

IMPROVING RETRIEVAL RESULTS THROUGH CITATION ANALYSIS

S. Raman Venkataraman
Science Division, Main Library
University of British Columbia
Vancouver, B.C. V6T 1Y3

ABSTRACT

Present information retrieval techniques are designed to match keywords in a query with keywords in a document or document representation in order to retrieve documents which would contain the information desired by the user. Since relevance of a document to a user depends on many variables, a document retrieved through matching keywords may or may not be relevant to a user. In the quest for better retrieval results, many models have been developed. Attempts have been made to improve the range of useful relations between terms by ranking documents in the order of their potential relevance to a user query and by the use of thesaurus terms. These techniques are based on statistical relationships among the words in a query and the collections of documents. They fail to bring the keywords or thesaurus terms semantically close to those of the query.

Information searchers in addition to using databases indexed by keywords and descriptors also use citation databases. Citation searches for information retrieval are based on the rationale that citing documents are related to cited documents. Very little overlap is found between citation and semantic retrievals. Search output is augmented by combining the subject search with author search and a co-citation search. The two types of approaches are not combined, but used independently of each other.

At present, information retrieval is usually carried out using keyword and/or thesaurus terms to match the query with the document. Not all the documents retrieved by this approach are found to be relevant by the user. The relevant documents are selected by the user by browsing through the titles as well as the abstracts of the documents retrieved. In the present study, it is assumed that if a document retrieved through keyword/s and or descriptor/s is authored or cites a paper/s by researchers who were/are actively publishing in the specific area of research in which the user is interested, then that document will be relevant to the user. The main object of this paper is to examine whether citation analysis can be used to select papers relevant to the user from the set retrieved by keyword or descriptor search. The test involved matching documents selected by citation analysis (papers authored by or citing the work of selected researchers) with those selected by the user through browsing. The results are discussed in the context of improving retrieval and reducing the cumbersome task of browsing through numerous titles for selecting relevant documents from the set of documents retrieved by keyword and/or descriptor.

Amélioration des résultats du repérage grâce à l'analyse des citations

S. Raman Venkataraman
Science Division
University of British Columbia
Vancouver, BC

Les techniques de recherche documentaire actuelles sont conçues pour appareiller les mots clef d'une demande aux mots clef d'un document ou de sa représentation dans le but d'extraire des documents contenant l'information requise par l'utilisateur. Plusieurs variables entrent en ligne de compte lorsqu'il s'agit d'évaluer la pertinence d'un document, et un document repéré par combinaison de mots clef ne présente pas nécessairement d'intérêt pour un usager. Dans le but d'améliorer les résultats du repérage, plusieurs modèles ont été mis au point. Des tentatives ont été faites d'améliorer l'éventail de relations utiles entre les termes en classant les documents par ordre de pertinence éventuelle pour la demande d'un usager et en utilisant des termes tirés de thésauri. Ces techniques sont fondées sur les relations statistiques existant entre les termes d'une recherche et les collections de documents. Elles échouent dans leur tentative de rapprocher sémantiquement les mots clef ou les termes d'un thésaurus à ceux de la demande.

En plus de bases de données indexées par mots clef et par descripteurs, les professionnels de l'information font appel à des bases de données de citations. Dans le cadre de la recherche documentaire, la recherche par citation est fondée sur le raisonnement qu'il existe une relation entre documents citant d'autres documents et documents cités. On retrouve très peu de chevauchement entre recherche sémantique et recherche par citation. Les résultats de la recherche sont accrus en combinant la recherche par sujet à celle par auteur et par co-citation. Les deux approches sont utilisées indépendamment l'une de l'autre.

Aujourd'hui, la recherche documentaire se fait habituellement en combinant mots clef et/ou termes tirés de thésauri de manière à assortir la demande au document. Tous les documents repérés grâce à cette approche ne sont pas considérés pertinents par l'utilisateur. Ce dernier choisit les documents pertinents en parcourant les titres et les résumés des documents repérés. On suppose dans la présente étude qu'un document sera pertinent pour un usager si ce document, repéré au moyen de mots clef ou de descripteurs, a pour auteur ou cite un document produit par des chercheurs qui sont/étaient actifs dans le domaine de recherche qui intéresse cet usager. Le but premier de cette communication consiste à évaluer la justesse de l'utilisation de l'analyse des citations pour sélectionner des documents pertinents à partir d'un ensemble repéré par recherche sur mots clef ou sur descripteurs. L'expérience consistait à appareiller des documents repérés suite à l'analyse des citations (des communications ayant pour auteur des chercheurs choisis ou citant leurs travaux) à des documents choisis par balayage par l'utilisateur. Les résultats font l'objet de discussions visant à améliorer le repérage de documents et à réduire le travail fastidieux du balayage de longues listes pour choisir des documents pertinents à partir d'ensembles préalablement repérés à l'aide de mots clef et/ou de descripteurs.