Analyse des paramètres de recherche d'information: Etude de l'influence des paramètres sur les résultats

Josiane Mothe [*,**], Marion Moulinou[*]

[*] Institut de Recherche en Informatique de Toulouse, IRIT
UMR 5505 CNRS, Université de Toulouse,
118 Route de Narbonne 31062 Toulouse CEDEX 9
[**]ESPE, Ecole interne Université Toulouse II Jean Jaurès,
56 av. de l'URSS 31079, Toulouse
prenom.nom@irit.fr, http://www.irit.fr/~ Prenom.Nom

Résumé. Cet article présente une analyse détaillée d'un ensemble de 2 millions de résultats de recherche d'information obtenus par différents paramétrages de systèmes de recherche d'information. Plus spécifiquement, nous avons utilisé la plateforme Terrier et l'interface RunGeneration pour créer différentes exécutions (run en anglais) en modifiant les modèles d'indexation et de recherche. Nous avons ensuite évalué chacun des résultats obtenus selon différentes mesures de performance de recherche d'information. Une analyse systématique a été menée sur ces données afin de déterminer d'une part quels étaient les paramètres qui ont le plus d'influence, d'autre part quels étaient les valeurs de ces paramètres les plus susceptibles de conduire à de bonnes performances du système.

1 Introduction

Un système de recherche d'information (SRI) est un module logiciel qui sélectionne, à partir d'une collection de documents, une liste de documents potentiellement pertinents en réponse à une requête utilisateur. Le processus suivi par un SRI est composé de 3 étapes.

Indexation. Cette étape permet de passer d'un document textuel à un document qui peut être utilisé dans la RI. Elle se base sur l'extraction des mots les plus importants des textes. Lors de cette étape, les mots vides tels que *le*, *la*, *les* sont généralement éliminés ; les termes sont ensuite racinisés, c'est-à-dire que des règles de transformation sur les termes sont appliquées afin d'obtenir un radical, limitant les variantes des termes à une forme unique ; enfin une pondération reflète l'importance des différents radicaux obtenus. Dans un cadre non dynamique, l'indexation est réalisée sur l'ensemble des documents, avant toute recherche.

Calcul des scores de pertinence des documents. Lorsqu'une requête est soumise au système, des scores de pertinence sont attribués aux termes qui la composent, en tenant compte de leur présence dans les documents. Ces scores sont ensuite combinés pour calculer le score global de chacun des documents de la collection. Il existe de nombreux modèles de pondération. La plupart sont basés sur les facteurs TF et IDF. L'expression TF (Term Frequency) correspond à la fréquence du terme dans le document, tandis que l'IDF (Inverse Document