2G-2

N-gram 全文検索と概念検索を融合した文書検索方式の検討

亀代 泰三　　永井　明人　　谷垣　宏一　　平野　敬　　岡田　康裕
三菱電機（株）情報技術総合研究所

１．はじめに

文書検索の主な方式には、文書全体から検索キーワードの存在を検証する全文検索と、文書内容とキーワードとの類似性を検証する概念検索がある。全文検索は検索キーワードを含む文書を確実に取得できる反面、類義語・関連語を含む文書を取得するには予めシソーラスを定義する必要があり、検索もれを防止するにはこれらを充実させる必要がある。一方概念検索は検索キーワードに内容が類似する文書を出力するため、類義語・関連語を含む文書も取得できる反面、キーワードを含む文書であっても関連性が低いと検索もれとなる場合がある。本稿では、この全文検索と概念検索を融合することで互いの欠点を補完し、検索精度向上を図る方式について検討した。

２．検索方式詳細

本稿では、N-gram による全文検索方式とベクトル空間モデルを用いた概念検索の融合方式を検討した。以下に個々の検索方式とその融合方式を示す。

２．１　N-gram 全文検索

N-gram 全文検索はキーワード抽出、キーワード検索、スコア計算の３つの処理で構成する。

(1) キーワード抽出：検索クエリを形態素解析し、名詞と未知語を抽出する。名詞は全て検索キーワードとするが、未知語は字種により検索キーワードとしての利用有無を決定する。更に各キーワードに対してシソーラスを用いて類義語展開する。

(2) キーワード検索：抽出したキーワードおよび類義語を用いて N-gram インデックスを検索する。

(3) スコア計算：個々のキーワードによる検索結果から検索クエリに対するスコアを計算する。スコアは、性質の異なる２種類のスコア i)キーワードの TF-IDF 重みより算出したスコア（一般語の影響の抑制）ii)検索クエリ内のヒットしたキーワード数に比例したスコア、により求める。文書 D_i に対する一致度 SZ_{Di} の算出は、式(1)を用いる。

$$SZ_{Di} = k_1 \sum_{t} \log(1+tf(t,D_i)) \cdot (\log \frac{M}{df(t)} + \gamma) \quad \cdots(1)$$

ここで、t は検索キーワード、tf は文書 Di 内の出現頻度、df は t が出現する文書数、M は全文書数、γ は定数、k_1 は正規化係数である。

２．２ 概念検索

一般に、検索クエリに類似する文書を検索する方式としては、ベクトル空間モデルを利用した関連文書検索が知られている。この方式は、文書中の単語出現頻度に基づいた文書ベクトルと、検索クエリベクトルとの類似性が高い文書を検索するものである。しかし、検索クエリの単語が含まれない文書は類似性が低くなるため、検索目的に合致した類似内容であっても、検索クエリの単語と表記が異なる文書を検索できないという課題がある。

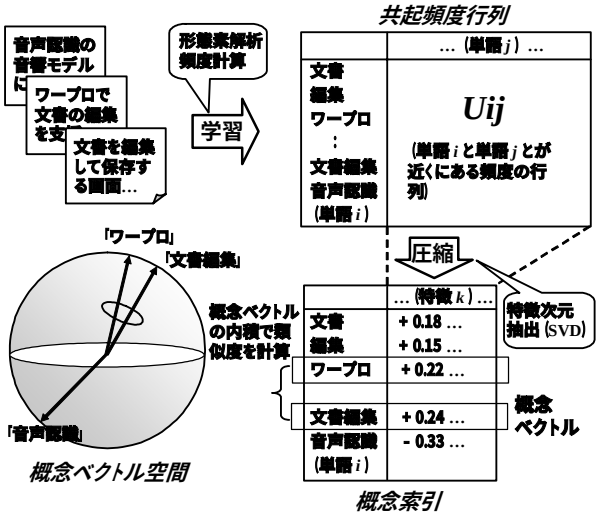

図１ 概念検索方式

これに対し我々の概念検索では図１に示すような特異値分解(SVD：Singular Value Decomposition)によって共起頻度行列から特徴的な次元を自動抽出して概念索引を生成する[1][2]。概念索引には単語間の高次の相関関係が含まれており、「ワープロ」と「文書編集」のような、テキスト中に潜在する重要な関連性が抽出可能となる。文書ベクトルは文中に現れる単語に対応する概念ベクトルの和を正規化して作成する。検索時には、検索クエリを形態素解析し、概念索引を用いて検索クエリベクトルを生成した後、文書ベクトルとの一致度を計算する。検索クエリベクトル $\boldsymbol{q}$ と文書 Di の文書ベクトル $\boldsymbol{d_i}$ との一致度 SG_{Di} は、式(2)で算出する。k_2 は正規化係数である。

$$SG_{Di} = k_2 (\boldsymbol{q} \cdot \boldsymbol{d_i}) / (|\boldsymbol{q}| \cdot |\boldsymbol{d_i}|) \quad \cdots(2)$$

A Study on Document Retrieval Method based on Full Text Search and Concept Search.
Taizo Kameshiro, Akito Nagai, Koichi Tanigaki,
Takashi Hirano, Yasuhiro Okada
Information Technology R&D Center,Mitsubishi Electric Co.
5-1-1 Ofuna, Kamakura, Kanagawa,　247-8501, Japan

