単語間の係受け情報を用いた文献検索手法

4K－2

池田 和幸†
東京大学大学院工学系研究科

安達 淳††
学術情報センター研究開発部

1 はじめに

計算機技術やネットワークの発達によって、電子化文書は急激な増大を見せているが、その中から目的とする情報を見つけ出すことは現在のキーワードベースの検索では容易ではない。その理由としては

- 検索者が検索要求を問合せに正確に反映することの難しさ
- キーワード文字列のみによる検索の性能的限界

を挙げることができる。

上記の問題の解決策として、我々は文献データベースを対象として、擬似的な自然言語で問合せを発行することで検索要求を問合せに正確に反映させるとともに、単語間の係受け関係を検索に利用することで単語間の語間に存在する意味関係を検索に反映し、検索効果を高めることを提案する。本稿では我々の提案する手法を用いた文献検索システムの概要及び検索実験の結果について論じ、手法の有効性を検証する。

2 検索システムの全体構成

図1に提案する検索システムの全体構成を示す。

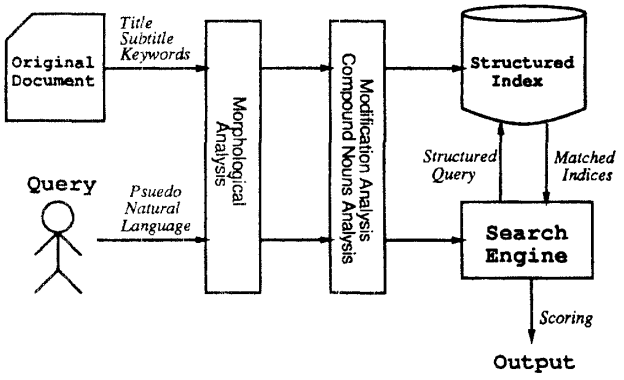

図 1: 提案する検索システム

検索対象文書としては、文献に付与された表題や副題、キーワードを用いる。これらは文献の主題を的確に表現している一方、文書自体は名詞と付属語が交互に出現するという比較的単純な構造を有するため、文書の構成単語間の係受けの付与が通常の自然文と比べて容易に行えるという特徴があり、検索インデックスとして用いるのに適している。検索対象文書は形態素解析、名詞間の係受け解析、複合名詞解析を通じて二分木の形で構造化され、検索インデックスとして格納される。(図2)。

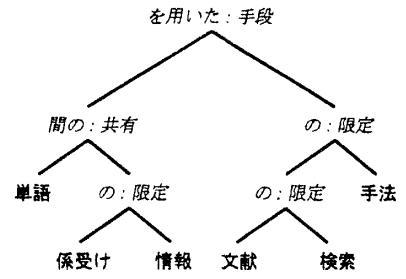

図 2: 構造化された検索インデックス

この検索インデックスは、文書の構成単語を葉、文書中の付属語を内部ノードに配するデータ構造を有し、内部ノードが語句の「係り関係」、内部ノードの左右の子孫がそれぞれ「係る語句」「係られる語句」を表す。本インデックスでは、それらの係受け関係を文全体について階層的に表現できることが大きな特徴である。

一方、問合せは文献表題に見られるような擬似的な自然言語で発行され、検索対象文献と同様の手法で二分木に構造化される。問合せ・検索インデックス双方の二分木の構造のマッチングをとり、適合度に応じた得点付けを行うことで、検索要求に近い文書順に検索結果が出力される。

3 係受け解析

係受け解析は、付属語によって区切られる名詞同士の係受け解析と名詞の構成単語間の係受け解析(複合名詞解析)の2段階に分かれている。名詞同士の解析には、文書中の付属語の並びと文末に位置している語の特徴によってパターン化された係受け規則を用いている。この規則によって、3名詞の文書の92.0%、4名詞の文書の84.7%に正しい係受けが付与された。係受け規則に当てはまらない文に対しては、いくつかの経験則を適用することにより文中の局所的な係受けの成立を考慮し、係受け精度の劣化を防いでいる。

複合名詞解析においては、複合名詞の構成単語同士の意味的な結合関係を正確に把握することが重要である。本システムでは統計的に得られた単語2-gram

---

A Method of Document Retrieval using Dependency Relationships between Words.
Kazuyuki IKEDA†, Jun ADACHI††
†Graduate School of Engineering, University of Tokyo
††Research & Development Department, National Center for Science Information Systems

