自然言語処理 115－18
情報学基礎 43－10
（1996. 9. 13）

フルテキストと抽出キーワードを利用した情報検索

木谷 強† 高木 徹† 木原 誠‡ 関根 道隆‡

†ＮＴＴデータ通信株式会社 情報科学研究所
‡日本経済新聞社 データバンク局

情報検索システム評価用ベンチマーク BMIR-J1 を使用し、抽出キーワードを検索対象とする従来のキーワード検索と、全文を検索対象とするフルテキスト検索との検索精度を比較した。検索式は検索者の経験と背景知識によって異なることを考慮し、検索式を計算機による自動生成、および筆者と専門家による作成という３種類の方法で作成した。評価の結果、キーワードと本文の両方を検索対象とした場合に最も高い検索精度が得られた。専門家は論理条件を複雑に組み合わせて検索式を作成し、シソーラスを利用して高い再現率を得ていた。また、BMIR-J1 の検索要求文と検索精度の関係を分析し、知識を必要とする検索要求ほど検索精度が低いことを確認した。

Information Retrieval Using a Full-text and Extracted Keywords

Tsuyoshi Kitani† Toru Takaki† Makoto Kihara‡ Michitaka Sekine‡

†Laboratory for Information Technology, NTT Data Corporation
‡Databank Bureau, NIHON KEIZAI SHIMBUN, INC.

Using BMIR-J1, a benchmark for evaluating information retrieval systems, a retrieval accuracy is compared when texts are retrieved from extracted keywords and/or from a full-text. Since queries may differ depending on user's experiences and background knowledge, queries created by computer, the author, and an expert are tested. The highest accuracy is achieved when both the keywords and full-text are used for the retrieval. The expert generates complicated queries using boolean operators and thesaurus words to improve recall of the retrieval. The retrieval accuracy is found to be poor to the queries requiring some knowledge to understand them.

1 はじめに

特許、論文、新聞記事などの情報検索サービスが広く一般に利用されている。これらのサービスは、テキストの全文を検索対象とするものではなく、あらかじめ抽出しておいたキーワードを検索対象とするものが多い。このようなキーワード検索では、重要語であるキーワードが検索対象となることから、検索結果における適合率（出力中の正解テキストの割合）は比較的高いと考えられる。しかし、検索対象は本文の一部であるため、再現率（全正解に対する出力された正解の割合）が低くなる傾向がある。再現率を高めるために、表記のゆれ、同義語、シソーラス語を検索語に追加して検索することが多い。

一方、近年になって急速にテキストが電子化されており、TREC に代表されるようなフルテキストを検索対象とした研究が盛んに実施されている [1]。フルテキスト検索はキーワード検索と異なり、あらかじめキーワードを抽出しておく処理が不要となる利点がある。しかし、重要でない部分も含むテキスト全体が検索対象となることから、再現率は高いが適合率は低くなる傾向がある。適合率を高めるため、テキスト中の検索語の出現頻度と出現位置などを利用しテキストの重要度を判断するスコアリング処理の研究がなされている [2]。これは、検索語と一致する度合（スコア）が高い順にテキストを出力し、上位の出力結果における適合率を向上させるものである。

今後はフルテキストの電子化が進み、抽出済みのキーワードとフルテキストの両方を利用できる環境が整ってくる。本稿では、検索システムに対するシステム利用者の経験と背景知識の違いを考慮し、検索式のレベルを 3 段階に分け、キーワード検索とフルテキスト検索の検索精度を比較する。また、日経シソーラスを使用したシソーラス展開および検索結果に対するスコアリング処理の有効性について評価する。評価用ベンチマークとしては、情報検索システム評価用ベンチマーク BMIR-J1 を使用する。BMIR-J1 の検索要求文は 6 つの機能に分類されており、機能分類と検索精度との関係についても考察する。

