Une méthode d'évaluation de la pertinence des pages Web dans WebSum

Olfa Jenhani El Jed

118 route de Narbonne- Université Paul Sabatier-
Institut de recherche en Informatique à Toulouse.
Jenhani@irit.fr
http://www.irit.fr/ilpl

Résumé. Dans cet article nous présentons une méthode d'évaluation de la pertinence des pages Web retournées par un moteur de recherche.

Ce travail s'inscrit dans le cadre du projet de recherche WebSum qui est un système de résumé automatique de pages Web offrant un moyen de visualisation rapide et structuré des réponses retournées par un moteur de recherche suite à une requête utilisateur.

Afin de produire le résumé, WebSum procède par le classement des réponses récupérées depuis un moteur de recherche (Google) par ordre de pertinence à l'aide d'une métrique qui fait l'objet de ce présent article.

Pour la définition de notre métrique et de ces différents paramètres, nous nous sommes basés sur un corpus de 300 pages Web collectées à partir de réponses d'un moteur de recherche (Google) à différentes requêtes (20 requêtes) représentant des recherches simples sur le Web autour de domaines grand public (médecine, société et éducation). Ce corpus est réparti en trois échantillons de 100 pages chacun. Un échantillon contenant des pages pertinentes, un deuxième contenant des pages de qualité moyenne et le troisième de pages non pertinentes.

L'étude analytique de ce corpus nous a permis de définir notre métrique qui est donnée par l'équation (1) et détaillée dans ce qui suit.

$$R_{doc} = C_f \times (\alpha \,.\, C_{stat} + \beta.\, C_{synt}) \quad (1)$$

Cette métrique se base sur différents critères permettant de juger la pertinence d'une page qui viennent compléter ceux habituellement utilisés par la plupart des moteurs de recherche (popularité, proximité des termes de la requête, etc.). Nous avons identifié trois critères complémentaires pour l'identification d'une page pertinente: 1) C_f critère de forme de la page. C'est un critère booléen qui vérifie si la page contient (=1) ou non (=0) du texte exploitable, 2) C_{stat} critère statistique. Il désigne la fréquence d'occurrence des termes de la recherche dans la page et 3) C_{synt} critère morpho-syntaxique. Il vérifie la bonne forme linguistique de la page en privilégiant l'utilisation des pronoms de la troisième personne et du temps présent. A travers plusieurs expérimentations, nous avons constaté que les valeurs de α et β doivent être fixer à 0.5 afin de donner la même importance pour l'évaluation de la page aux deux critères C_{stat} et C_{synt}. Autrement dit, une page peut avoir une fréquence d'occurrence des termes de la requête élevée mais une mauvaise forme linguistique (forum de discussion, etc.) et vice versa (page traitant un document autre que celui de la recherche).

Finalement, notre métrique vient compléter les critères de pertinences utilisés par les moteurs de recherche et sélectionne les pages pertinentes par rapport à la requête utilisateur et par rapport aux besoins du résumé (document bien formé et exploitable par le système).