係り受け関係を用いる高精度全文検索

4N－10

新美和彦　兵藤安昭　池田尚志
岐阜大学工学部

1　はじめに

近年、多くの情報が発信され、大量の文書が電子化されている。しかし、その大量の文書から利用者の要求する文書を検索する技術は、まだ十分であるとは言えない。特に現在の大規模テキスト検索においては単語あるいは文字列のブール検索が主流で、単語間の修飾関係はあまり考慮されていない。しかし、検索における精度向上には語間関係の利用は効果的であると思われる。

そこで、本研究では語間関係として、『ある単語がある単語にかかる』という係り受け情報を利用した文書検索システムを構築した。そして、係り受け検索を用いたときの効果や影響について考察した。

2　システムの概要

本システムは従来の単語検索の他に、係り受け関係を指定した検索を可能とした。これにより検索語間に意味的繋がりを持たないものを排除し、検索結果を絞り込むことができる。検索例を図3に示す。

システムはインデックスとして単語インデックスと単語出現情報データベースを持つ。

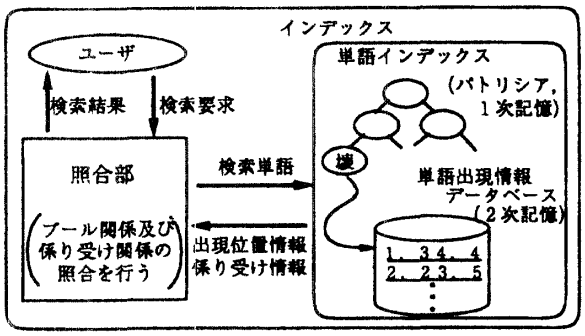

図１：検索システムの概要

High Precision Full-Text Search by Using Dependency Relations in a Sentence. Kazuhiko Niimi, Yasuaki Hyodo, Takashi Ikeda
Faculty of Engineering, Gifu University
Gifu-shi,501-11,Japan

これらのデータは文書集合に対して [1] の方法で構文解析を行ない、自立語のみをインデキシングの対象としたものである。

単語インデックスはメモリ上にパトリシアを使って構成し、高速な検索を可能にしている。単語出現情報データベースは単語の出現位置とその係り先の単語の位置の情報等を記述している。　検索結果照合部ではブール関係と係り受け関係の照合を行う。マッチした文に関しては文書ごとに集計し、ランク付けして検索結果を出力する。

3　単語出現情報データベース

単語出現情報データベースは以下のデータから構成される。

1. 単語の出現した文数
2. 単語の出現した文ID
3. 文内での単語の出現位置
4. (単語の出現位置に対する) 係り先の相対位置

文IDはどの文書の何番目の文であるかを表現している。データ3,4は1つの出現情報としてまとめて出力される (以後、係り受け情報と呼ぶ)。単語出現情報データベース作成の際にはこれらのデータを出現した全ての自立語について求め、単語毎にまとめあげる。そして、出現した文の数、文ID、係り受け情報の順に書き出す。この際、文IDは昇順にした上で、前の文IDとの差分情報にて表現している。さらに、単語毎に文IDの表示ビット数を変え、少ないビット数で表現できるようにしてある。係り受け情報は文IDの順に単語の出現位置、係り先の相対位置を表現している。

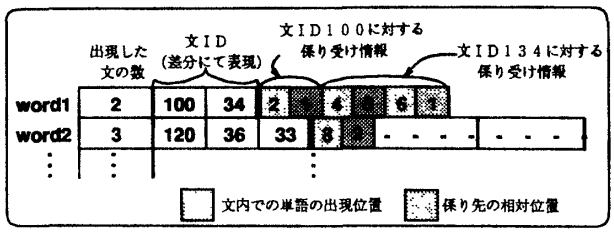

図２：単語出現情報データベースの内部

検索要求：" 会う" と" 来る"
－－－－－－－－検索結果 －－－－－－－－
（係り受けを使用しない検索)
だれか会いにきたら留守だといってくれ。
もう少し早く来てさえいれば彼に会えたのだが。
⇓
(係り受けを使用した検索)
だれか会いにきたら留守だといってくれ。

検索要求：" きみ" と" 行く"
－－－－－－－－検索結果 －－－－－－－－
（係り受けを使用しない検索)
きみの足じゃ、そこまで行くには日暮れになる。
きみは何度アメリカへ行きましたか。
⇓
(係り受けを使用した検索)
きみは何度アメリカへ行きましたか。

図３：検索結果の例

4 検索実験

4.1 実験方法

以上で作成されたシステムに関して以下の実験を行った。

1. 構文解析されたデータより係り受け情報を持つインデックスと持たないインデックスを作成し、その作成時間とインデックス容量の増加率を比較する。

2. 係り受け関係にある２単語の組を任意に 300 組抜きだし、係り受け検索 (検索１) と係り受けを使用しない検索 (検索２) の間での検索実行時間と検索された文の比較を行う。

実験には講談社和英辞典およびオーム科学技術大辞典の用例 97,336 文 (総単語数：949,636 個, 異なり単語数：58,441 種) を構文解析システム [1][2] にて構文解析したものを使用した。そして、このうちインデキシングされた単語は 58,272 種類、556,954 個となった。

尚、以下の実験はすべて SPARC Station20 互換機 (CPU：SuperSPARC-II 75MHz, メモリ：64M,OS：SunOS4.1.4) にて行ったものである。

4.2 実験結果

1. 上記のデータから 100 文,1000 文,10000 文, 全文 (97336 文) を抜きだしてインデキシングした時のインデックス作成時間・容量の増加率を調べた。結果を表１に示す。

文の数	作成時間の増加率	容量の増加率
100 文	0 %	16.95 %
1000 文	2.38 %	19.39 %
10000 文	7.94 %	42.43 %
97336 文	6.53 %	40.86 %

表１：増加率の比較

作成時間は７％前後の増加、容量は 40 ％くらいの増加率であった。

2. 検索実験の結果は表２のようになった。尚、結果に関しては一件当たりの平均検索時間、平均検索件数を表示している。

項目	検索１	検索２
平均検索件数	12.71 件	18.19 件
平均検索時間	48.9(ms)	40.5(ms)

表２：１件当たりの平均の検索結果

検索件数は係り受け検索では係り受けを使わない検索の 69.9 ％に絞り込めた。尚、検索実行時間の増加は 21 ％程度であった。

5 終わりに

係り受け情報の照合を加えた結果、検索結果を約７割に絞り込む事ができた。その際の検索時間は 1.2 倍程度で済むことが確認された。

今回は文を対象に行ったが、今後は特許等の大規模な文書集合を対象とした上で、インデックス容量や検索時間の増加を調べ、係り受けを使用した検索の有効性を検討していく予定である。

参考文献

[1] 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,Vol36,No.9pp2091-2101(1995)

[2] 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)