01
|

Apprentissage incrémental des profils dans un système de filtrage d'information

M. BOUGHANEM, H. TEBRI, M. TMAR

UPS-IRIT-SIG
118, route de Narbonne
F-31062 Toulouse Cedex 4
{*boughane*, *tebri*, *tmar*}@ irit.fr

Résumé. Cet article présente une méthode d'apprentissage des profils dans les systèmes de filtrage d'information. Le processus d'apprentissage est effectué d'une manière incrémentale au fur et à mesure que les informations sont filtrées et jugées par l'utilisateur. Des expérimentation effectuées sur une collection de test de référence TREC [1], montrent que la méthode permet effectivement l'amélioration des profils.

1 Introduction

Le Filtrage d'Information (FI) est un processus dual à la Recherche d'Information (RI) comme le montre Belkin dans (Belkin 1992). Il traite des documents provenant de sources dynamiques (News, Email, etc.) et décide à la volée, si le document correspond ou pas aux besoins en information des utilisateurs, besoins modélisées au travers du concept de profils utilisateurs. Dans les deux cas, l'objectif est de sélectionner les informations répondant aux besoins des utilisateurs.

Compte tenu de la dualité RI et FI, bon nombre de modèles de filtrage d'information sont basés sur des modèles de recherche d'information augmentés par une fonction de décision, le plus souvent de type seuil. D'une façon générale, les documents et les profils sont représentés par des listes de mots pondérés. Le filtrage d'information revient à comparer chaque document, qui arrive dans le système, aux différents profils. Ceci consiste, d'une façon générale, à mesurer un score de similarité entre le document et le profil, si le score est supérieur au seuil le document est accepté sinon il est rejeté. La difficulté majeure en FI vient du fait qu'en l'absence de collection de référence, la détermination de ce seuil et des pondérations adéquates associées aux profils et aux documents est tout simplement impossible. Car dans un système de filtrage d'information, au démarrage du processus, on ne dispose d'aucune connaissance sur les documents à filtrer pour pouvoir construire une fonction de décision, ni pour identifier les mots clés pouvant représenter les profils. La solution adoptée, dans la majorité des travaux actuels, consiste à démarrer le processus de filtrage en initialisant le profil avec des mots clés extraits du texte du profil et le seuil à une valeur arbitraire, puis adapter et apprendre le seuil et le profil au fur et à mesure que les documents arrivent. Cette approche est appelée filtrage incrémental ou "*Adaptive filtering*" dans la terminologie TREC (Voorhees 2001).

La majorité des techniques d'adaptation de profil proposées dans la littérature sont inspirées du principe de reformulation de requêtes. Les techniques utilisées sont principalement basées sur une version incrémentale de l'algorithme de Rocchio (Rocchio 1971), on y trouve les travaux de (Callan 1998), (Shapire et al. 1998), ou des techniques basées sur les classifieurs Bayesiens (Kim et al. 2000), les réseaux de neurones (Kwok et al. 2000) et les techniques génétiques (Boughanem et al. 1999).

1. Text REtrieval Conference

Concernant le seuillage, les méthodes proposées tentent de définir un seuil qui permet d'optimiser une fonction d'utilité. Une fonction d'utilité permet de mesurer la capacité d'un SFI à sélectionner que des documents pertinents(Voorhees 2001).

Nous nous intéressons dans cet article à l'apprentissage du profil dans un cadre purement incrémental. Contrairement à ce qui se fait dans les autres travaux, qui souvent utilisent des collections d'apprentissage ou effectuent l'apprentissage sur un lot de documents, nous proposons une approche adaptative et incrémentale. Aucune information autre que le profil initial n'est connue au démarrage du processus de filtrage. Les statistiques des termes, important pour les mesures de poids, sont actualisées au fur et à mesure que le système reçoit des documents.

Cet article est organisé comme suit : la section 2 décrit le modèle de filtrage d'information, en terme de représentations des profils et des documents et le processus d'appariemment. Dans la section 3, nous présentons la méthode d'apprentissage des profils. Cette méthode incrémentale est basée sur un principe d'apprentissage par renforcement. Enfin la section 4, est consacrée à l'expérimentation et aux résultats. Les expérimentations sont effectuées sur une collection de test de référence TREC.

2 Le modèle de filtrage

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.

Un profil $p^{(t)}$ est un ensemble de termes sans les mots vides. Il est représenté sous une forme vectorielle, où à chaque terme tp_i est associé un poids $w_i^{(t)}$, t représente l'instant où le système reçoit un document.