２．３　融合方式

N-gram 全文検索と概念検索を用いた融合方式のブロック図を図 2 に示す。提案方式では検索クエリに対して全文検索、概念検索を独立に実行する。そこで得た各検索結果（文書の一致度）から統合一致度を算出し、一致度の高い順に検索結果を出力する。文書 D_i の統合一致度 ST_{Di} は前記 SG_{Di}、SZ_{Di} を用いて式(3)で算出する。

$$ST_{Di} = \alpha * SG_{Di} + (1-\alpha) * SZ_{Di} \quad \cdots(3)$$

$(0 \leqq \alpha \leqq 1)$

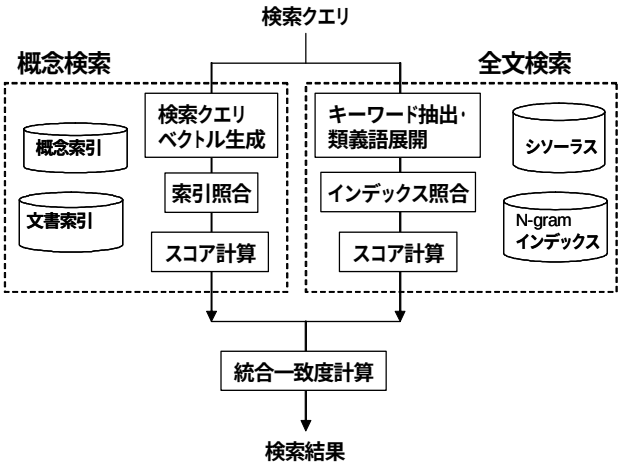

図 2 提案方式ブロック図

３．評価

本方式の有効性を検証するために、特許明細書約 1 万文書を用いて検索を行い、再現率・適合率を算出した。検索に用いたキーワードは「自動販売機」「浄水器」の 2 種類である。正解文書は上記キーワードに関連する文書を人手により抽出した。例えば「自動販売機」では「自動取引装置」「自動交付機」など、「浄水器」では「水処理器」「水質改善方法」などに関する文書も正解とする。また統合一致度を評価するために、提案方式についてαと検索結果出力文書数 N を変化(50～500)させて再現率・適合率を算出した。なお、本実験では全文検索でシソーラスを用いずに検索処理を実行した。上位 N 位における再現率・適合率の算出方法を以下に示す。

再現率=(上位 N 位までに存在する正解数)*100 / (全文書中の正解文書数)

適合率=(上位 N 位までに存在する正解数)*100 / N

図 3 に各方式の再現率を、図 4 に各方式の適合率を示す。比較のために、全文検索および概念検索の単独実行による結果も併せて示す。

図 3 より、提案方式は上位 200 文書以上で再現率が全文検索、概念検索単独に比較して向上している。またα=0.5 での上位 400 文書での再現率は 94.8%であり、この再現率を全文・概念検索の OR 出力で取得するには上位 1300 文書以上を必要とする。

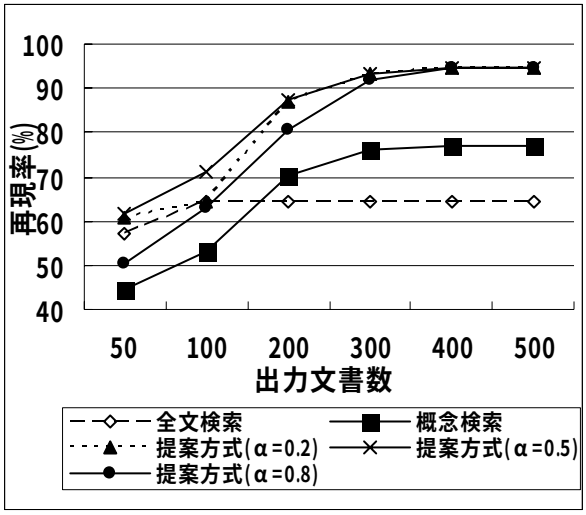

図 3 各方式の再現率

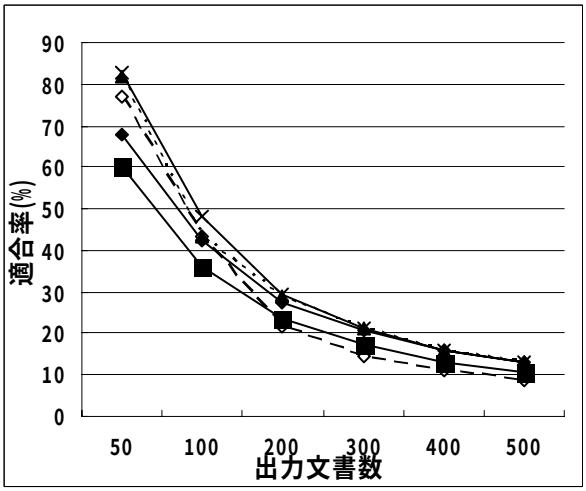

図４ 各方式の適合率(グラフの凡例は図 3 と同一)

検索結果を個別に解析すると、概念検索では「自動販売機」の再現率が低く、全文検索は「浄水器」の類義語の再現率が低いが、融合方式ではこれらをうまく補完して再現率を向上している。

４．まとめと今後の課題

性質の異なる２つの検索方式である全文検索と概念検索を融合することで検索精度が向上することを確認した。今後は、より詳細な評価を行い更なる検索精度の向上を図る。

参考文献

[1]高山 他, “単語の連想関係に基づく情報検索システムＩｎｆｏＭＡＰ”, 情報処理学会 情報学基礎研究会(SIGFI), FI53-1 (1999).

[2]永井 他, “単語共起行列の次元圧縮に基づく概念検索方式の評価”,第 67 回情処全大, 2G-3,2005