抄録検索における構造化インデクスの効果†

2P－6

松村 敦‡　高須 淳宏‡　安達 淳‡
学術情報センター 研究開発部§

1 はじめに

従来からの情報検索の主流となっている検索手法は，単語の出現頻度等の統計的特徴量のみに依存している．このような検索手法では，単語間の意味関係を正確に表現することは出来ず，意味的な曖昧性が残ることによる検索精度の劣化が問題となる．そこで，我々は単語間の意味的な関係として係受け関係に注目し，これを二分木の形に構造化し，文の内部表現として扱う検索システムを検討した．特に，この二分木を検索インデクスとして保持するため，本手法を構造化インデクス検索システムと呼んでいる．これまでタイトルを検索対象としてシステムを開発し，予備的な検索実験で再現率が低いところで TF-IDF 法に比べて適合率向上を実現してきた [1]．ここでは，より高精度の検索システム実現のために構造化インデクスの抄録検索への適用を試みた結果とその分析，今後の構想について述べる．

2 構造化インデクス検索システム

構造化インデクスの一例を図 1 に示す．構造化インデクスは係受け関係を二分木の形で構造化したものであり，三つの要素で成り立っている．一つは文中で概念を表す「概念語」で二分木の葉の部分におかれる．もう一つはそれらの関係を表す「関係語」で内部ノードに配置される．三つ目の要素は，関係語を意味の類似性によって 18 に分類した「カテゴリ」の名前で，これは対応する関係語とともに内部ノードに保持される．

文をこのような形に構造化するために，まずはじめに，形態素解析によって，文を「概念語」と「関係語」の並びに変換する．この際，あらかじめ事例分析で作成した関係語を定義した辞書を用いる．次に，「関係語」の並びと事例分析から得られた係受けパタン辞書を比較し係受けを付与する．複合名詞もその構成要素間の係受けを判定し，二分木の枠組に組み込まれる．こうして，文書内の各文は二分木の形に構造化され，構造化インデクスとしてシステムに保持される．

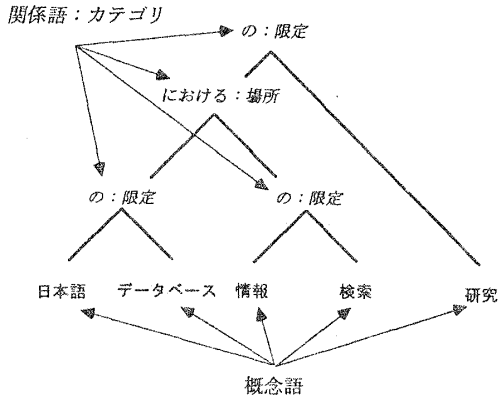

図 1: 構造化インデクス

一方，検索は以下の流れで行われる．はじめに，論文タイトルのような表現の自然文を問合せとして受け付け，インデクス作成と同様な二分木の形の「構造化問合せ」を作る．次に，「構造化問合せ」と「構造化インデクス」の類似度を判定し，これに基づいて得点付けを行い結果を出力する．手法の全体像は図 2 である．

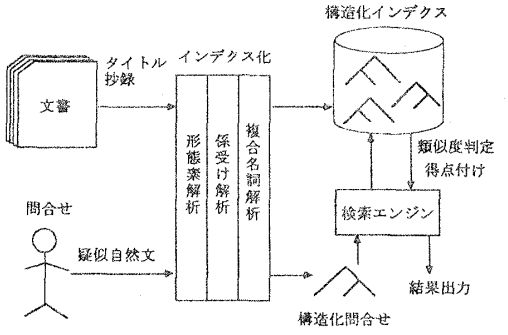

図 2: 構造化インデクスシステムの処理の流れ

抄録検索への適用において最も重要となる問題は，問合せと文書の類似度の定義である．本手法では以下の三つのレベルで係受けを尺度とした類似度を定義している．

†Effect of the Structured Index on Abstract Retrieval
‡Atsushi MATSUMURA, Atsuhiro TAKASU, Jun ADACHI {atsushi, takasu, adachi}@rd.nacsis.ac.jp
§Research & Development Department, National Center for Science Information Systems

問合せと文書の類似度　文書に含まれる複数文中で問合せと最も類似度の高い文を代表文として扱う.