Initialement, les termes du profil peuvent être saisis par un utilisateur ou extraits à partir d'un ensemble de documents représentant le centre d'intérêt de l'utilisateur. Le poids du terme dans le profil à l'étape initiale est calculé comme suit : $w_i^{(0)} = tfp_i * (\max_j(tfp_j))^{-1}$, où tpf_i est la fréquence du terme tp_i dans le profil. Ce poids sera ajusté par apprentissage à chaque fois un document est sélectionné pertinent.

A chaque arrivée d'un document, celui-ci est indexé. Le résultat de cette opération est une liste de termes. Le poids $d_i^{(t)}$ de chaque terme dans le document est calculé par une fonction de pondération utilisée dans le système de recherche d'information **Mercure** (Boughanem 2000).

Le processus de filtrage consiste à mesurer un score, noté $rsv(d^{(t)}, p^{(t)})$ entre le document et le profil, défini par le produit scalaire entre le document $d^{(t)}$ et le profil $p^{(t)}$. Ce score est ensuite comparé à un seuil de filtrage, pour décider si le document est accepté ou non : si $rsv(d^{(t)}, p^{(t)}) \geq seuil^{(t)}$, où $seuil^{(t)}$ est le seuil à l'instant t, alors le document $d^{(t)}$ est sélectionné, sinon il est rejeté. Le profil et les statistiques liées à la pondération des termes des documents sont appris à chaque arrivée d'un document pertinent. Ainsi, si un document est jugé pertinent alors l'apprentissage est déclanché.

Nous présentons dans la section suivante la méthode d'apprentissage du profil. Nous ne détaillons pas l'adaptation du seuil, les lecteurs intéressés peuvent se référer à (Tmar 2002).

3 Apprentissage des profils

L'apprentissage des profils que nous utilisons est basé sur un principe de renforcement (Sutton 1998). A cet effet, on considère que quand un document $d^{(t)}$ est jugé pertinent, il faut trouver une représentation du profil $p_x^{(t)}$ qui permet de retrouver le document avec un score fort, soit β. Ceci revient donc à trouver le profil tel que $\mathrm{rsv}\big(\mathrm{d}^{(t)}, p_x^{(t)}\big) = \beta$. Autrement dit, il faut chercher les $pw_j^{(t)}$ qui satisfont l'équation suivante :

$$\sum_{t_i \in d^{(t)}, tp_j \in p^{(t)}, t_i = tp_j} d_i^{(t)} pw_j^{(t)} = \beta \tag{1}$$

Cette équation admet évidemment une infinité de solutions. Pour pallier ce problème, nous proposons d'ajouter une contrainte pour réduire le nombre de solutions et donc arriver à une solution unique.

Avant de donner cette contrainte, nous précision la notion du profil et du poids idéal. Nous appelons profil idéal à l'instant t, le profil qui sélectionne tous les documents pertinents et que les documents pertinents à l'instant t du filtrage . Le poids idéal est le poids d'un terme dans le profil idéal.

La contrainte à intégrer est : si le poids idéal d'un terme t_i est $f_i^{(t)}$=f($\mathrm{d}_i, r_i^{(t)}, s_i^{(t)}$) et si le poids du terme dans le profil est $pw_i^{(t)}$, alors $pw_i^{(t)}/f_i^{(t)}$ est une constante, où $r_i^{(t)}$ (resp. $s_i^{(t)}$) représente le nombre de documents pertinents (resp. non pertinents) contenant le terme tp_i à l'instant t. Le système à résoudre devient alors :

$$\begin{cases} \sum_{t_i \in d^{(t)}, tp_j \in p^{(t)}, t_i = tp_j} d_i^{(t)} pw_j^{(t)} = \beta \\ \forall (t_i, t_j) \in {d^{(t)}}^2, \ \frac{pw_i^{(t)}}{f\left(d_i^{(t)}, r_i^{(t)}, s_i^{(t)}\right)} = \frac{pw_j^{(t)}}{f\left(d_j^{(t)}, r_j^{(t)}, s_j^{(t)}\right)} \end{cases} \tag{2}$$