の頻度を単語同士の結合関係の強さとみなし、単語2-gram頻度の高い語の組に優先的に係受けを付与している。

4 マッチング処理

本手法では、単語間に成立している係受け関係を検索時に考慮することで語間関係を反映した検索を行い、検索要求に近い結果出力を行うことを可能にしている。マッチングでは、問合せに含まれる「係る語句 - 係り関係 - 係られる語句」の三項関係を有する検索インデックスを抽出する。ここで、係り関係については、それを表現する付属語に類義表現が多数存在する。例えば「〜に関する」と「〜については」は表層文字列こそ違え意味はほぼ等しい。そこで、付属語をその意味に応じて20種のカテゴリに分類しマッチングに用いることで、係り関係に「文字列の一致」「意味の一致」「意味も不一致」という3段階のマッチングの段階を定めている。

マッチングの有無は、問合せの二分木構造の部分木が検索インデックスに包含されるか否かで判定される。しかしながらこの場合、いずれかの文書に挿入を含む場合[1]などには係受け関係が一致するにも関わらずマッチしない場合がある。そのため、日本語の係受け規則に則って成立している係受けの一部を削除することで二分木の縮退を適宜行い、検索の柔軟性を高めている。

5 検索実験

検索結果は得点順に整列されて出力される。ここで得点付けにおいては

- 「係る語」「係られる語」の統計的・意味的性質
- 三項関係の二分木中での位置
- 「係り関係」のマッチングの段階(類似度)

を考慮し、それぞれを重み付けすることで得点を与える。また、係受け関係のみを検索に用いると係受け付与の誤りによる検索漏れが懸念されるため、検索時には単語レベルのマッチングもサポートし、マッチした単語に対してTF-IDF法によって得点を付け、係受けによる得点に上乗せしている。

本システムの有効性を検証するために、全く同じ単語を用いた2つの意味の異なる検索文を発行し、その検索結果の違いを見た。問合せは以下の2つである。

Query 1 関係データベースを用いた並列結合演算処理

Query 2 並列関係データベースを用いた結合演算処理

それぞれの検索結果を表1、表2に示す。

表1では「並列結合演算」、表2では「並列関係データベース」がそれぞれ検索要求の中で特徴的な部分である。その部分を持った文書が得点付けの結果上位に来たり、あるいは得点が増加していることが結果から見てとれる。これは、係受けの成立が検索に大きく影響したことを示しているものと考えられる。

[1]例)「データベースの問合せ」と「関係データベースの最適な問合せ」

表 1: Query 1 の検索結果

得点	表題
66.33	スーパーデータベースコンピュータSDC−IIにおける並列結合演算処理に関する性能評価
52.30	並列計算機AP1000DDVにおける多重結合演算の実装とその評価
44.37	Symmetry81における結合演算の並列処理に対する考察
44.37	主記憶データベースに関する研究
39.90	並列SQLサーバSDC-IIにおけるトランスポーズ型ファイル編成適用の検討

表 2: Query 2 の検索結果

得点	表題
53.70	トランスピュータを用いた並列データベースマシンにおける結合演算の性能評価
53.70	並列関係データベースにおけるバッチ問合せ処理最適化技法の検討
50.09	並列計算機AP1000DDVにおける多重結合演算の実装とその評価
50.09	並列データベースシステムにおける多重結合演算の静的最適化技法の一考察
50.09	疎結合並列計算機における多重結合演算の評価

6 まとめ

本稿では、文書中の単語間の係受け関係を検索に用いることで、従来のキーワード型の検索手法よりも検索要求に適合する文書を検索者に提示できることを示した。今後の課題としては、得点付けの評価、検索速度やインデックスのオーバヘッド等の検索効率の問題の改善、文献の全文データからのインデックスの自動抽出が挙げられる。

参 考 文 献

(1) 宮崎:「係受け解析を用いた複合語の自動分割法」, 情報処理学会論文誌, Vol.25, No.6, pp.970–979, 1984.
(2) Douglas P. Metzler et al. "Conjunction, Ellipsis, and Other Discontinuous Constituents in the Constituent Object Parser", *Information Processing & Management* Vol.26, No.1, pp.53–71, 1990.
(3) 池田, 安達:「表題の意味構造を考慮した文献検索手法」, 第53回情報処理学会全国大会, 2T-3, 1996.