_j = p(score_j < score < score_{j+1}))$ *avec* $j \in P \setminus \{i\}$,

ii. $P \leftarrow \{i\}$,

iii. $c \leftarrow c+1$,

Ce processus permet seulement de représenter la distribution de probabilités des documents sous forme de plusieurs segments de droite. Cependant, il est nécessaire de transformer cette représentation pour former une distribution de probabilités continue :

1. Premièrement, il faut relier les deux extrémités de chaque deux intervalles adjacents. Cette liaison s'effectue comme suit : pour deux classes linéaires adjacentes C_c et C_{c+1}, relier f_c et d_{c+1} par une droite $y = \alpha_c + \beta_c.x$. Cette doit passer par les points $(f_c, a_c + b_c.f_c)$ et $(d_{c+1}, a_{c+1} + b_{c+1}.d_{c+1})$.

   Soit g la fonction définie par :

$$g : [score_0, score_n] \rightarrow R, \quad x \rightarrow \begin{cases} a_c + b_c.x & si\ \exists c, d_c \le x \le f_c \\ \alpha_c + \beta_c.x & sinon,\ avec \end{cases}$$

2. Deuxièmement, il faut normaliser les coefficients a_c, b_c, α_c et β_c pour que :

$$\int_{score_0}^{score_n} g(x)\,dx = 1$$

Puisque $\int_{score_0}^{score_n} g(x)\,dx$ représente la surface de l'aire formée par la représentation graphique de g et l'axe des abscisses, il suffit de diviser les coefficients a_c, b_c, α_c et β_c par cette valeur. L'aire est calculée comme la somme des aires de chaque surface d'une classe linéaire.

La figure 5 illustre une linéarisation effectuée sur l'ensemble des documents pertinents et non pertinents de la base **Reuters** dans le cas du profil 1. Elle montre que la linéarisation des probabilités de scores tend à avoir une allure exponentielle pour les documents non pertinents et une allure gaussienne pour les documents pertinents.

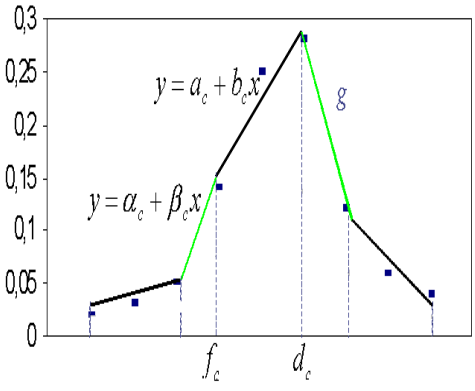

Figure.4 : linéarisation de la distribution de probabilités

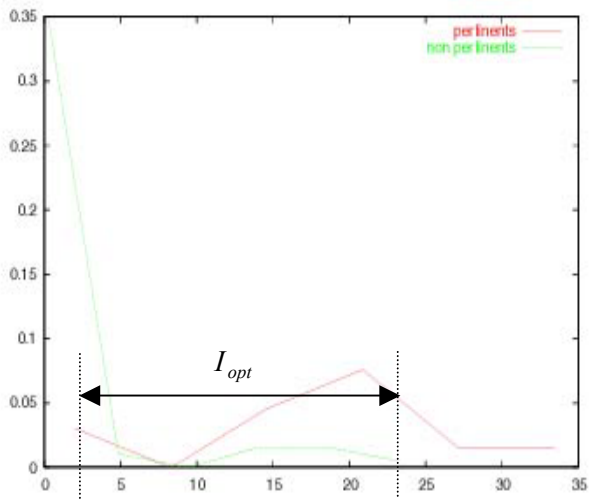

Figure. 5 : Densité de probabilités des scores des documents pertinents et non pertinents