2 実験環境

本章では、使用したベンチマークの概要と、検索エンジンの特性を考慮した新聞記事の登録フォーマットについて述べる。

2.1 情報検索ベンチマーク BMIR-J1

本評価に使用する BMIR-J1 は、日本経済新聞の 600 記事と、それに対する 60 件の検索要求文、および期待される正解集合からなる[1]。検索要求文は短いフレーズで記述されており、検索のために必要な主要機能にもとづき、以下の 6 グループに分類されている [3]。

- **基本機能** (10 検索要求)：検索語およびシソーラス展開した語の論理式 (AND, OR 条件など) で検索可能な検索要求
  (例)「国内航空大手 3 社」
- **数値・レンジ機能** (5 検索要求)：数値の大小比較や単位の理解・変換により検索可能な検索要求
  (例)「1 ドル＝105 円以上の円高」
- **構文解析機能** (6 検索要求)：動作およびその主体・対象が記述されている検索要求
  (例)「コンピュータメーカーの人員削減」
- **言語知識利用機能** (12 検索要求)：言語の深い意味や文脈を理解して可能となる検索要求
  (例)「株価動向」
- **世界知識利用機能** (10 検索要求)：主に常識的な判断や、蓄積された事実からの推論により検索可能となる検索要求
  (例)「多角化事業の低迷」
- **言語知識利用と知識処理機能** (17 検索要求)：言語知識と世界知識の両方により検索可能となる検索要求
  (例)「現地調達」

正解集合は記事番号で示され、正解レベルとして A ランクと B ランクがある。A ランクは、検索要求の内容が主題として記述されているテキストに付与されている。B ランクは、検索要求の内容は記述されているが主題ではないものに付与されている。今回の評価では、両者を区別することなく正解として扱った。

2.2 記事の登録フォーマットと検索処理

本評価では、市販のフルテキスト検索パッケージを検索エンジンとして使用する。本文のほかキーワードを含む任意のフィールドがフルテキスト検索の対象となる。使用する検索エンジンは、文字列を単語ではなく文字の並びで照合する。したがって、検索もれは発生しない反面、過剰検索（検索ノイズ）が発生することがある。たとえば、検索語「色」は「赤色」に照合

[1] 株式会社日本経済新聞社の協力によって、社団法人情報処理学会・データベースシステム研究会・情報検索システム評価用データベース構築ワーキンググループが、1993 年 9 月 1 日から 12 月 31 日の日本経済新聞記事を基に構築した情報検索用データベース（テスト版）を利用。

するが、検索語「京都」は「東京都」にも照合してしまう。記事に付与されたキーワードに対する過剰検索の可能性を低減するため、各キーワードの前後にASCII文字のカンマを付け、キーワード間の区切りを明確にする。図1はBMIR-J1の新聞記事データを登録フォーマットに整形したものである。検索時には検索対象フィールドを識別するSGMLタグを指定することにより、任意のフィールドに対する検索が可能となる。

```
<NEWS>
  <TITLE> 日銀、公定歩合引き上げ </TITLE>
  <DATE>1995/2/22</DATE>
  <BODY> 日銀総裁は 21 日、...</BODY>
  <KEYWORD>, 単語 1, 単語 2, 単語 3,...</KEYWORD>
</NEWS>
```

図1 新聞記事の登録フォーマット

3 評価実験

本章では、検索者の経験や知識の違いを反映した3種類の検索式の作成方法、検索対象とする新聞記事フィールドの選択、および検索結果に対するスコアリング処理アルゴリズムを説明する。

3.1 検索式の作成

我々は検索式を組み立てる際、過去の検索システムの使用経験や検索対象分野の背景知識を用いて、検索語を選択し検索語間の論理条件を決定している。このため経験や知識によって組み立てる検索式が異なり、検索精度も変化すると考えられる。本評価では、検索システムが現実に使用される様々な状況を想定し、検索要求文から以下の3通りの方法で検索式を作成し、検索精度を評価する。

