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