

5W-7 全文検索における検索結果の可視化による文書選択支援手法の提案

佐藤 亮介 三石 大 佐々木 淳 船生 豊
岩手県立大学ソフトウェア情報学部

1 はじめに

大量の文書から、必要とする文書の検索を行うための手法として全文検索がある。しかし、検索結果として得られた文書が必ずしも自分の求める文書であるとは限らず、その中から目的の文書を選び出すことは容易ではない。そこで本稿では、全文検索における検索結果を可視化することにより、利用者の文書選択を支援する手法を提案する。

2 既存の検索手法

大量の文書から目的の文書を検索する手段として全文検索がある。この全文検索には、単純に与えられたキーワード列に一致する文書を検索する手法から、より効果的な検索を目的としたシソーラス検索や、文書間の関連性に着目した検索、データマイニング技術等を用いた検索など、多様な手法が提案されている [1][2][3]。

これらの検索手法では、検索結果として得られた文書の一覧が何らかの評価関数に基づき順序付けされた形で提示され、利用者はその中から目的の文書を推定し選択する。しかし多くの検索システムでは、利用者に対しこれらの文書が何を基準に順序付けされているのかが明確でなく、その結果、一般の利用者が大量の検索結果の中から求める文書を選択するのは容易ではない。

これに対し、利用者の文書選択を容易にするために、個々の文書の内容や文書間の関係等を可視化する手法が提案されている [4]。例えば文書レンズでは、文書一覧を2次元平面上に配置し、その中で選択された文書を拡大表示する。これにより利用者は、図が多い、表がある、等の各文書の形態を直感的に判断することが可能となり、その結果目的の文書を容易に選択することが可能となっている。

しかしながらこれらの提案手法の多くは、個々の文書の内容や、複数の文書間の関係を可視化することを目的

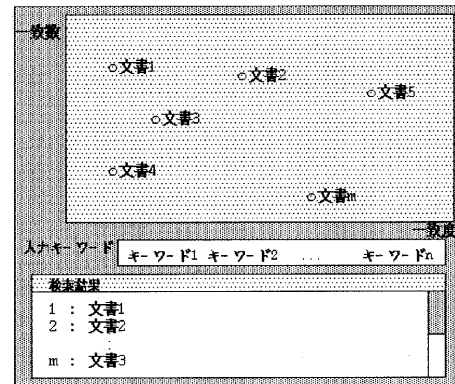


図 1: 文書の2次元平面上への配置

としており、文書検索時における利用者の検索内容と、その検索結果との関係を明確にするものではない。そのため、検索結果として得られた文書集合の全体を確認することが必要となり、大量に得られた文書からの目的の文書の選択は容易ではない。全文検索における大量の検索結果の中から利用者が効率的に目的の文書を選択できるためには、その利用者が行った検索内容と検索結果との関係を明確にすることが必要である。

3 検索結果の可視化による文書選択支援

3.1 検索結果の2段階による可視化

我々は、全文検索における検索内容と検索結果の関係に基づく可視化による、利用者の文書選択支援手法を提案する。

本手法は検索のために与えられたキーワードに対し、1) 検索結果として得られた各文書における重みの合計と、このとき一致したキーワードの個数に基づく各文書の2次元平面上への配置、および、2) 各文書内における個々のキーワードの重みの提示、の2段階からなる。まず、利用者が複数のキーワードによる全文検索を行うと、検索結果として適合する文書の一覧と、各文書毎にその文書における各キーワードの $tf \cdot idf$ 値を取得する。この検索結果に基づき、 $tf \cdot idf$ 値の合計 (一致度) を x 軸、 $tf \cdot idf$ 値が 0 以外のキーワードの個数 (一致数) を y 軸とする2次元平面上に各文書を配置する (図1)。これにより利用者は、検索の結果として得られた文書集合における個々の文書が、検索内容として与えられたキー