1. プログラムにより機械的に作成した検索式
2. 一般的な利用者が作成した検索式
3. 日経キーワードに精通した専門家が作成した検索式

上記の分類は、検索システムの利用経験および知識に応じて、それぞれ初心者レベル、中級（一般）レベル、専門家レベルを想定している。

初めに、プログラムにより機械的に検索式を作成する方法を説明する。まず、検索要求文を形態素解析プログラムMajesty [4]により単語に分割し、名詞、サ変名詞および未知語（解析に失敗した単語）を取り出す。検索式は抽出した単語をAND条件で結合して作成する[2]。検索要求文「業績悪化を原因とする企業の合併の事例」は、次のような検索式に変換される。

検索式(1)：

```
"業績" AND "悪化" AND "原因" AND
"企業" AND "合併" AND "事例"
```

次に、一般的な利用者が作成する検索式を考える。検索システムの一般的な利用者は、思いついた数語の単語を簡単な論理条件で組み合わせて検索することが多い。本評価では、検索要求文から検索に有効と考えられる単語を筆者が選択した[3]。検索語の選択基準として、総称的な単語（「企業」、「メーカー」、「業界」など）や抽象的な単語（「原因」、「対策」、「影響」など）を含めないこととした。これは、総称的な単語は多くのテキストで使用される可能性が高く、テキストの弁別力が弱いためである。また、抽象的な単語は、テキスト中では別の具体的な表現となることが多く、それ自身がテキスト中に出現することが少ないためである。この方針で作成した上記の検索要求文に対する検索式は次のようになる。

検索式(2)：

```
"業績悪化" AND "合併"
```

第3の検索式として、日経新聞記事のキーワードを日常の業務で保守している専門家は、上記の検索要求文から以下の検索式を作成した。

検索式(3)：

```
"企業合併%" AND
("業績悪化" OR "経営不振" OR
 "計上損益%" OR "欠損%")
```

この検索式から明らかなように、専門家は検索要求文に出現しない単語を検索式に追加している。検索語に付加された記号(%)は、その単語がシソーラス展開可能なエントリであることを示している。

上記の3通りの検索式に対して、日経シソーラスを使用して検索語の下位語をシソーラス展開する場合と、シソーラス展開しない場合の検索精度を測定する。シソーラス展開は、検索語とシソーラスのエントリが完全に一致したときのみ実施し、展開される単語はOR条件で検索式に追加する。日経シソーラスには主に一般語と業界コードによる体系があるが、今回は一般語のみを使用した。表1は3通りの検索式について、60個の検索要求文から作成した検索式に対し、含まれる

[2]検索対象フィールドをキーワードと本文とし、全てOR条件で検索式を作成した場合、再現率77%、適合率9%と適合率が極端に低かったため、OR条件での検索式は評価対象としなかった。

[3]筆者はBMIR-J1の作成に従事していたため、背景知識としては一般レベルではない。

	計算機	筆者	専門家
検索単語数	179	102	158
OR 演算子数	0	4	73
AND 演算子数	125	38	33
シソーラス展開語数	12	12	43

表 1: 60 個の検索式における検索語数と演算子数

検索語数、OR 演算子と AND 演算子の数、シソーラス展開の対象となる検索語数を示す。表 1から、専門家はシソーラス展開可能な検索単語を検索語として加え、AND 演算子と OR 演算子を積極的に組み合わせて検索式を作成していることがわかる。専門家が使用した検索単語のうち、検索要求文に現れない単語は、検索単語全体の約 6 割にあたる 94 単語であった。

3.2 検索対象フィールドの選択

検索対象フィールドによる検索精度を違いを評価するため、以下の 3 フィールドを検索対象とする。

- 日経キーワード (KW)
  BMIR-J1 の新聞記事には、日経新聞社が付与したキーワードが 1 記事あたり平均 42.2 語含まれている。本文に出現しないキーワードは、1 記事あたり平均 3 語登録されている。
- 本文 (BD)
  記事見出しと記事本文。BMIR-J1 の新聞記事は、1 記事あたり平均 733 文字を含んでいる。
- 日経キーワードと記事本文 (KW+BD)
  日経キーワードと記事見出しおよび記事本文。