二分木間の類似度　各係受けの類似度の和を用いる.

係受けの類似度　係受けを表す関係語とそのカテゴリの一致度，および係る語，係られる語の TF–IDF の積を利用した係受けの重要度によって定義する.

一方，単語の出現頻度による類似度を TF–IDF 法による得点付けで与える．問合せと文書との類似度を示す得点として，これら二つの要素の重み付け線形和を用いた.

3　抄録検索への適用

NTCIR1 [2] を使って予備的な検索実験を行ったところ，様々な要因が検索結果に影響し，またこれらは問合せ毎に異なる効果を与えることが分かった．いくつかはタイトルベースで開発した本システムの改良すべき点を示す結果となっている.

係受けの一致の条件　現在のシステムでは，タイトルベースでの経験から係受けの一致条件を緩くしている．例えば,「WWW トラヒックの分析」という問合せ文に含まれる係受けは，厳密には「WWW トラヒック」と「トラヒックの分析」であるが,「WWW の分析」という係受けも考慮している．これは，タイトルという情報量の少ない対象に対してはノイズよりも関連文書に適合するという効果があり有効であった．しかしながら，抄録程度の情報量を持つ場合には，逆にノイズが非常に多くなる傾向があり検索精度を落す結果になっている.

一文の長さ　一文が長い場合には同じような係受けが複数一致し，係受け得点が非常に高くなる場合がある．問合せ「分散ネットワーク環境におけるメディア同期問題の解決」では,「問題解決」という係受けが一文で 5 回一致する文と「メディア同期」が 1 回だけ一致する文とでは，前者のほうが係受けの得点が高くなってしまう．その結果，重要な係受けをもちながらランキングが低くなる文書が出てきて検索精度が劣化する.

TF–IDF と係受け点の組合せ　係受けの得点計算には係る語と係られる語の TF–IDF による得点の積を係受けの重要度として考慮しているが，この効果が強過ぎて係受けの得点が TF–IDF に大きく影響されるようになっている．例えば,「ATM 網を用いた TCP/IP 通信のスループット特性」という問合せに対して,「ATM」が 9 回出現する文書と 2 回出現する文書とでは前者のほうが「ATM」に関する係受け点が高くなる．この場合，前者は「ATM 網」という係受けしかない不正解にも関わらず,「TCP/IP」や「スループット特性」を含む正解である後者よりも得点が高くなるということが起こっており，検索精度の劣化につながっている.

4　今後のシステム構想

ここでは，分析結果で明らかになった問題点から，今後のシステム構想について考える.

まず，係受けの一致基準の改良を考える必要がある．係受けの基準の変更による検索結果中のノイズ文書と関連文書を分析し，適当な一致基準を求める必要がある．また，文の長さによる影響を考慮した係受け得点の与え方，係受けの重要度の定義についても考える必要がある.

一方，複数の自然言語文を含む抄録を対象とした検索では問合せと文書の類似度の定義の妥当性も検討する必要がある．現在は最も類似度の高い文を代表する形をとっているが，この方式が情報検索に与える影響を分析し，より有効な類似度の定義として全ての文についての係受けの類似度を反映した形を検討する.

これまでの分析では，以上のような要素が情報検索の精度へ及ぼす影響は問合せや文書集合に完全に依存するものと思われる．今後は，各問合せや文書集合毎にパラメタの最適化を行い，そのパラメタと問合せや文書集合との関係を明らかにしていく必要がある．その結果を利用し，係受けを用いた有効な情報検索システムの構築を試みる計画である.

なお，本研究は，著者らが NTCIR ワークショップに参加し，学術情報センター研究開発部が「学会発表データベース」のデータの一部を使用してデータ提出学会¶の理解の下に構築した「テストコレクション 1」を利用して行った [2].

参考文献

[1] Atsushi Matsumura, Atsuhiro Takasu, Jun Adachi : “Information Retrieval Method using Structured Index for Japanese Text”, *Proceedings of the 3rd International Workshop on Information Retrieval with Asian Languages(IRAL'98)*, 15-16 October 1998, p.109-115.

[2] 情報検索システム評価用テストコレクション構築プロジェクト
http://www.rd.nacsis.ac.jp/~ntcadm/

¶http://www.rd.nacsis.ac.jp/~ntcadm/acknowledge/thanks1-ja.html