文書検索システム *Inforium* における自然言語処理ベースの支援機能

亀田雅之　小川泰嗣　松田透
(株)リコー・研究開発本部・情報通信研究所
{kameda,ogawa,matsuda}@ic.rdc.ricoh.co.jp

1　はじめに

大規模文書データベースの検索システムのニーズの高まりとともに、任意の文字列を検索できる高速全文検索システムが注目されている。しかし、文書検索機能単体では、ユーザの視点から、次のようないくつかの問題点が見出される。(1)検索条件式は、検索文字列を論理演算子形式で指定するが、演算子の使用や単語選択が難しい。(2)検索文書のタイトル一覧からさらに文書を選択したり、絞り込む支援、(3)検索された多くの文書の内容を素早く把む支援、といった機能が乏しい。

これらの問題に対し、既に、(1)自然言語による検索条件の指定[1]を許したり、(2)検索条件との類似度に応じたランキング[2]や検索文書群の関連[3]、(3)自動的に生成した抄録[4]を提示する、といった機能が提案、装備されるようになってきた。こうした機能は、(特に、日本語の場合は、)言語処理がベースとなるが、全文検索の高速性や検索精度を損ったり、大規模辞書の保守等が必要になるという問題がある。

我々は、これまで、1文字及び2文字の文字成分表による高速全文検索法[5]とともに、独自の日本語処理手法[軽量・高速な日本語解析系QJP[7, 8]とその上の日本語文書読解支援機能QJR[9, 10]]を提案・研究開発してきた。

今回、全文検索の高速性や検索精度を大きく損なわず、また、辞書保守を必要とせずに、日本語処理をベースとした支援機能を広範に備えた文書検索システムのプロトタイプ *Inforium*[11]を開発した。*Inforium* では、(1)日本語文による検索条件指定、(2)検索結果表示ウィンドウでのランキングと重要キーワード提示、(3)文書閲覧ウィンドウで重要文・キー文節・キーワード等の強調表示を行なう読解支援の機能群を搭載した。

2　*Inforium* の概要

Inforium の全体構成を図1に示す。

Inforium の中枢は、文字成分表を用いてBoolean形式の文字列検索条件を満足する文書を高速に検索する全文検索エンジンであり、さらに、これに、検索文字列ごとの文書頻度と文書内頻度に基づいた検索スコアによる文書ランキング系[6]を加えた[1]。

ユーザI/F系には、検索文解釈系、検索結果表示系、文書閲覧表示系があり、そのベースには、日本語解析系QJP、日本語文書読解支援系QJRがある。

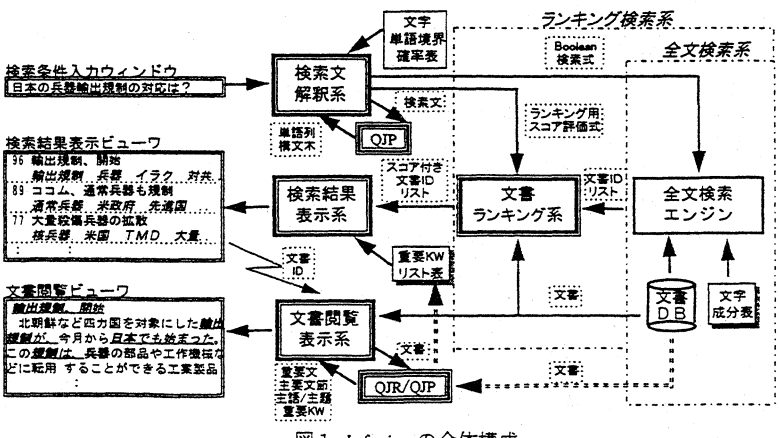

図1. *Inforium* の全体構成

[0] NLP-based support functions in a document retrieval system *Inforium*; KAMEDA Masayuki, OGAWA Yasushi, MATSUDA Toru

[1] 全文検索及び文書ランキング法については、本稿では省略する

3 QJPとQJR

まず、ベースとした日本語処理－簡易日本語解析系QJPと日本語文書読解支援系QJR－について述べる。

3.1 簡易日本語解析系QJP

QJP[7, 8]は、日本語文の字種の使い分けに着目し、機能語主体の小規模辞書による形態素解析と、意味情報に依らない構文レベルのif-then規則による係り受け解析により、軽量性[辞書約50KB, 実行メモリ260KB]、高速性[約500語/秒(Pentium133MHz)]、頑強性[新語登録不要, 長文解析]を実現した日本語解析系である[2]。

