文書の分析結果に基づく検索機能を備えた個人向け文書管理機能の開発

蒐場　広翔

近年のパーソナルコンピュータ（PC）の普及に伴って、個人所有文書の効率的な管理の需要が高まり、個人向け文書管理機能の拡充が必要だろうと考えられる。本研究ではそのような文書管理機能のうち、PC 内に保存された文書をその PC 上で検索する個人向け文書検索機能に存在する課題に着目し、それを解決する機能の開発を目的とする。

既存の文書検索機能は、ファイル名や本文内容を対象とした全文検索機能と、作成日時や作成者などの文書属性の指定による絞り込み検索機能が主流である。しかし、前者の全文検索機能はユーザの指定したキーワードと文書内容の部分一致検索を行うため、文書内容に誤字があったり、ユーザの記憶違いによってキーワードと文書内容に含まれる単語が完全一致しなかったりして目的の文書を見つけ出せない場合がある。特に個人が作成する文書では、ユーザ自身で作成と使用が完結するメモであることもあり、誤字脱字が起こる可能性が高いと考えられる。一方、後者の文書属性の指定による絞り込み検索機能はファイル形式や文書作成日時などのメタデータ依存の抽出であり、文書内容を反映した検索のためには前者の全文検索機能を使用することになる。

そこで本研究ではそのような目的の文書が見つからない課題を解決するために、文書群を分析することで単語を抽出し、その単語をユーザが検索に指定するキーワード候補として表示する検索機能の開発を目的とし、その開発を通じて必要な文書属性と必要な機能を明らかにする。具体的には、検索対象となる文書群から単語を集計・抽出してキーワード候補として画面に表示し、ユーザがキーワードを選択することで目的の文書を探し出す文書検索機能を開発する。文書内に含まれた単語のみをキーワードとして提案することで、ユーザが指定した検索キーワードと検索対象の文書内容が不一致になり結果にたどり着けないという課題を解決することができると考えた。

単語群の表示方法として 3 種類のワードクラウドを用いた検索機能を実装し、tf 値、df 値、tf-idf 値に基づく語に基づく 3 つのワードクラウドを作成して、筆者自身が過去に作成・編集した講義のメモやレポートを中心とした個人文書 407 件について 10 通りの検索要求を作成して実験を行った。その結果をもとに検索結果として正しかった結果を検索結果一覧で割ったものを適合率、検索結果として出力された結果数を検索結果として出力が望まれる結果の文書数で割ったものを再現率として計測し集計し評価した。

結果としてワードクラウドを用いたキーワード提案による文書検索機能は tf-idf 値によって各文書の特徴的な単語を集計して表示することで効果的な検索結果が得られた。また、特定の 1 つの文書を絞り込む場合に効果的な機能である可能性があることが分かった。

（指導教員　阪口　哲男）