

DOCUMENT ANALYSIS AND SYSTEM DESIGN

An enhancement of Boolean retrieval systems based on term co-occurrence frequencies / Un amélioration des systèmes d'information basée sur la cooccurrence de termes

S. K. M. Wong
W. Ziarko

V. V. Raghavan
(Computer Science Dept., University of Regina)

Most commercial systems for information retrieval employ the standard Boolean retrieval strategy for providing response to user queries. One of the major problems in this context is the inability of such systems to provide ranked output. Furthermore, there have not been very many studies that attempt to incorporate dependencies between the index terms into the retrieval process. In this presentation, a scheme is proposed by which the Boolean retrieval strategy can be enhanced by using dependencies based on term co-occurrence frequencies. Experiments performed on two experimental collections demonstrate that the retrieval performance of the proposed scheme is significantly better than the standard approach. If the preprocessing time required to determine the amount of term-term dependencies is ignored, then the computing time to process a query in the proposed environment is of $O(1)$; i.e. it is independent of the number of terms in the query.

La majorité des systèmes d'information documentaire emploient la méthode booléenne afin d'obtenir une réponse suite à une réquisition. Un des problèmes majeurs de cette méthode est l'incapacité de fournir une pondération attachée à chaque résultat. Il y a très peu d'études visant à incorporer un facteur d'indépendance entre les éléments de l'index lors du processus de recherche de documents. Ce document de recherche propose une technique permettant d'améliorer la méthode booléenne conventionnelle en introduisant un facteur de dépendance basé sur la cooccurrence de documents. Les résultats d'expériences faites sur deux collections limitées de documents en utilisant la technique proposée ont démontré une amélioration significative en comparaison avec la méthode booléenne conventionnelle. Si le temps-machine utilisé lors de la détermination du facteur de dépendance entre les documents est ignoré, alors le temps-machine requis pour répondre à une réquisition en utilisant la technique proposée est de l'ordre $O(1)$ ce qui veut dire que le nombre de termes utilisé lors de la réquisition n'influence pas la vitesse d'exécution du processus.