3.3 スコアリング処理

本評価で使用するスコアリング処理は単語の頻度情報を利用する TF ／ IDF 法に基づくものであり、TREC で高い検索精度が確認されている SMART システムのアルゴリズムを利用している [6]。検索式とテキストの適合度合はベクトル空間モデルにより、検索式の単語ベクトルと、比較するテキストの単語ベクトルの角度から求める [5]。

TF/IDF 法によると、あるデータベース db において、単語 t に対するテキスト d の重要度 $W(d,t)$ は、次の式で求めることができる。

$$W(d,t) = TF(d,t) \times IDF(t) \qquad (1)$$

ここで、$tf(d,t)$ をテキスト d における単語 t の出現頻度、$length(d)$ をテキスト d の長さ（文字数）とすると、$TF(d,t)$ は次の式で求めることができる。

$$TF(d,t) = \frac{1.0 + \log(tf(d,t))}{\log(length(d))} \qquad (2)$$

また、$DBsize(db)$ を db 内の総テキスト数、$freq(t,db)$ を db 内で単語 t が出現するテキスト数とすると、$IDF(t)$ は次の式で求めることができる。

$$IDF(t) = \log \frac{DBsize(db)}{freq(t,db)} \qquad (3)$$

次に、検索式とテキストの適合度合を計算するため、データベース db 内に存在する全ての単語からなる m 次元ベクトルを考える。あるテキスト d の単語ベクトルは、(1) 式から求めた単語の重要度を要素とするベクトルとして表現できる。検索式の単語ベクトルは、たとえば、検索語の重みを 1、その他の単語の重みを 0 とする m 次元ベクトルとして作成する。適合度合（スコア）は、両者のベクトルの成す角度を内積計算によって求める。

評価実験では上記のスコアリングアルゴリズムを使用し、本章で述べた 3 通りの検索式に対してスコアリング処理の有効性を評価する。

3.4 評価指標

評価指標として、情報検索で一般的に使用されている再現率と適合率を採用する。それぞれの定義は次の通りである。

$$\text{再現率} = \frac{\text{出力中の正解数}}{\text{全正解数}} * 100 \qquad (4)$$

$$\text{適合率} = \frac{\text{出力中の正解数}}{\text{出力数}} * 100 \qquad (5)$$

また、総合的な処理精度を表す評価指標として、情報抽出の分野で知られている F 値 (F-measure) を使用する [7]。F 値は再現率 (R) と適合率 (P) から、次式で求めることができる。

$$F = \frac{(\beta^2 + 1.0) * P * R}{\beta^2 * P + R} \qquad (6)$$

ここで、β は適合率に対する再現率の相対的な重みである。本評価では、再現率と適合率の重要度を同等と考え、$\beta = 1.0$ とした。

4 実験結果と考察

本章では、実験から得られたデータを分析し、実験結果と BMIR-J1 の機能分類との関係を考察する。

	KW	BD	KW+BD
計算機	10/53 (17)	20/41 (27)	22/41 (29)
計算機 (シソーラス)	11/53 (18)	21/41 (28)	22/40 (28)
筆者	22/46 (30)	32/44 (37)	35/43 (38)
筆者 (シソーラス)	22/41 (29)	32/43 (37)	35/40 (37)
専門家	31/37 (34)	37/36 (36)	46/33 (39)
専門家 (シソーラス)	36/35 (36)	41/35 (38)	52/32 (40)

表 2: 再現率 / 適合率による検索精度 (括弧内は F 値)

