2U－6　関連文書検索システムの開発 (4) －関連文書検索－

倉持 勉† 永峯猛志† 梅基 宏† 石飛康浩† 倉橋政之† 増市 博†‡ 舘野昌一†

†富士ゼロックス(株) IT 事業開発部 ‡スタンフォード大学 CSLI

1．はじめに

今回開発した関連文書検索システムは、全文検索と関連文書検索の２種類の検索機能を備える。関連文書検索とは、検索者が指定した適合文書（入力適合文書と呼ぶ）に関連した内容の文書を関連度が高い順に出力する機能である。このような検索として、語の文書内頻度（tf）と文書頻度の逆数（idf）の積を語の重要度とし、語の重要度に基づいて検索する方法が提案されている[1]。この方法では、複数件ある入力適合文書の一部の文書に集中的に出現する語は tf が大となり、結果的にそのような語の重要度が過度に大となった場合に、検索ノイズが多く発生することが考えられる。本稿では、入力適合文書が複数件である場合に、高い適合率を得られる関連文書検索方式について報告する。

2．関連文書検索方式

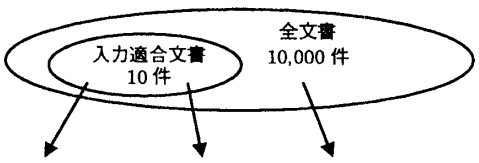
本システムは、入力適合文書に含まれる語の重要度を計算し、語の重要度に基づいて入力適合文書に対する各文書の関連度を計算する関連文書検索機能を備える。次に、関連文書検索における語の重要度計算処理と、文書の関連度計算処理の内容について説明する。

2.1　語の重要度計算処理

語の重要度計算処理では、入力適合文書の内容をよく表している語が高いスコアとなるように、入力適合文書に含まれるすべての語の重要度を計算する。入力適合文書の内容をよく表している語は、各入力適合文書に共通して出現するが、検索対象の全文書における出現文書数が小であるという仮説を立てた。入力適合文書が複数件の場合に、一部の入力適合文書に集中的に出現する語が重要ではないと断定することはできない。しかし、検索者が検索と結果確認を進めて、入力適合文書がある程度の件数に達した場合、一部の文書に集中的に出現する語が重要である確率は低くなると考えた。

語の重要度を判定する方法として、tf*idf という式が提案されている。入力適合文書が複数件である場合に、一部の文書に集中的に出現する語は tf が大となり、その語の重要度が大となる可能性があるため、前記の仮説に反する。

語が入力適合文書に共通して出現するか否かを反映させるため、入力適合文書における出現文書数を df_a、検索対象の全文書における出現文書数の逆数を idf_b として、df_a*idf_b という式を考える。しかし、この式は出現頻度が少ない語の重要度が過度に大となるという問題が指摘されている[2]。そこで、本方式では $df_a^2*idf_b$ という式に基づいて語の重要度を計算した。図1 に本方式により語の重要度を計算した例を示す。



入力適合文書に含まれる語	df_a^2	df_b	重要度 df_a^2/df_b
装置	9^2	9,000	0.009
車両	8^2	200	0.32
自動車	4^2	200	0.08
ナビゲーション	2^2	100	0.04
:	:	:	:

図 1　語の重要度

Relevant Document Retrieval System (4) -- Relevant Document Search --
Tsutomu KURAMOCHI†, Takeshi NAGAMINE†, Hiroshi UMEMOTO†, Yasuhiro ISHITOBI†,
Masayuki KURAHASHI†, Hiroshi MASUICHI†‡, Masakazu TATENO†
†IT Business Development, Fuji Xerox Co., Ltd. ‡CSLI, Stanford University

図1において、df_a が相対的に大であり、かつ、df_b が相対的に小である「車両」という語の重要度が相対的に大となり、df_b が大である「装置」や df_a が小である「ナビゲーション」という語の重要度が小となる。

2.2 文書の関連度計算処理

文書の関連度計算処理では、語の重要度に基づいて検索対象となっている各文書の入力適合文書に対する関連度を計算し、関連度が大きい順に文書をソートする。文書の関連度は、各文書に含まれる語の重要度の総和とした。ただし、同一文書中に2回以上出現する語については、文書の関連度に語の重要度を1回だけ加算した。

3. 実 験

本関連文書検索方式と tf*idf に基づいた検索方式の検索結果を比較した。具体的には、関連度が大きい順に出力された文書列の中で、検索者が検索目的に合致すると判断した文書が出現する順位を比較した。実験では、検索目的を環境にやさしい複写機に関する特許の収集、入力適合文書を検索目的に合致した国内公開特許公報26件、検索対象を全文検索（IPC の OR 検索）で絞り込んだ複写機に関する国内公開特許公報6,480件とした。その結果を図2に示す。

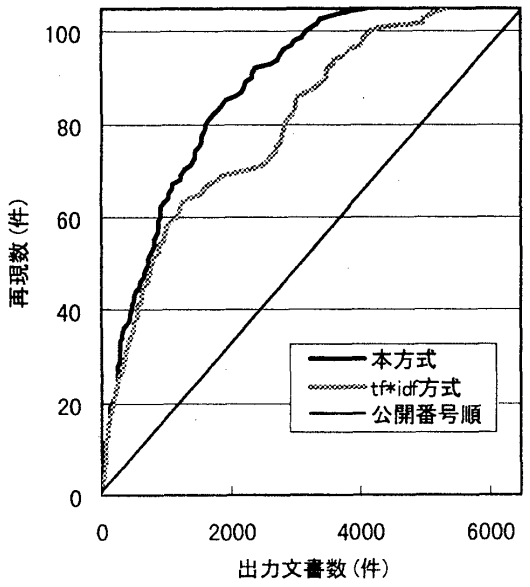


図2 検索結果の比較

図2において、横軸は関連度が大きい順に出力された文書の件数、言い換えると、出力された各文書の順位であり、縦軸は検索者が検索目的に合致すると判断した文書数の累計である。

検索者が上位の文書から順に調査したと仮定すると、図2のグラフは立ち上がりがよい程、一定件数を調査した場合に、検索目的に合致する文書をより多く抽出できることを表す。また、検索目的に合致する一定件数の文書を抽出するために調査する文書数がより少なくなることを表す。

全文検索により得られた6,480件の特許を、例えば公開番号順に調査した場合、調査件数に比例して検索目的に合致する文書を抽出できると考えられる。図2から、本方式と tf*idf に基づいた方式は公開番号順に調査した場合に比べて有効であることがわかった。さらに、本方式は tf*idf に基づいた検索方式に比べてグラフの立ち上がりがよく、検索目的に合致する文書が比較的上位に位置することがわかった。したがって、検索者は関連文書検索結果の上位の文書から順に調査することにより、有用な情報を効率よく抽出できると考えられる。

4. まとめ

検索者が指定した文書に関連した内容の文書を検索する関連文書検索機能を実現した。特に、複数件の適合文書が指定された場合に、高い適合率で検索することを可能にした。このことにより、本システムでは、まず全文検索機能により数件の適合文書を探し出すことができれば、あとは関連文書検索機能によりその適合文書群に関連した内容の文書を容易に抽出することができる。

参考文献

[1]G.Salton et al.："Introduction to Modern Information Retrieval", McGraw-Hill, New York, 1983.

[2]丹羽：「動的な共起解析を用いた対話的文書検索支援」，情報処理学会研究報告 96-NL-115.