こうした性質から、応用系への負担が小さく、未知語登録のような辞書の保守も原則として不要である。

3.2 日本語文書読解支援系QJR

QJR[9, 10]は、QJPの形態素/構文解析結果を利用した、次のような日本語文書の読解支援機能群である。

Screening支援：文書中でのキーワード候補単語の強調表示、あるいは、重要キーワードのリスト表示により、直感的な文書のふるい分けを支援する。

キーワード候補単語(擬似キーワード)は、QJPの形態素解析結果を基に機能性名詞等を除いた名詞から抽出する[9]。さらに、修正単語頻度と予測構成単語数と呼ぶ構成単語に着目した指標により重要度付けを行なう。ただし、QJPが複合名詞を分割しないことから、構成単語の認定は、キーワード候補単語同士の文字列の重複により代替的に扱う(擬似キーワード相関法[10])。

Skip Reading支援：文書中での重要文の強調表示により、文書の飛ばし読み/拾い読みを支援する。

文については、キーワード候補の重要度と同時に、文ごとのキーワード候補群同士の構成単語の重複を計数して得た文間関連度[3]等の指標に基づいて、重要度を付与する。重要度が上位の文が重要文と判定される[10]。

Skimming支援：文書中での各文の骨格になる文節群を強調表示し、文書の斜め読み/速読を支援する。

この文節は、文の係り受け構造を基に、「～が～を～し、～が～を～した」のヘッドとなる文節[主要文節]や、文脈上の新情報を担う主節中の「が」句と主題化の「は」句と対応する用言文節を抽出する[9]。

Analytic Reading支援：長文の構造を適切に構造化表示することにより、分析的な読みを支援する。

構造化は、QJPの係り受け構造を基に、主要文節や一定の深さの文節に括弧付きのノード縮退等を施す[9]。

[2]ただし、原則として複合語を分割しない。

[3]関連文等も得られる

4 自然言語処理ベースの支援機能

本節では、ユーザI/F系の検索文解釈系、検索結果表示系、文書閲覧表示系と各系に組み込まれた自然言語処理ベースの支援機能について示す。

4.1 検索文解釈系

検索条件は、普通、Boolean形式や簡易的な表形式で指定するが、*Inforium*では、日常用いている日本語自然言語文を受け付けるので、ユーザは、複雑な条件式の指定を避けることができる。

入力された日本語文検索条件は、QJPにより解析される。形態素解析された単語のうち、機能性名詞等を除いた名詞やサ変名詞、形容動詞語幹等を抽出する。

複合語は、QJPが原則として分割を行なわないため、文字単位の単語頭及び単語末の確率表に基づく単語境界確率により分割する。即ち、複合語の各文字間の境界確率を前方文字の単語末確率と後方文字の単語頭確率の積として求め、その閾値により分割する[4]。例えば、「平(.018)和(.140)維(.047)持(.227)活(.029)動」(括弧内が境界確率)は、閾値が0.1の場合、「平和|維持|活動」と切断する。

抽出・分割された単語(検索文字列)群は、原則として、単純結合(+)の式に変換されるが、QJPの係り受け構造で、特殊な並列詞「かつ」や「または」等の並列が検出されていると、連言的(AND)結合や選言的(OR)結合になる。この式に基づき、文書ランキング系がスコア計算を行なう[5]。一方、全文検索エンジンには、すべてOR結合にしたBoolean式が渡される。

◇検索条件文

「高速かつ低価格の文書検索システムは？」

↓ QJP［形態素解析, 構文解析］＋複合語分割

◇解析＆複合語分割の結果

高速(名詞) かつ(並列詞)
低価格(名詞) の(格助詞)
文書|検索|システム(名詞) は(係助詞) ?(句点)

↓ 検索条件/評価式変換

◇検索条件/評価式

・文書ランキング用スコア評価式
　+(AND(高速, 低価格),+(文書, 検索, システム))
・全文検索エンジン用Boolean式
　OR(高速, 低価格, 文書, 検索, システム)

図2. 検索条件文の処理

[4]分割の閾値は、「ノイズなく」(適合率優先)～「もれなく」(再現率優先)のユーザ指定に応じて変化する

[5]現状では、どの結合も同等に扱い、検索文字列ごとの文書頻度と文書内頻度に基づくスコアの総和を文書スコアとしている

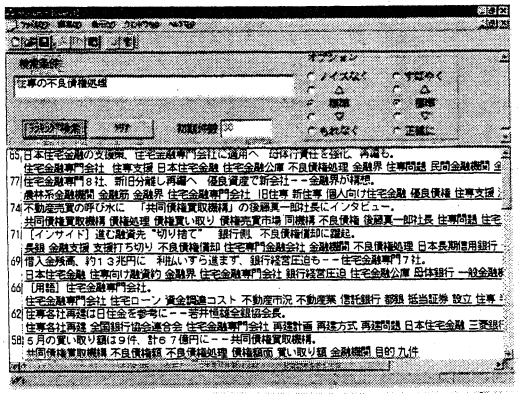

図3.検索結果表示例

4.2 検索結果表示系

全文検索の検索結果を受けて、文書ランキング系が、検索文書ごとにスコアを与えてランキングする。

検索結果表示系では、ランキング順にスコア付きで文書の一覧を表示すると同時に、文書登録時にあらかじめ文書ごとにQJP/QJRにより抽出しておいた**重要キーワードのリスト**(下線部)を並べて表示する(図3)。