La solution du système 2 est l'ensemble des poids du profil qui permet de retrouver le document $d_i^{(t)}$. Pour retrouver tous les documents pertinents, il faut combiner les solutions de l'équation pour tous les documents pertinents. Par conséquent, une solution correspond à des poids provisoires qui vont intervenir dans le poids global du profil. Soient n le nombre de termes distincts dans le document à l'instant t, et $f\left(d_i^{(t)}, r_i^{(t)}, s_i^{(t)}\right)$. Le système 2 peut être réécrit en : $\forall i \in \{1 \dots n\}$

$$\begin{cases} \frac{pw_1^{(t)}}{f_1^{(t)}} = \frac{pw_i^{(t)}}{f_i^{(t)}} \quad \Leftrightarrow \quad pw_1^{(t)} d_{j_1}^{(t)} = f_1^{(t)} d_{j_1}^{(t)} \frac{pw_i^{(t)}}{f_i^{(t)}} \\ \qquad\qquad \vdots \\ \frac{pw_n^{(t)}}{f_n^{(t)}} = \frac{pw_i^{(t)}}{f_i^{(t)}} \quad \Leftrightarrow \quad pw_n^{(t)} d_{j_n}^{(t)} = f_n^{(t)} d_{j_n}^{(t)} \frac{pw_i^{(t)}}{f_i^{(t)}} \end{cases} \tag{3}$$

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

$$\forall i, \ pw_i^{(t)} = \frac{\beta f\left(d_i^{(t)}, r_i^{(t)}, s_i^{(t)}\right)}{\sum_j f\left(d_j^{(t)}, r_j^{(t)}, s_j^{(t)}\right) * d_j^{(t)}} \tag{4}$$

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 testé deux fonctions pour l'estimation du poids idéal :

- La première fonction que nous avons proposée est représentée par f_1 :

$$f_1(d_i^{(t)}, r_i^{(t)}, s_i^{(t)}) = d_i^{(t)} * \frac{exp(\gamma * \frac{r_i^{(t)}}{R^{(t)}}) * (1 - exp(\gamma(\frac{s_i^{(t)}}{S^{(t)}} - 1)))}{exp(\gamma)} \quad (5)$$

Où :γ est un paramètre correcteur utilisé pour renforcer l'allure de la courbe représentative de f_1, il vaut 3, $R^{(t)}$ ($S^{(t)}$) est le nombre de documents pertinents (non pertinents) sélectionnés à l'instant t.

Dans le dénominateur de l'équation de f_1, $exp(\gamma)$ est introduit pour normaliser ce poids afin de forcer son appartenance à l'intervalle $[0, 1]$.

L'application de cette fonction montre que les termes qui apparaissent dans tous les documents pertinents et qui n'apparaissent pas dans les documents non pertinents ont un poids idéal maximal (= 1). Au contraire, Les termes qui apparaissent dans aucun document pertinent et dans tous les documents non pertinents ont un poids nul.

- la fonction f_2 est une forme de la formule **BM25**(Robertson et al.,1976) :

$$f_2(d_i^{(t)}, r_i^{(t)}, s_i^{(t)}) = d_i^{(t)} * log(1 + \frac{r_i^{(t)}(S^{(t)} - s_i^{(t)})}{(s_i^{(t)} + 1)(R^{(t)} - r_i^{(t)} + 1)}) \quad (6)$$

On constate, avec f_2, que plus le terme apparaît dans les documents pertinents et moins il apparaît dans les documents non pertinents, plus son importance croît.

Enfin, L'adaptation du profil consiste à utiliser les poids provisoires $pw_i^{(t)}$ pour contribuer à l'apprentissage des termes dans le profil. Nous utilisons la formule de distribution de gradient suivante :

$$w_i^{(t+1)} = w_i^{(t)} + log(1 + pw_i^{(t)})$$

4 Expérimentation et résultats

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 collection Reuters, fournie par **TREC'10**. La corpus de test de **Reuters** est constitué d'un ensemble de 783484 documents et de 84 topics. Ces derniers sont utilisés pour construire les profils initiaux (un topic=un profil).

Le but de cette expérimentation est de comparer et évaluer l'effet des fonctions f_1 et f_2 mesurant le poids idéal d'un terme dans le profil. En utilisant à chaque fois une fonction, on fait passer tous les documents par le filtre sans modification du seuil ($seuil^{(t)} = 0 \forall t$). Ainsi,pour chaque profil, nous déterminons quelle fonction permet de mieux discriminer les documents pertinents et les documents non pertinents. Mais comme il n'existe pas de mesure standard pour l'évaluation de l'aptitude de discrimination si on n'applique pas le seuil, nous proposons d'appliquer un seuil virtuel variable et de mesurer l'utilité à chaque variation du seuil.

