Etude de La Pertinence lors de La Sélection de Collections dans les Systèmes Distribués

Kheira Mechach*, Lougmiri Zekri*, Mustapha Kamel Abdi* ,**

*Département d'informatique
Université d'Oran1 BP 1524 El-M'naouer Maraval, Oran, Algérie
mechach.kheira@gmail.com, lougmiri@gmail.com
**abdi.mustapha@univ-oran.dz

1 Introduction à Sélection basée sur le Degré de Pertinence

Les librairies digitales sont actuellement très répandues. Elles renferment des quantités d'informations énormes et nécessitent des mécanismes efficaces d'indexation et de manipulation. Les moteurs de recherche du type général ne peuvent pas les indexer car ils exigent que l'information qu'ils manipulent soit composée d'entités indépendantes. Dans le besoin de traiter rapidement et efficacement les requêtes, des méthodes basées des approches différentes ont été inventées. On rencontre alors, des méthodes se basant sur les réseaux bayésiens comme CORI Callan et al. (1995), d'autres méthodes qui se basent sur les statistiques TF*IDF. Il existe aussi des méthodes qui se basent sur le modèle de langage et la pseudo-pertinence. Ces méthodes utilisent des résultats déjà obtenus pour de réponses futures. Puisque le modèle centralisé souffre du problème de passage à l'échelle, certaines méthodes ont été mises pour tourner sur les systèmes pair-à-pair. La méthode CORI a été une source d'inspiration et a été utilisée comme moyen de classification dans beaucoup de travaux. Cette méthode fonctionne sur un système bayésien pour localiser des réponses probables aux utilisateurs. La fonction de score donnée dépend de certains paramètres obtenus à partir d'expérimentations sur des data-sets. Ce paramétrage fait que CORI est devenue instable. Ces paramètres doivent être réajustés pour chaque nouvelle collection. Afin de réduire le nombre de collections interrogées, Abbaci et al. (2002) présente la méthode CS. Celle-ci définit ndoc le nombre de documents à retourner et tient compte uniquement des deux premiers termes lors de l'évaluation des requêtes longues. Bien que l'objectif de réduction de flux est atteint, CS produit des faux positifs et faux négatifs importants à cause des restrictions imposées.
Soit un système distribué où un serveur appelé courtier est lié à un ensemble de serveurs. Le courtier détient un index Terme/Serveur qui indique pour chaque terme t_i la liste des serveurs qui le manipulent. Chaque serveur S_i est responsable d'une collection de documents c_i et manipule un index Terme/Documents. Cet index définit pour chaque terme t_i la liste des documents où il figure. Par cette définition, le courtier sélectionne de façon déterministe le sous-ensemble de serveurs pertinents. Ces index permettent de réduire la charge du système. Un document est jugé pertinent s'il partage au moins un terme avec la requête. Plus un document partage de termes avec la requête plus son degré de pertinence s'élève, induisant ainsi que le score d'une

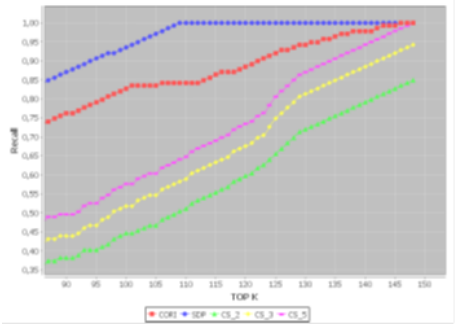

FIG. 1 – *Comparaison selon le Recall.*

collection est proportionnel aux nombre de documents pertinents qu'elle contient. Sur cette définition, pour une requête q, le score d'une collection c_i se calcule selon la fonction SDP suivante :

$$SDP(c_i, q) = (\frac{1}{Nd_{ji}} + Nd_{ki}) * (\sum_{t=0}^{|q|} TF_{ti})$$

Où Nd_{ji} est le nombre de documents de $c_i/d_j \cap q \neq \emptyset$. Nd_{ki} est le nombre de documents de $c_i/d_j \cap q = q$. TF_{ti} est la fréquence du terme t dans la collection c_i. L'expérimentation des trois méthodes sur le dataset Reuters21578, sur un système distribué. La figure Fig. 1 présente la comparaison entre les trois méthodes en fonction du recall. Nous avons réalisé des expérimentations intensives en faisant varier le Top-k. Nous remarquons que les valeurs pour cette métrique sont plus grandes dans SDP que dans les autres méthodes. CS (CS2 pour ndoc=2, CS3 pour ndoc=3, CS5 pour ndoc=5) a retourné un recall plus faible. C'est certainement à cause du ndoc qui influence la recherche. Avec ndoc=2 et 3 ; le recall n'atteint pas 1 c-à-dire il existe des documents pertinents et rares où le système n'arrive pas à les sélectionner. CORI c'est placé au-dessus de CS.

Références

Abbaci, F., J. Savoy, et M. Beigbeder (2002). Méthodes pour la sélection de collections dans un environnement distribué. *Congrès Documents dans les systèmes d'information mobiles* , 227–238.

Callan, J., Z. Lu, et W. Croft (1995). The dynamics of chain formation in oecophylla longinoda. *Journal of Insect Behavior* , 679–696.

Summary

This paper presents a new function of collection selection. Our function is free of any extra-collection parameter and is based on the documents relevance. The ranking of a collection is proportional to its number of relevant documents.