スコアと重要キーワードリストは、文書を閲覧する前にユーザが文書への閲覧の要不要を判断するための有用な手がかり情報となる。

4.3 文書閲覧表示系

検索結果表示中に示された文書を指定すると、その文書の閲覧表示が行われる。この表示では、検索文字列が強調表示される他に、読解支援機能として、**重要文、主要文節、主語・主題**(及び対応する用言文節)、**重要語句**(重要キーワード)の4種類の要素を検索語(下線)とともに強調表示することができる(図4～図6)。

これらは、文書指定時に即時に指定文書のQJP/QJR処理が行われ、文書中の各文の係り受け構造からの主要文節、主語・主題(「は/が」句)と対応用言文節の抽出[9]、擬似キーワード相関法による文とキーワードの重要度付け[10]が行われる。

ユーザは、これらの支援機能を適宜切り替え、Screening, Skip Reading, Skimming等の読みの支援として利用できる。また、文とキーワードの重要度に基づき、強調する割合を5段階に指定できる。

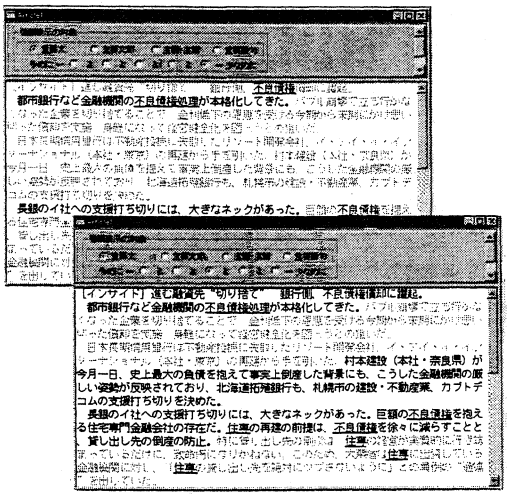

図4.文書閲覧表示例-重要文["少なめに"と"ふつう"]

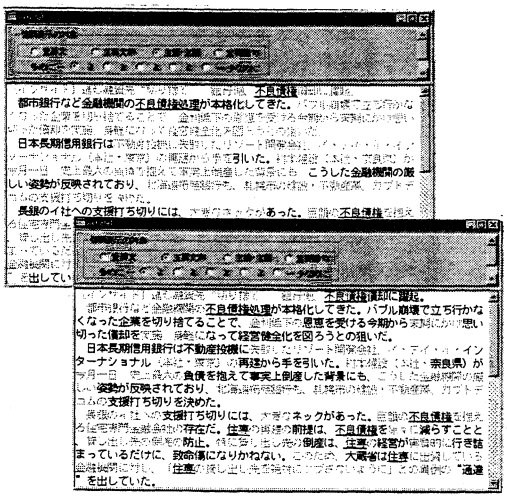

図5.文書閲覧表示例－主語・主題と主要文節

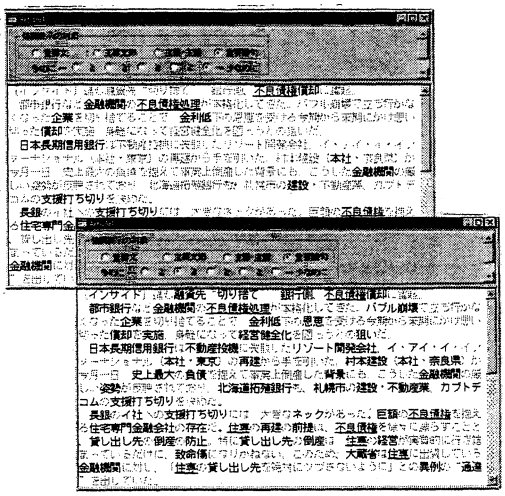

図6.文書閲覧表示例-重要語句["少なめに"と"多めに"]

5 評価

inforium 全体としての評価は行なっていないが、次のような要素ごとの評価実験を行なった。

ユーザ待ち時間

検索は、1年分約10万件の新聞記事(約100MB)に対し、Pentium/133MHzのPCで1秒程度で終了する[6]。この時間には検索条件式の解析/解釈時間も含まれるが、10数語程度の文ならば、ほとんど無視できる。

