Expérimentations sur un modèle de recherche d'information utilisant les liens hypertextes des pages Web

Bich-Liên Doan*, Idir Chibane**

* Supélec, Plateau de Moulon, 3 rue Joliot Curie, 91 192 Gif/Yvette, France
Bich-Lien.Doan@supelec.fr
** Supélec, Plateau de Moulon, 3 rue Joliot Curie, 91 192 Gif/Yvette, France
Idir.Chibane@supelec.fr

Résumé. La fonction de correspondance, qui permet de sélectionner et de classer les documents par rapport à une requête est un composant essentiel dans tout système de recherche d'information. Nous proposons de modéliser une fonction de correspondance prenant en compte à la fois le contenu et les liens hypertextes des pages Web. Nous avons expérimenté notre système sur la collection de test TREC-9, et nous concluons que pour certains types de requêtes, inclure le texte ancre associé aux liens hypertextes des pages dans la fonction de similarité s'avère plus efficace.

1 Introduction

Les systèmes de recherche d'information (SRI) sont composés essentiellement de deux modules. Un module d'indexation qui représente les documents, et un module d'interrogation qui représente la requête. La fonction de correspondance permet de calculer le degré d'appariement entre les termes de la requête et les termes d'indexation des documents afin d'évaluer la pertinence des documents par rapport à la requête. Avec le succès grandissant du Web (Google recense plus de 4 milliards de pages Web) le classement des réponses devient critique. Aussi des fonctions de correspondance prenant en compte les liens hypertextes ont vu le jour. En réalité, la plupart des fonctions de correspondance utilisées par les systèmes de recherche hypertextes combinent une mesure de pertinence calculée en fonction du contenu de la page et de la requête utilisateur avec une mesure de popularité de la page qui elle, est indépendante de la requête. Cette dernière mesure repose sur la structure du Web, considéré comme un graphe orienté de pages et de liens. L'hypothèse (Savoy et Rasolof 2000) stipule qu'une page est supposée être de bonne qualité si elle a beaucoup de liens entrants, en particulier, si les pages qui pointent vers elle sont aussi de bonne qualité. Un certain nombre de systèmes qui tiennent compte de la structure du web dans la fonction de correspondance ont été développés. Les systèmes les plus connus sont InDegree, PageRank (Brin et Page 1998) utilisé dans Google, HITS (Kleinberg 1998) et SALSA (Lempel et Moran 2000). Dans ces systèmes, la fonction de correspondance, qui relie la requête aux documents est remplacée par une fonction de classement des résultats qui elle est indépendante de la requête. L'étude des systèmes existants nous a permis de conclure que toutes les fonctions de correspondance basées sur les liens hypertextes ne dépendent pas des termes de la requête. Cela a diminué considérablement la précision des résultats retrouvés. En effet, l'analyse du comportement des utilisateurs dans leur recherche montre qu'ils ne s'intéressent pas aux pages populaires, si ces dernières ne contiennent aucun terme de la requête.