2G-3

単語共起行列の次元圧縮に基づく概念検索方式の評価

永井明人† 相川勇之† 高山泰博† 今村誠†
三菱電機株式会社 情報技術総合研究所†

1. はじめに

大量の電子化文書から、自然言語の質問文により目的の類似文書を検索する技術として、単語共起行列の次元圧縮に基づく概念検索方式[1]の開発を行なっている。本方式の特徴は、単語共起行列の次元圧縮で学習された単語特徴ベクトルの類似性に基づいて、異なる表記の単語を含む類似文書が検索可能な点である。本稿では、上記概念検索方式の評価を目的として、特許明細書データに対する検索精度を実験により検証した結果を報告する。

2. 概念検索方式の概要

一般に、質問文に類似した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、質問文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、質問文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、質問文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。

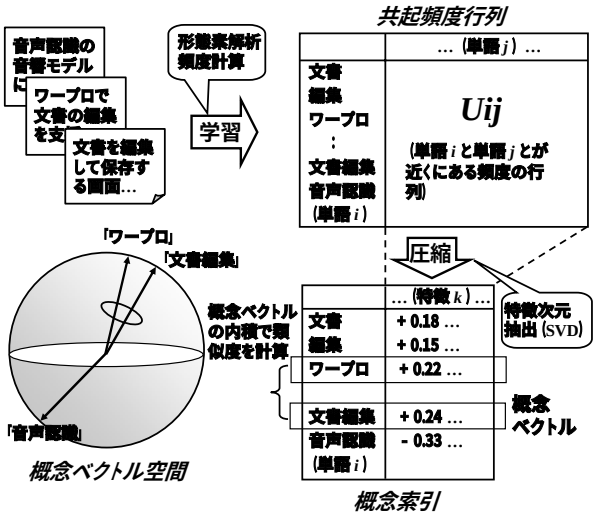

図 1：概念検索方式の原理

"Evaluation of a method of concept search based on dimensionality reduction of word co-occurrence matrix"
† Akito Nagai, Takeyuki Aikawa, Yasuhiro Takayama, Makoto Imamura
† Information Technology R&D Center, Mitsubishi Electric Corporation

これに対し我々は、特異値分解(SVD：Singular Value Decomposition)によって単語共起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概念検索方式[1](図1)を提案した。この概念検索方式の基本アプローチは、文献[2]に基づいており、テキスト中の単語の共起頻度行列(単語共起行列)をベースにして単語間の類似性を学習するものである。学習では、計算上の扱いやすさとデータのスパース性への対処として、単語共起行列を SVD により次元圧縮する。SVD は、任意のサイズの行列を分解する線形代数の手法であり、得られた縮退行列を、単語の概念ベクトルの索引(概念索引)として用いる。この概念索引には単語間の高次の相関関係(association)が含まれており、図 1に示すように、「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になる。

3. 評価

本概念検索方式の検索精度を実験により評価した。以下に評価方法と評価結果を述べる。

3.1. 評価方法

対象データは特許明細書 (1996～2001 年分)の要約 200 万件を用いた。また、比較のために文献[1]の概念検索方式を方式 A とし、ベクトル空間モデルの関連文書検索を比較対象の方式 B とした。質問文は、表 1に示すような 5 単語と 5 文からなる 10 種類を用いた。

検索精度の評価では、検索結果順位の上位 100 位までの妥当性を主観的に○×で判定し、適合率と再現率で評価した。なお、ここで用いた再現率は、方式 A、B の検索結果の和集合の内、○判定とした文書を正解とみなした擬似的な再現率である。

表 1：評価に用いた質問文 10 種類 (5 単語+5 文)

5 単語	「OCR」「文字認識」「ワープロ」「かな漢」「自動販売機」
5 文	「帳票の文字データを正確に自動認識する OCR 方式」 「タブレットのペン入力による手書き文字認識装置」 「文書の編集時に文書構成の全体を容易に把握できる表示方法を備えたワープロ」 「かな漢の候補一覧制御に関する方式」 「自動販売機の硬貨の返却方法」

3.2. 評価結果

検索精度の評価結果を図 2に示す。横軸は検索結果の順位、縦軸は検索精度の適合率と再現率である。検索精度は、ある順位までの検索結果の累積に対するものであり、10 位～100 位までの各順位に対応する適合率と再現率を方式A、B毎に示す。

