係り受け情報を用いた全文検索とその評価

新美和彦, 兵藤安昭, 池田尚志
岐阜大学工学部
〒501-11 岐阜県岐阜市柳戸 1-1
Tel: 058-293-2740
Fax: 058-293-2740
E-Mail: {kazuhiko,hyodo,ikeda}@ikd.info.gifu-u.ac.jp

概要

大量の電子化文書に容易にアクセスできる環境が整ってくるにつれて，その中からユーザが必要とする情報のみを正確に検索する技術がますます必要不可欠となってきている。従来の検索手法では，文書中に単語が出現するか否かに関するブール演算による絞り込みが主として用いられてきた。そのほか、単語間の関係による絞り込みとしては近接演算が用いられてきたが，近接演算では単語間の正確な関係を表現することは出来ない。本論文では，単語間の係り受け情報を用い，高精度な絞り込みが可能な全文検索システムについて述べる。特許データを対象とした検索実験で，係り受け関係を用いた検索精度は適合率９２．１１%，再現率９６．０１%を示し，近接関係を用いた検索と比べ良好な結果を得た。また，インデックス容量の増加も２７%程度に抑えることが出来た。

キーワード

全文検索，係り受け情報，骨格構造解析，近接演算

Full-text retrieval using dependency structure and its evaluation

Abstract

Today we can easily access a lot of large scale electoronic documents, and with these advance the eager wish for high precision text retrieval is increasing. In this paper we propose , to improve the precision , the full-text retrieval using dependency relation between words instead of proximity operation. The proximity relation has been used as a substitution for syntactic relation so far, because the syntactic analysis is still a difficult task for a computer. We apply our method of skeletal syntactic analysis for Japanese to full-text retrieval and evaluate the index size, response time, accuracy of retrieval and others verifying usefulness of this method.

Keyword

Full Text Retrieval, Dependency Structure, Skeletal Syntactic Analysis, Proximity Operation

１．はじめに

大量の電子化文書に容易にアクセスできる環境が整ってくるにつれて，その中からユーザが必要とする情報のみを正確に検索する技術がますます必要不可欠となってきている。従来の検索手法では，文書中に単語が出現するか否かに関するブール演算による絞り込みが主として用いられてきた。そのほか、単語間の関係による絞り込みとしては近接演算が用いられてきたが，近接演算では単語間の正確な関係を表現することは出来ない。

検索精度向上のためには言語情報の活用が効果的である。[兵藤 ９６] は構文解析情報を活用した翻訳支援のための類似用例検索について述べている。本論文では,『ある単語がある単語に係る』という係り受け情報を利用した高精度な全文検索システムについて述べる。

係り受け情報は形態素解析及び構文解析を施すことにより作成されるが，長文に対する安定した構文解析はまだ困難である。本システムにおいては表層的情報のみを用いて行う骨格構造解析法 [兵藤 ９５] によって、文書データベースの係り受け解析を行った。

特許データを対象とした検索実験で，係り受け関係を用いた検索精度は適合率９２．１１%，再現率９６．０１%を示し，近接関係を用いた検索と比べ良好な結果を得た。また，インデックス容量の増加も２７%程度に抑えることが出来た。

２．全文検索システム

２．１ システム概要

本システムは，図１に示すように文書データベース，係り受け解析部，インデックス部，照合部，インタフェース部から構成される。インデックス部は，１次記憶上の単語エントリと，２次記憶上の単語出現情報及び係り先情報とから成る。これらのインデックスは文書データベースに対する係り受け解析処理の結果から作成される。照合は２段階に分けて行われる。ユーザが，係り受け関係を含む検索パターンを入力すると，まず始めに，１次記憶上の単語エントリを検索し，単語が出現する文を抽出する。次に検索されたすべての文を対象として，検索パターンと係り受け構造が一致するか否かの照合を行い，検索結果をインターフェース上に表示する。検索システムはサーバ上にあり、ユーザはＷｅｂ上のインターフェースを通して検索する。