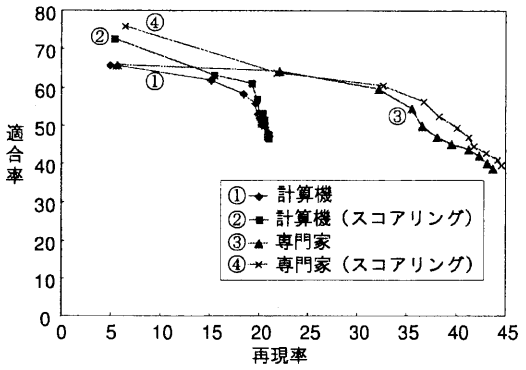

図 1: スコアリング処理の効果

4.1 検索式と検索対象による精度

作成した 3 通りの検索式を使用し、3 つの検索対象フィールドに対して検索を実行した結果、表 2に示す適合率と再現率を得た。表中の括弧内の数値は F 値である。

まず、検索者別にみると、検索精度は専門家、筆者、計算機の順に高く、検索システムの利用経験と背景知識の違いが現れた。キーワードと本文 (KW+BD) を検索対象としシソーラス展開をしない場合、F 値では筆者と専門家の検索式は、それぞれ 38 と 39 で近い精度を得ている。内訳を見ると、筆者は適合率が高く専門家は再現率が高いという特徴がある。一般的に、検索語を絞り込むことによって、比較的容易に適合率を上げることができる。一方、再現率を高くしかつ適合率の低下を一定のレベルに保つことは難しい。これは、本実験で専門家が作成した検索式から明らかなように、再現率を上げるためには検索要求に明示的に出現しない検索語を検索式に追加することが必要であり、幅広い知識が要求されるためである。

次に、検索対象フィールドによる検索精度の違いを考察する。適合率が最も高いのは、筆者によるシソーラス利用の結果を除いて、キーワード (KW) を検索対象とした場合である。しかし、本文 (BD) を検索対象とした場合と比較すると、筆者と専門家では、適合率の低下度合よりも再現率の向上度合が大きい。筆者や専門家のように検索条件を工夫できる場合、キーワードよりも本文を対象とした方が再現率の面からメリットが大きいことを示している。また、キーワードと本文 (KW+BD) を検索対象とした場合、本文 (BD) のみを検索対象とした場合に比べ、特に専門家において 9 ～ 11 もの再現率の向上がみられる。検索式を作成した専門家は、実際に日経キーワードの保守業務をしており、登録キーワードについて熟知していることが再現率向上の理由と考えられる[4]。登録キーワードのうち、本文に存在しないものは 1 記事あたり平均 3 語であるが、専門家はそれらを有効に利用しているといえる。F 値から判断すると、本ベンチマークに関する限り、どの検索式に対しても、キーワードと本文を検索対象とする効果が認められる。

シソーラス展開の効果は、計算機と筆者が作成した検索式に関しては認められない。これは、表 1に示すとおり、シソーラス展開の対象となった単語数が少なかったためと考えられる。専門家の場合、どの検索対象フィールドにおいても F 値で 1 ～ 2 のシソーラス展開の効果が確認できた。

4.2 スコアリング結果

図 1は、計算機と専門家が作成した検索式について、スコアリング処理の有無による検索精度を比較したグラフである。スコアリング処理をしない場合は検索結果の順番に意味はないが、スコアリング処理の効果を比較するため出力の上位から検索精度を調べた。図 1は検索対象フィールドをキーワードと本文 (KW+BD) とし、シソーラス展開をしない場合の結果である。グラフから、スコアリング処理が特に再現率の低い部分 (再現率で 5 ～ 20, 出力順位で 1 ～ 5 位に相当) で適合率の向上が確認できる。図 1では筆者が作成した検索式の結果は省略したが、若干のスコアリング処理の

[4] 商用の日経キーワード検索サービスを使用して専門家が同様の検索式で検索した場合、再現率 41%、適合率 37% (F=39) であった。本実験でのキーワードフィールドを対象とした検索精度が低い (再現率 36%, 適合率 35%, F=36) のは、シソーラス展開した単語が一般語のみであったことと、全ての下位語を展開しなかったことが原因と考えられる。

