

全文・引用検索システムの構築 - 図書館情報学分野を対象として -

金子 千容

研究活動において、他の研究者の研究成果を利用することは不可欠である。他の研究者の過去の研究成果を入手する方法としては、論文の引用文献を参考にすることが多いことが分かっており、論文検索システムにおける引用検索の必要性は明らかである。近年、Webや電子ジャーナルの普及によって研究に関する情報の入手は容易になった。しかしながら、現在普及している論文検索システムの中に全文検索・引用検索を十分な形で提供しているものはない。そこで本研究では、全文検索・引用検索可能な論文検索システムを構築した。対象分野は日本の図書館情報学とする。

本システムには、日本図書館情報学会誌と Library and Information Science の論文 93 本を収録した。PDF 形式のファイルで入手した各論文からテキストデータを抽出、さらに転置索引ファイルを作成し、全文検索の検索対象とした。引用検索を可能にするために、各論文の引用索引も作成した。各論文の引用文献・被引用文献を表示する機能に加えて、全文検索結果として表示された複数の論文が共通して引用している文献・共通して引用されている文献の表示も行なう。以上の機能を同時に提供することで、より多くの適合文書を提示することを目指す。

本研究では全文検索機能の有効性を検証するために、各論文中で検索対象とする範囲が異なる論文検索システムを作成し、その性能を比較した。それぞれの検索対象範囲は、(1) タイトル、(2) 第 1 章、(3) 最終章、(4) 第 1 章と最終章、(5) 重要文(TF-IDF)、(6) 重要文(出現率)、(7) 全文、の 7 種類である。(2)~(7)は、本文中の重要文に着目した検索範囲である。それぞれの検索範囲で、広い検索テーマ 3 種類と狭い検索テーマ 3 種類の、計 6 種類の検索テーマに関する正解文書を適切に検索できるかテストした。評価には精度、再現率、F 値の 3 つの指標を用いた。また引用検索機能の有効性を検証するために、全文検索結果のみの再現率と、引用検索結果も含めた再現率の比較を行なった。

結果、狭い検索テーマにおいては本文中の重要文に着目した検索範囲の有効性が示せた。広い検索テーマに関しては全体として検索範囲をタイトルのみにした方が有効であった。しかし、検索結果の上位だけを見ると検索範囲を全文にしても同等の有効性が示された。さらに、全文検索結果の共通する引用文献リストや被引用文献リストによって再現率の上昇が確認出来た。全文検索と引用検索を併用することによって、より多くの正解文書を検索できることを示せた。

以上のことから、本研究では論文検索システムに全文検索と引用検索を導入する有効性を示すことが出来た。

(指導教員 辻慶太)