Extraction de la localisation des termes pour le classement des documents

Annabelle MERCIER*, Michel BEIGBEDER*

*École des Mines de Saint-Etienne
158 cours Fauriel F 42023 Saint-Étienne Cedex 2 FRANCE
mercier,beigbeder@emse.fr

Résumé. Trouver et classer les documents pertinents par rapport à une requête est fondamental dans le domaine de la recherche d'information. Notre étude repose sur la localisation des termes dans les documents. Nous posons l'hypothèse que plus les occurrences des termes d'une requête se retrouvent proches dans un document alors plus ce dernier doit être positionné en tête de la liste de réponses. Nous présentons deux variantes de notre modèle à zone d'influence, la première est basée sur une notion de proximité floue et la seconde sur une notion de pertinence locale.

1 Introduction

Le domaine de la recherche d'information, bien connu à travers les moteurs de recherche sur le Web, utilise différents modèles. Ces derniers précisent comment sélectionner et ordonner les documents qui répondent aux besoins d'informations des utilisateurs. Il en existe principalement trois familles (Baeza-Yates et Ribeiro-Neto, 1999) : (a) les modèles ensemblistes (booléen, à ensembles flous et booléens étendus), (b) les modèles algébriques (vectoriel et indexation sémantique latente) et (c) les modèles probabilistes (basés sur les réseaux d'inférence, les réseaux bayésiens et les réseaux de croyance). Notre modèle est basé non seulement sur les familles de modèle ensemblistes et algébriques, mais aussi sur une des premières idées fondatrice de la recherche d'information formulée par Luhn (Luhn, 1958) qui consiste à s'appuyer d'une part, sur la fréquence des termes et d'autre part sur la position relative des termes de la requête dans les documents. Le premier aspect relatif à l'utilisation de la fréquence des termes a été beaucoup développé dans le cadre des modèles algébriques, par contre, le second concernant la proximité entre les occurrences des termes n'a reçu que peu d'attention, notre étude permet d'approfondir ce dernier point.

Tout d'abord, nous rappelons certains modèles classiques ainsi que les quelques méthodes qui utilisent la proximité. Ensuite, nous présentons les deux variantes de notre modèle à zone d'influence avant de conclure.

2 État de l'art

La méthode d'indexation associée à un modèle de recherche d'information permet de construire les représentants des documents et s'appuie généralement sur les occurrences des termes trouvés dans les documents. Nous notons T l'ensemble des termes et D celui des documents.