2.2.4. Optimisation de la fonction de seuillage

La méthode que nous proposons pour choisir le meilleur seuil de filtrage est basée sur l'optimisation d'une fonction d'utilité. Une fonction d'utilité permet de mesurer la performance d'un système de filtrage d'information. L'optimisation de la fonction d'utilité consiste à déterminer un seuil de convergence, c'est-à-dire à quel moment il ne devient plus nécessaire de continuer l'adaptation. Ce seuil doit donc permettre au système de sélectionner le maximum de documents pertinents et d'éliminer le maximum de documents non pertinents. Le principe est le suivant :

i. Définir un intervalle de seuils $I_{opt} = [score_{\min p}, score_{\max np}]$ (figure. 5), où $score_{\min p}$ et $score_{\max np}$ représentent respectivement la valeur minimale et maximale des scores des documents pertinents et non pertinents,

ii. Détecter la valeur du seuil dans l'intervalle I_{opt} qui optimise la fonction d'utilité F :.

$$F = a.R_+ - b.S_+ \qquad (15)$$

Alors, notre objectif est de déterminer le seuil qui permet de maximiser théoriquement la fonction F, soit :

$$seuil^* = \arg \max_{seuil \in I_{opt}} F$$

Où :

a, b : deux constantes positives,

R_+ (S_+) : nombre de documents pertinents (non pertinents) sélectionnés,

Les valeurs R_+ et S_+ dépendent évidemment du choix du seuil et sont estimés par :

$$R_+ = p(r|score > seuil) * R$$

$$S_+ = p(s|score > seuil) * S$$

Où R et S représente le nombre total de documents pertinents et documents non pertinents.

Par application de la transformation de la règle de Bayes, nous obtenons :

$$R_+ = \frac{p(score > seuil|r) * p(score > seuil)}{p(r)} * R$$

$$S_+ = \frac{p(score > seuil|s) * p(score > seuil)}{p(s)} * S$$

Avec $p(score > seuil|r)$ (resp. $p(score > seuil|s)$) représente la probabilité qu'un document soit sélectionné sachant qu'il est pertinent (resp. non pertinent) et elle est définie par l'aire de la surface formée par la courbe de g correspondant à la densité de probabilité des scores des documents pertinent à partir du seuil.

3 - EXPERIMENTATION ET RESULTATS

Les expérimentations que nous avons effectuées ont été réalisées sur une collection issue de la compagne **TREC-10**. Dans ce paragraphe, les expérimentations sont concentrées sur la dernière base fournie par **TREC** : la base **Reuters**. Cette base est constituée d'un ensemble de documents de 783484 documents en format XML (eXtensible Markup Language). Les tests sur cette base doivent porter sur 84 profils. Le but de ces expérimentations est de mettre en valeur principalement la technique d'adaptation du seuil précitée.

L'adaptation du seuil exploite directement les scores des documents filtrés, plus les paramètres de la fonction d'utilité *TU10* [Robertson & al., 2000].Ainsi, pour chaque profil, nous appliquons l'algorithme d'adaptation incrémentale du seuil suivant :

1. $echantillonP \leftarrow \Phi$ (un échantillon de documents pertinents)
2. $echantillonN \leftarrow \Phi$ (un échantillon de documents non pertinents)
3. $seuil^{(0)} = 0$
4. *pour chaque document* $d^{(t)}$
   a. *calculer* $rsv(d^{(t)}, p^{(t)})$
   b. *si* $rsv(d^{(t)}, p^{(t)}) > seuil^{(t)}$
      i. *si* $d^{(t)}$ *est jugé pertinent*

A. $R_+ \leftarrow R_+ + 1$
B. $echantillonP \leftarrow \{d^{(t)}\}$
C. Apprendre le profil
D. Apprendre le seuil en utilisant $echantillonP$ et. $echantillonN$

ii. *Sinon*

$S_+ \leftarrow S_+ + 1$

5. *évaluer :* $T10U = 2R_+ - S_+$

Pour mesurer la performance de notre méthode de seuillage, nous avons utilisé les 15 profils de la base **Reuters**, et avons comparé les valeurs d'utilité obtenues, à celles obtenues par les meilleurs systèmes ayant participé à **TREC-10**.

Le tableau 1 montre que sur les 11 profils, on obtient un résultat meilleur que tous les autres participants, et pour la totalité des profils on obtient un résultat qui dépasse la moyenne des résultats des autres participants.