評価結果より、方式 A(概念検索)の適合率 A は、上位 60 位程度まで約 80%の妥当性を持つこと、及び、再現率 A では、上位 100 位までに、約 70%の正解をカバーしていることが分かる。さらに、方式 B(関連文書検索)との比較では、適合率、再現率ともに方式 A(概念検索)の方が良い結果を示しており、文献[1]の概念検索方式の検索精度における優位性を確認した。

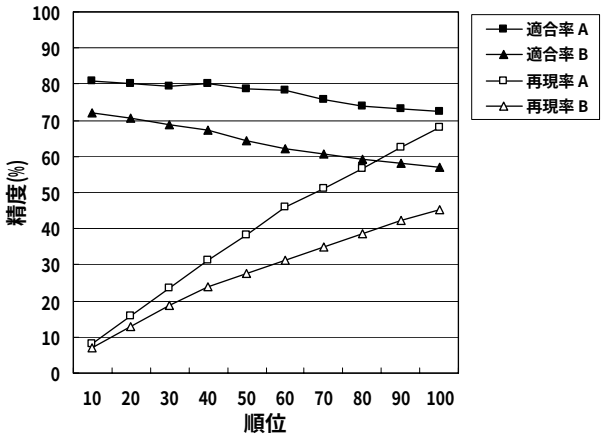

図 2：検索精度(適合率・再現率)の評価結果

3.3. 考察

図 2の評価結果に対して、方式 A(概念検索)による検索結果の内容を分析して、単語の連想関係による検索精度向上の効果を評価した。評価方法は、表 1に示した 5 単語の質問文の内、「OCR」「文字認識」「ワープロ」「かな漢」の 4 単語による検索結果の上位 100 位を視察し、質問文の単語を含まない場合の、検索結果の妥当性を調査した。なお、「自動販売機」については、全ての検索結果で質問文の単語を含んでいたため割愛する。

図 3に、質問文の単語を含む／含まないの割合、及び、含まない場合の検索結果の内容の妥当性を調査した結果を示す。この結果より、各検索結果の中で、質問文の単語を含まない場合は、全体の 46%を占めており、そのうちの 79%は、妥当な内容の検索結果であることが分かる。このため、連想的に検索された異なる表記の単語により、全体の精度に対して、46% × 79% = 36% 分相当の精度向上に寄与があったことが分かる。

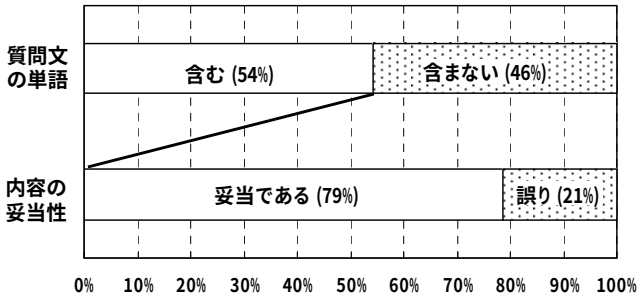

図 3：異なる表記の単語の妥当性

さらに、妥当であると判定された上記 79% の異なる表記の単語を調査した結果、表 2に示すような具体例が学習されていることが分かった。単語の文字列としての類似性のみならず、「OCR」と「文字認識」や、「ワープロ」と「文書編集」といった連想的な単語が学習されていることが分かる。

表 2：異なる表記の単語の具体例

質問文	異なる表記の単語
OCR	文字読取装置、文字認識、手書き文字等の認識
文字認識	文字切り出しおよび認識、手書き文字を認識
ワープロ	文書編集、ワードプロセッサ、文書作成装置
かな漢	仮名漢字変換、漢字かな混じり文に変換

4. おわりに

単語共起行列の次元圧縮に基づく概念検索方式を、特許明細書データを対象として精度評価を行なった。評価の結果、単語の連想関係による検索精度向上の効果を確認した。今後は、異なる分野の文書データに対する精度評価などの、より詳細な評価を進めていく。

[参考文献]

[1] 高山泰博, Raymond Flournoy, Stefan Kaufmann, Stanley Peters, “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] Hinrich Schutze, “*Ambiguity Resolution in Language Learning: Computational and Cognitive Models,*” CSLI Lecture Notes 71, CSLI Publications, 1997. (Ph.D. thesis, Stanford Univ., Dept. of Linguistics, July 1995.)