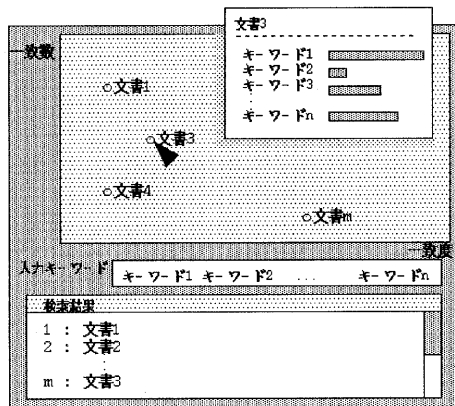


図 2: 各文書毎の各キーワードの $tf \cdot idf$ 値の提示

ワードに対してどの程度一致しているかの概要を掴むことができる。次に、2次元平面上に配置された文書の一つを利用者が選択すると、検索のために与えた個々のキーワード毎のその文書における $tf \cdot idf$ 値が提示される(図2)。これにより、利用者は各文書において個々のキーワードがどの程度の重要性を持っているかを確認することが可能となる。

この様に2段階に検索内容と検索結果との関係を可視化することにより、得られた検索結果全体の概要と、各文書毎の詳細な検索結果の両方を直感的に把握することが可能となると考えられる。

3.2 TSSV システム

今回提案した検索結果の可視化手法に基づき、我々は全文検索における文書選択支援システム:TSSV(Text-Search Support with Visualization) システムを設計した。TSSV システムは、全文検索エンジンに対し、我々の提案する2段階の可視化手法に基づき、得られた検索結果を可視化し、文書選択を行うためのユーザインタフェースを提供するシステムである。

本システムでは、利用者が検索を行うと、その検索結果を先ず2つの領域に提示する。一方は、我々の提案する可視化手法の第1段階の可視化領域であり、また一方は通常の全文検索システムと同様の、文書一覧を提示するための領域である。可視化領域に配置された文書を選択すると、その文書の第2段階の可視化が行われ、各文書毎の詳細な検索結果を得る。また、文書一覧から文書を選択すると、その文書を得ることができる。設計した TSSV システムのシステム構成を図3に示す。

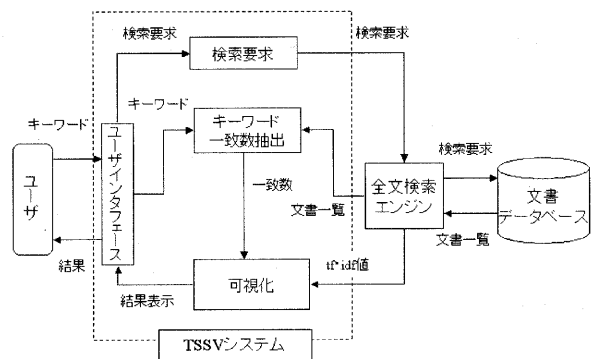


図 3: TSSV システムの構成図

4 まとめ

本稿ではキーワードによる全文検索の検索結果をキーワードの一致数と $tf \cdot idf$ 値に基づいて可視化する手法を提案し、これを基に TSSV システムを設計した。

しかしながら、実際の検索の際には、利用者がキーワードをあまり多く入力しないことも予想され、その場合、今回提案した手法は必ずしも有効とはいえない。このような問題を解決するためには、シソーラス検索や関連性を用いた検索などの手法を併用し、入力された少数のキーワードから充分の数のキーワードを導出し、検索を行う事も必要であると考えられる。

今後、これらの検討を行うと同時に、設計したシステムの実験を行ない、その有効性確認のための実証実験を行っていく予定である。

参考文献

- [1] 西村英樹, 河野浩之, 長谷川利治: WWW データ資源検索システムの実装と評価, 情報処理学会研究会報告, Vol.96, No.68, pp.263-268(1996).
- [2] 金沢輝一, 高須淳宏, 安達淳: 文書関連性を考慮した検索方式, 情報処理学会研究報告, Vol.98, No.58, pp.263-268(1998).
- [3] 川原 稔, 河野浩之: 文献二次情報データベースにおける検索支援, 情報処理学会研究報告 Vol.97, No.64, pp.239-244(1997).
- [4] 武田浩一, 野美山浩: テキスト情報の可視化を利用した情報検索, 情報処理, 41 巻 4 号 pp.343-350(2000).