２．２ 係り受け解析部

対象とする文書には，形態素解析および係り受け解析を施す。係り受け解析には骨格構造解析 [兵藤９５] を用いた。骨格構造解析とは必ずしも完全な係り受けの構造を求めるものではなく，並列構造の解析など意味に立ち入らなければ解析できない部分は曖昧なブロックとしてそのまま残し，文全体の構造を把握しようとするものである。解析例を図２に示す。

２．３ インデックス部

インデックスは，１次記憶上の単語エントリと，２次記憶上の単語出現情報及び係り先情報から成る（図３参照)。

単語エントリはパトリシア構造を用いて構築している。現在のところ、数字・記号を除くすべての自立語を登録している。

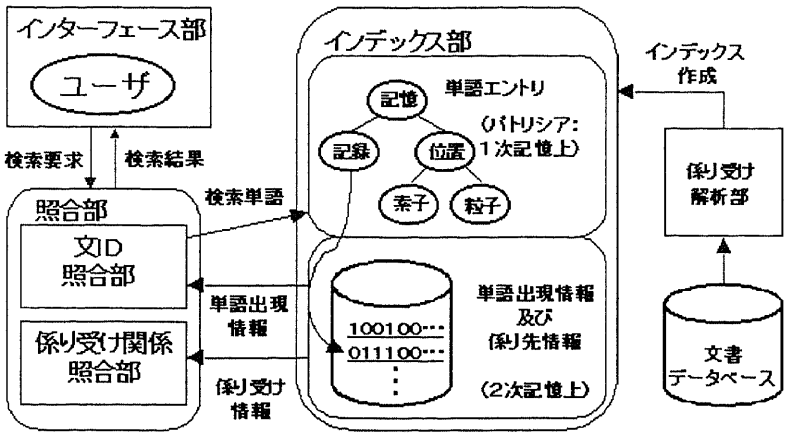

図1：システムの概要

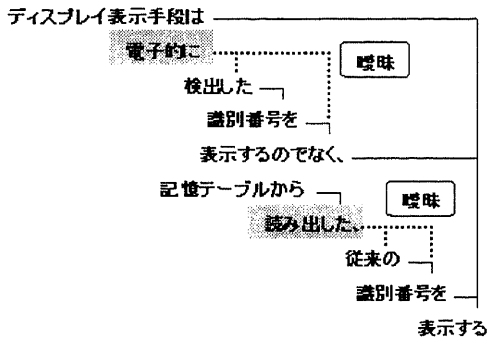

図2：解析例

単語出現情報は文ＩＤと，文内での単語位置情報から構成されている。文ＩＤは文書番号，文書内の項目番号，項目内での文番号から成り、[泓田９７]の手法による階層化ビットベクトル用いて実装した。また，単語位置情報と係り先情報は，それぞれ文ＩＤごとに単語番号列として登録する。単語番号は文内でのインデックス対象語を一意に表した番号である。

２．４ 照合部

検索は２段階に分けて行う。ユーザが，係り受け関係を含む検索パターンを入力すると，まず始めに，１次記憶上の単語エントリを検索し，２次記憶上の文ＩＤおよび単語位置情報と係り先情報を読み込む。そして，読み込まれた文ＩＤベクトル（階層化ベクトル）の論理積を実行し，指定した単語が出現する文を検索する（１次検索)。次に検索されたすべての文を対象として，検索パターンと係り受け構造が一致するか否かの照合を行う（２次検索)。係り受け構造の照合は，単語位置情報と係り先情報をビットベクトルに展開し論理積を実行することで行う。これにより，１つの単語が複数の位置に出現している場合や，係り先が特定出来ず複数の解析結果が得られている場合でも高速に照合が可能である。(図４参照)

