Recherche ciblée de documents sur le web

Amar-Djalil MEZAOUR

LRI, Université Paris Sud, 91405 Orsay Cedex
mezaour@lri.fr, http://www.lri.fr/∼mezaour

Résumé. Les langages de requêtes mots-clés pour le web manquent souvent de précision lorsqu'il s'agit de rechercher des documents particuliers difficilement caractérisables par de simples mots-clés (exemple : des cours java ou des photos de formule 1). Nous proposons un langage multi-critères de type attribut-valeur pour augmenter la précision de la recherche de documents sur le web. Nous avons expérimentalement montré le gain de précision de la recherche de documents basé sur ce langage.

1 Introduction

De par sa croissance et son développement, le web représente aujourd'hui une source importante de données hétérogènes (news, articles, photos, vidéos...). Les informations y sont stockées sous forme de documents identifiés d'une manière unique par urls et reliés entre eux par des liens hypertextes. Rechercher ou consulter une information particulière consiste à retrouver les urls des documents susceptibles de la contenir. Les moteurs de recherche ont été développés pour offrir aux utilisateurs des outils simples, mais néanmoins puissants, pour rechercher des documents sur le web. Un moteur de recherche (ex Google [Google, 2003]...) se décompose grossièrement en deux parties : un index web et un langage de requêtes utilisateur. L'index peut être vu comme un immense entrepôt de données où les documents webs sont stockés et indexés par mots clés après avoir été rapatriés par un robot explorateur. Un langage de requêtes mots-clés est proposé aux utilisateurs pour interroger l'index et accéder aux documents web qu'il contient. Pour cela, l'utilisateur spécifie une requête dans laquelle il précise l'ensemble des mots-clés caractérisant, selon lui, le ou les documents à rechercher. Cet ensemble de mots clés est soumis à l'index afin de retrouver les urls de documents contenant le plus d'occurences de ces mots. Les réponses renvoyées sont généralement très nombreuses, peu précises et ne correspondent pas nécessairement aux pages souhaitées par l'utilisateur. Il y a à cela deux raisons majeures. D'une part, le pouvoir expressif des requêtes d'un langage mots-clés ne permet pas de cerner avec exactitude les pages souhaitées. En effet, une requête mots-clés est limitée à la spécification des mots pertinents que doivent contenir les pages pour être considérées comme réponses, sans autre possibilité de décrire d'autres caractéristiques d'une page. Ainsi, pour la recherche de documents à faible contenu textuel (exemple : images, pdf...) ou pour la recherche de documents caractérisables autrement que par des mots clés (exemple : cours java ou c++), les requêtes mots-clés se montrent inappropriées. D'autre part, l'approche même qui considère toute page contenant les mots clés fournis dans une requête comme pertinente, sans tenir compte de la localisation de leur présence ni de la structure du document ni de son contexte accentue d'avantage l'imprécision des réponses. Par exemple, pour une recherche de documents de cours en c++, l'utilisateur soumet naïvement la requête « cours c++ » à un moteur de recherche sans autres alternatives pour décrire des cours c++. Le moteur de recherche renvoie en réponses quelques cours c++ mais aussi des documents