Profil	Utilité obtenue	Utilité maximale	Utilité moyenne
1	0.11	0.10	0.02
2	0.58	0.30	0.13
3	0.03	0.14	0.02
4	0.55	0.24	0.07
5	0.84	0.35	0.06
6	0.63	0.33	0.15
7	0.28	0.37	0.14
8	0.28	0.51	0.26
9	0.36	0.34	0.15
10	0.52	0.77	0.30
11	0.27	0.11	0.01
12	0.45	0.41	0.11
13	0.26	0.21	0.06
14	0.26	0.10	0.03
15	0.52	0.22	0.06

Tableau 1 : Résultats du seuillage selon la distribution des scores des documents

CONCLUSION

Nous nous sommes intéressés dans cet article, plus particulièrement au problème d'adaptation incrémentale du profil et du seuil de décision. L'adaptation est effectuée à chaque sélection d'un document pertinent.

L'adaptation du profil est basée sur la pondération des termes candidats extraits du document en arrivée. Pour chaque terme, un poids provisoire est calculée par un système d'équations sous contraintes. Ce poids provisoire contribue partiellement au poids final de chaque terme dans le profil en utilisant une technique de distribution de gradient.

L'adaptation du seuil, est basée sur une technique probabiliste, elle se base sur le distribution des probabilités des scores des documents d'un échantillon choisi et mis à jour continuellement, à chaque arrivée d'un document.

Des expérimentations ont été réalisées sur une Base **Reuters** issue de TREC-10. Les résultats obtenus montrent l'efficacité de notre processus adaptatif par rapport aux autres systèmes ayant effectués des tests sur la même base.

Nos futurs travaux concernent, l'intégration de la technique de détection de nouveauté durant le filtrage. Autrement dit, chaque fois qu'un document pertinent est sélectionné le système doit détecter si le document est porteur de nouvelles informations ou non. Ainsi, notre processus d'adaptation du profil et du seuil doit agir en fonction du type d'information que le document contient.

REFERENCES

[Belkin & al., 1992] N. J. Belkin, W. B. Croft. "*Information retrieval and information filtering: Two sides of the same coin?*". CACM, pages 29-38, 1992.

[Boughanem, 2000] M. Boughanem. "*Formalisation et spécification des systèmes de recherche et de filtrage d'information*" . HDR de l'université Paul Sabatier de Toulouse, 2000.

[Boughanem & al., 2002] M. Boughanem, M. Tmar. "*Incremental adaptive filtering: Profile learning and threshold calibration*". Proceedings of ACM-SAC, pages 640-644, 2002.

[Hoashi & al., 1999] K. Hoashi, K. Matsumoto, N. Inoue, K. Hashimoto, "*Experiments on the TREC-8 filtering track*". Proceedings of TREC-8, 1999.

[Kwok & al., 1999] K. L. Kwok, L. Grunfeld, M. Chan. "*TREC-8 Adhoc, query and filtering track experiments using PIRCS*". Proceedings of TREC-8, 2000.

[Malone & al., 1987] T. W. Malone, K. R. Grant, F. A. Turbak, S. A. Brobst, M. D. Cohen. "*Intelligent information sharing systems"*. CACM, 30(5), pages 390-402, 1987.

[Robertson & al., 1976] S. Robertson, K. Sparck Jones. "*Relevance weighting of search terms*". JASIS, 27(3), pages 129-146, 1976.

[Robertson & al., 1999] S. E. Robertson, S. Walker. "*Okapi-Keenbow at TREC-8*". Proceedings of TREC-8, 1999.

[Robertson & al., 2000] S.E.Robertson, H. Hull. "*The TREC-9 filtering track final report"*. TREC-P, 2000..

[Saporta, 1996] G. Saporta. *"Probabilités, analyse des données et statistiques*". édition Technip, 1996.

[Sutton & al., 1998] R. S. Sutton, A. G. Barto. *"Reinforcement learning: An introduction*". MIT Press, Cambridge, MA, 1998.

[Tmar, 2002] M. Tmar. *"Modèle auto-adaptatif de Filtrage d'Information: Apprentissage incrémental du profil et de la fonction de décision* ". Thèse de l'Université Paul Sabatier de Toulouse, 2002.

[Zhai & al., 1998] C. Zhai, P. Jansen, E. Stoica, N. Grot, D. Evans. "*Threshold Calibration in Clarit Adaptive Filtering*". Proceedings of TREC-7, pages 149-156, 1998.