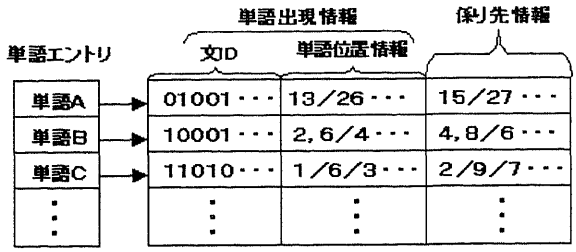

図3：インデックス

2．5 インタフェース部

インターフェース部はＪＡＶＡを用いて構築しており，Ｗｅｂ上で使用できる。検索単語の入力，係り受け関係の指定は，ＧＵＩ上で容易に行うことができる。図５にインターフェース画面を示す。

3．検索実験

係り受け情報を用いた検索の有効性を実証するため，ブール検索，近接関係を用いた検索との比較を行った。検索対象には表１に示す公開特許公報の一部を用い，サーバには，ＳＰＡＲＣ Ｓｔａｔｉｏｎ２０（ＣＰＵ：ＳｕｐｅｒＳＰＡＲＣⅡ，７５ＭＨｚ，メモリ：６４ＭＢｙｔｅ，ＯＳ：ＳｕｎＯＳ４．１．４）を使用した。

3．1 インデックス容量，検索実行時間

係り受け検索と近接関係を用いた検索及びブール検索との間で検索実行速度，インデックス容量の比較を行った。結果を表２に示す。

検索速度は，係り受け検索で１件あたり約２１．４５（ＭＳ）を要し，近接関係を用いた検索より高速で，ブール検索と比べても検索実行時間の増加を約１２％に抑えることができた。インデックス容量はブール検索で使用するインデックスの約１．６７倍，近接関係を用いた検索で使用するインデックスの約１．２７倍の増加となった。

3．2 係り受け検索の絞り込み精度

本検索システムで用いた骨格構造解析では，意味情報を利用しないと正確に係り受け解析できない部分は，曖昧なまま係り先を特定しないため，検索の際に正しく絞り込みができない可能性がある。図６に誤った絞り込みを行った例を示す。また，誤って解析した場合には，検索洩れを生じる場合もある。係り受け検索の絞り込み精度を評価するため，「[メモリ] が [記憶する] に係る文」のような１０件の検索要求に対し適合率と再現率を求めたところ，適合率９２．１１％，再現率９６．０１％という結果を得た。

3．3 近接関係を用いた検索との絞り込み精度比較

３．２で述べた同じ検索要求１０件に対し，近接関係を用いた検索との絞り込み精度比較を行った。表３に係り受け検索と近接距離を１～５まで変化させた時の適合率・再現率を示す。

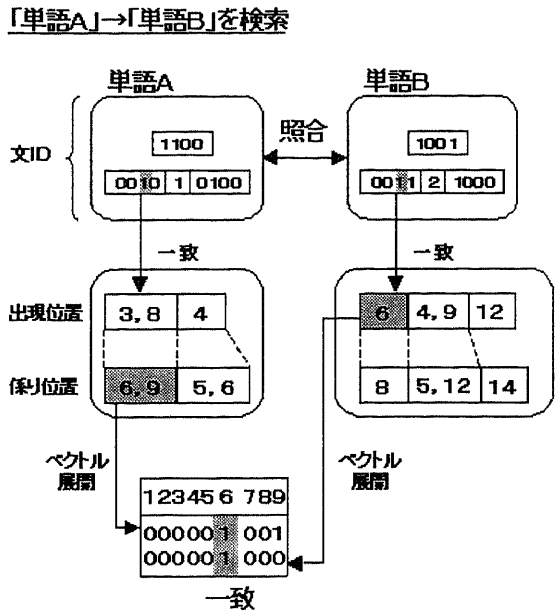

図4：係り受け関係の照合

近接関係を用いた検索では近接距離が短い時は適合率が良いが再現率が低い。図７の例では、近接距離１での検索はＡしか検索出来ない。又，近接距離を３にすると、実際には係り受け関係にないＣまで検索してしまう。さらに，ＤやＥを検索するため近接距離を長くとると，適合率が低下してしまう。このことから係り受け検索が有用であることが分かる。