文書閲覧表示では、文書をその場でQJP/QJR処理するが、通常の新聞記事程度なら1秒前後で表示する。

ユーザの待ち時間は、特に、問題にはならない。

検索文解釈とランキング

1単語から5文節の20検索文について、547件の新聞記事に対する検索実験を行なった。*Inforium* の再現率/適合率優先(4.1節 脚注4参照)での検索とともに、検索文そのまま及び人間が同義語展開まで記述したBoolean式による検索での上位10文書の比較結果を表1に示す[7]。

シソーラスを用いていないことを考慮すると、十分実用に近いレベルである。

検索方法	再現率	適合率	検索件数
検索文そのまま	0.212	0.365	1.55
Inforium：再現率優先	0.918	0.495	9.90
Inforium：適合率優先	0.742	0.547	8.00
人間の検索式	0.958	0.856	6.00

表1. ランキング検索の評価結果

重要キーワード

キーワードの重要度については、177件の新聞記事ごとに人間が記事内から選択した上位10キーワードに対し、単語頻度、修正単語頻度、予測構成単語数、及びQJR方式(修正単語頻度と予測構成単語数)に基づく方法とを比較し、再現率/適合率が、各々、48.8/42.2%、53.1/48.6%、54.7/46.8%、57.5/52.0%と、QJR方式が優れていることを確認した[9]。

読解支援

4つの支援について、各々官能評価実験を行い、目的に応じた効果を確認し、改良[8]等を図った[9]。

ただし、新聞記事では、省略表現等のため、文節抽出の精度が十分でない問題がある。

[6]ランキング処理のため、全文検索処理よりはるかに時間がかかるが、逐次確定方式により、待ち時間の効率化を図っている[6]

[7]人間が文書に付与したキーワードでの検索文書を正解とした

[8]例えば、多くの被験者からのコメントにより、「は/が」句(主題・主語)に対する用言句も強調するようにした

6 まとめと今後の展開

全文検索システムに自然言語処理ベースのいくつかのユーザフレンドリな支援機能を付加した。即ち、(1)初心者には分かり難いBoolean検索式の代わりとして、日本語文での検索条件の入力を可能にした。(2)検索結果は、検索文書の単なる一覧でなく、文書の検索スコアと文書中に現れる重要キーワードリストを付与して、文書選択や絞り込みを支援する。さらに、(3)文書の閲覧に際し、文書中の様々なキー要素を強調表示する読解支援機能を搭載した。

ベースとしたQJP/QJRの特徴－[軽量性]システムに負担をかけず、[高速性]検索条件文や文書の処理時間も小さい、[頑強性]大規模文書DBで問題となる未知語登録等の辞書の保守が不要である－が、文書検索システムへの応用によく適合したといえる[9]。

文書検索システムとしての将来の支援機能の充実[10]以外に、本システムの当面の延長上では、検索結果表示での最重要文提示、重要文による抄録提示、強調表示法の改善等を検討している。

謝辞 本プロトタイプの実験は、毎日新聞CD-ROM(1993年分)の新聞記事を検索対象にした(図4～図6)。毎日新聞社には、研究用に使用させていただいたことを感謝する。

参考文献

[1] 石川：フルテキスト・データ検索機能の検討, デジタル図書館, No.3, pp.35-41, 1995.

[2] G.Salton, M.J.McGill：Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

[3] R.Rao, et al.:Rich Interaction in the Digital Library, Comm. of the ACM, Vol.38,No.4, pp.29-39, 1995.

[4] 住田 他：電子図書館のための効率的な文書検索, デジタル図書館, No.3, pp.35-41, 1995.

[5] 岩崎, 小川：文字成分表による文字列検索の実現と評価, 情処学会 研究会報告 DBS97, pp.1-10, 1993.

[6] 小川：文字成分表を用いた効率的文書ランキング法の提案, ADBS'95, pp.29-38, 1995.

[7] 亀田：軽量・高速な日本語解析ツール『簡易日本語解析系Q_JP』, 言語処理学会 第1回年次大会, 1995.

[8] Kameda：A portable & Quick Japanse Parser：QJP, COLING'96, pp.616-621, 1996.

[9] 亀田：日本語文書読解支援系ＱＪＲの検討, 情処学会 自然言語処理研究会報告 110-9, 1995.

[10] 亀田：擬似キーワード相関法による重要キーワードと重要文の抽出, 言語処理学会 第2回年次大会, 1996.

[11] Ogawa, Kameda, Matsuda： Inforium： A user-friendly document retrieval system, IROL96, 1996.

[9]QJPが複合語を分割しない問題に対しては、文字レベルの確率表や擬似キーワード相関法により、辞書作成なしで対処した

[10]現在、注目されているさまざまな可視化機能等