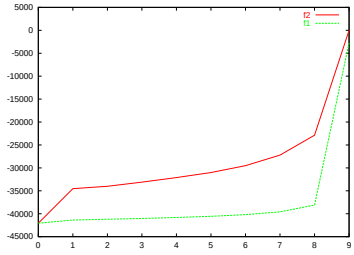

FIG 1. Evolution de l'utilité par f_1 et f_2

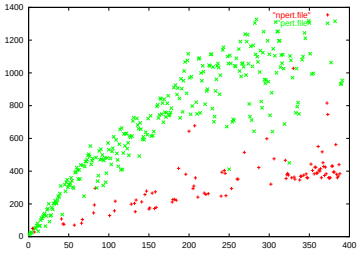

FIG 2. Représentation des scores ($f2$)

Cette façon n'est sans doute pas la meilleure méthode de seuillage, mais elle permet d'évaluer l'aptitude du système à discriminer les documents pertinents et les documents non pertinents. L'évolution des valeurs d'utilité U (ici $2 * R_+ - S_+$) permet de donner une idée sur cette aptitude. Où R_+ (S_+) est le nombre de documents pertinents (non pertinents) sélectionnés.

Les seuils sont exprimés par des fonctions linéaires, nous construisons alors 10 équations linéaires permettants de séparer l'espace des documents en 9 régions, chaque région est un ensemble de documents. L'équation linéaire de chaque seuil dépend alors de l'angle α que fait le seuil avec la droite des absisses, ici $n * \frac{\pi}{18}, n \in \{0, 9\}$. L'équation du seuil est alors : $D_s : y = tg(\alpha) * x$.

Nous avons mesuré l'aptitude de discrimination de notre modèle en utilisant les fonctions f_1 et f_2, les 15 premiers profils et les documents du corpus **Reuters**. Nous avons calculé la moyenne d'utilité à chaque valeur de n pour l'ensemble des profils. La figure 1 illustre l'évolution de l'utilité pour chacune des fonctions f_1 et f_2. Nous constatons que les deux fonctions permettent de discriminer plus ou moins bien les documents pertinents et les documents non pertinents. La fonction f_2 permet de mieux discriminer, par conséquent, cette fonction sera retenue pour la suite de nos expérimentations.

On constate également que notre approche d'adaptation permet effectivement de séparer les documents pertinents des documents non pertinents. La figure 2 illustre les scores des documents du corpus **Reuters** au cours du filtrage dans le cas du profil 1.

5 Conclusion

Nous nous sommes intéressés dans cet article, plus particulièrement au problème d'adaptation incrémentale du profil. L'adaptation du profil est déclenchée à chaque arrivée d'un document pertinent. Elle est basée sur l'apprentissage par renforcement . Ceci revient à résoudre une équation consistant à trouver le profil permettant de retrouver ce document pertinent avec un score fort. Les solutions de cette équation, appelés poids provisoires, vont représenter la contribution de ce document dans le profil global. Les poids provisoires des termes sont ajoutés aux termes du profil dans une équation de distribution de gradient. Des expérimentations ont été réalisées sur une collection Reuters issue de TREC'10. Les résultats obtenus montrent que la méthode d'apprentissage proposées permet effectivement l'amériolation des profils.

Références

N. J. Belkin, W. B. Croft. Information retrieval and information filtering : Two sides of the same coin?. CACM, pages 29-38.

M. Boughanem, C. Chrisment, L. Tamine, (1999), Query space exploration based on genetic algorithms, Information Retrieval Journal.

M. Boughanem. Formalisation et spécification des systèmes de recherche et de filtrage d'information . HDR de l'université Paul Sabatier de Toulouse.

J. Callan, (1998), Learning while filtering documents. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 224-231.

Y.H. Kim, S.Y. Hahn, B.T. Zhang, (2000), Text filtering by boosting Naïve Bayes classifiers, Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 168-175. ACM Press.

K.L. Kwok, L. Grunfeld, N. Dinstl, M. Chan, Queens College, CUNY, (2000), TREC-9 Cross Language, Web and Question-Answering Track Experiments using PIRCS, Proceedings of TREC-9, pp. 419.

S. Robertson, K. Sparck Jones, (1976), Relevance weighting of search terms. JASIS, 27(3), pages 129-146.

J. J. Rocchio, (1971), Relevance feedback in information retrieval, In The SMART retrieval system experiments in automatic document processing, Prentice Hall Inc., pages 313-323.

R.E. Schapire, Y. Singer, A. Singhal, (1998), Boosting and Rocchio applied to text filtering, Proceedings of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 215-223.

R. S. Sutton, A. G. Barto,(1998), Reinforcement learning : An introduction, MIT Press, Cambridge, MA.

M. Tmar, (2002), 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.

E.M. Voorhees, (2001), Overview of TREC'10, The 10th Text REtrieval Conference.

Summary

This paper presents a profile learning method in the information filtering. This method is based on an reinforcement process. Experiments carried out on TREC collection showed the effectiveness of the method.