４．おわりに

本論文では，係り受け情報を用いることによる高精度な全文検索システムについて述べた。特許公報に対する，係り受け情報の検索精度については、適合率９２．１１%，再現率９６．０１%であり、近接演算を用いた検索より良好な結果を得た。又、インデックス容量の増加は近接関係を用いた検索でのインデックスの約２７%に収まった。

参考文献

［兵藤９５］ 兵藤安昭, 池田尚志：表層的情報とＮ近傍ブロック化手法による日本語長文の骨格構造解析, 情報処理学会論文誌, Ｖｏｌ．３６, Ｎｏ．９, ｐｐ２０９１－２１０１（１９９５）

［兵藤９６］ 兵藤安昭, 河田実成, 應江黔, 池田尚志：構文つきコーパスの作成と類似用例検索システムへの応用, 自然言語処理, Ｖｏｌ３, Ｎｏ．２, ｐｐ７３－８８（１９９６）

［泓田９７］ 泓田 正雄, 溝渕 昭二, 獅々堀 正幹, 青江 順一：大規模文書データに対する用例文の効率的検索アルゴリズム，情報処理学会論文誌, Ｖｏｌ．３８, Ｎｏ．１０, ｐｐ２００４－２０１３（１９９７）

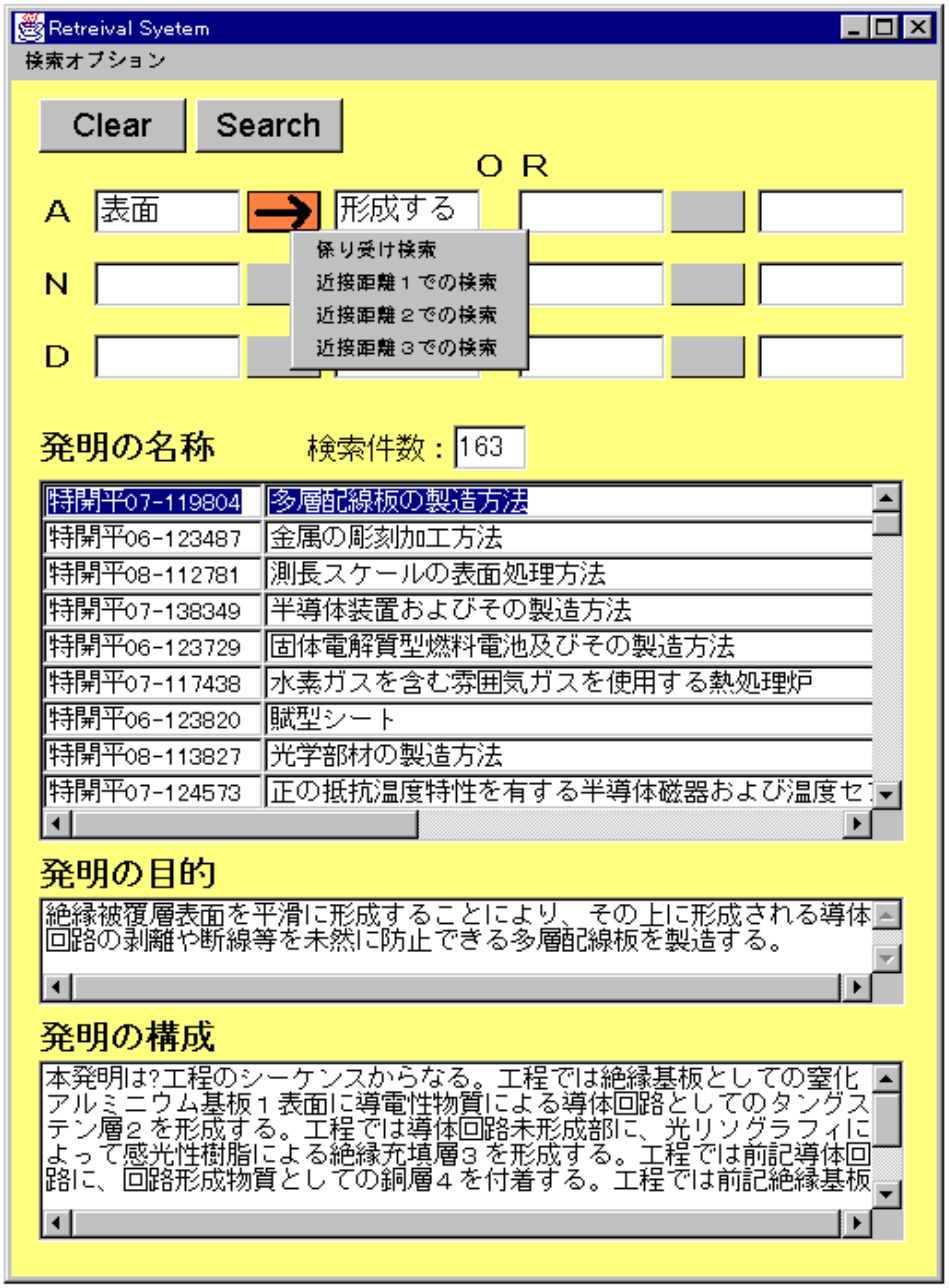

図5：インターフェース画面

対象	公開特許・実用新案中の2項目（「目的」、「構成」）
総特許数	19，870件 （総バイト数：9，021Kbyte）
総文数	51，070文 （1特許当たりの文数：2．57文）
総単語数	2，034，265個 （1文当たりの単語数：39．83個）
総単語エントリ数	867，496個
単語エントリの出現回数	8．75回／エントリ（MAX：「こと」，15，789回）

表1：実験に使用した特許文書のデータ

検索名	検索実行速度	インデックス容量
ブール検索	19. 10(MS)	3, 937(Kbyte)
近接関係を用いた検索	22. 56(MS)	5, 182(Kbyte)
係り受け検索	21. 45(MS)	6, 594(Kbyte)

表2：検索実行時間とインデックス容量の比較

「指示軸」→「備える」の検索

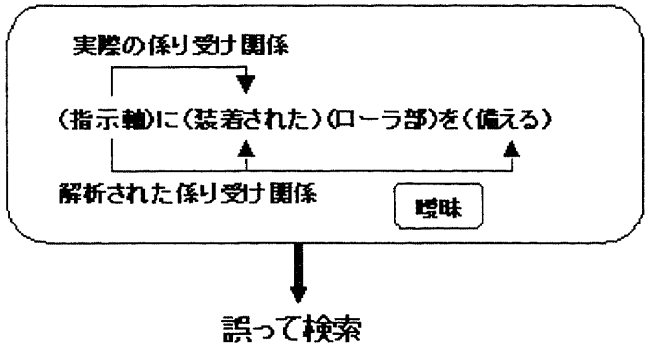

図6：検索誤りの例

検索名	適合率	再現率
係り受け検索	92. 11(%)	96. 01(%)
近接距離1での検索	96. 12(%)	74. 25(%)
近接距離2での検索	84. 06(%)	81. 55(%)
近接距離3での検索	75. 92(%)	87. 28(%)
近接距離4での検索	69. 94(%)	90. 52(%)
近接距離5での検索	65. 56(%)	93. 51(%)

表3：適合率・再現率の比較

検索要求：「表面→形成する」

	検索例文	近接距離2	近接距離3	係り受け
A	(樣材 表面)に(形成される)(皮膜)である	○	○	○
B	(ガラス 表面)に(多数)の(清状 凹部)が(形成された)(ガラス板)	×	○	○
C	(ZnSe 基盤)の(表面)を(酸化し)て(静電潜像)が(形成される)	○	×	○
D	(化粧材 表面)に(釘頭状)の(模様)を(簡単に)(安価に)(形成する)	×	×	○
E	(表面)には、(接着剤層)を(構成する)に(先立っ)て(印刷 インキ)で (木目 模様)を(印刷し)た(印刷 層)が(形成され)て(いる)	×	×	○

図7：検索例