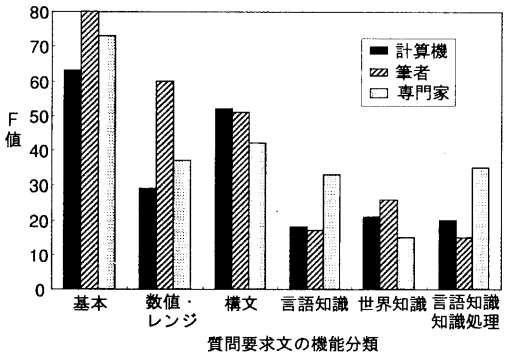

図 2: 検索要求機能別の F 値

効果が確認できた。

4.3 BMIR-J1 機能分類別の分析

図 2は、BMIR-J1 の機能別の F 値をグラフ化したものである。検索対象フィールドはキーワードと本文 (KW+BD) であり、シソーラス展開をしない場合の結果である。基本機能と構文解析機能の精度は比較的高いが、知識を必要とする機能の精度は総じて低い。全体的に見ると、検索要求の理解が困難であるほど検索精度は低くなっているといえる。検索式のレベル別に見ると、専門家は言語知識と言語知識+知識処理の機能で相対的に高い精度を出しており、専門家の知識が検索式に反映された結果と考えられる。数値・レンジ機能で筆者による検索式の精度が高い理由は、適合率を重視して検索語の数を少なくしたことが、数値の範囲制限の影響と偶然に一致したためと考えられる。日経シソーラスを利用した場合の機能分類と検索精度の向上度合については、明らかな相関関係は認められなかった。

5 おわりに

検索者の経験と背景知識の違いを考慮し、検索式を計算機による自動生成、および筆者と専門家による作成という 3種類の方法で作成し、情報検索システム評価用ベンチマーク BMIR-J1 を使用して検索精度を比較した。その結果、再現率を高めるためには、専門家が作成した検索式から明らかなように、シソーラス展開が可能な検索語を AND 条件と OR 条件を使用して積極的に組み合せることが有効であることが確認できた。また、スコアリング処理については、出力順位の上位において、その有効性が確認できた。

BMIR-J1 のテキストデータである日経新聞には、本文に出現しない単語がキーワードとして登録されている。実験から、キーワードと本文の両方を検索対象とした場合に、再現率は 3種類の検索式全てにおいて最も高く、再現率と適合率の総合評価 (F 値) も最も高かった。一方、適合率は検索対象をキーワードとした場合に最も高かったが、検索対象を本文としても筆者と専門家による適合率はほとんど低下せず、逆に再現率が増加するメリットが大きかった。

BMIR-J1 における検索要求文の機能別分類と検索精度の関係を分析した結果、検索要求が知識を必要とするほど検索精度が低くなることを確認した。知識処理にもとづく検索処理はこれからの課題であるが、数値・レンジ機能については、情報抽出コンテスト MUC-5 で実施されたように、テキストから該当する数値情報を抽出することで、検索精度の向上が期待できる [8]。BMIR-J1 を使用することで、本実験のような定量的な評価が実施できた。しかし、統計的に意味のある結果を得るためには大規模なベンチマークが必要であり、本格版のベンチマークに期待したい。

参考文献

[1] Harman, D. editor. "The 3rd Text Retrieval Conference (TREC-3)." National Institute of Standards and Technology, 1994.

[2] 高木 徹, 木谷 強. "単語出現共起関係を用いた文書重要度付与の検討." 情報処理学会情報学基礎研究会, 96-FI-41-8, pp. 61-68, 1996.

[3] 芥子 育雄ほか. "情報検索システム評価用ベンチマーク Ver1.0(BMIR-J1) について." 情報処理学会研究会報告, 96-DBS-106, pp. 139-146, 1996.

[4] 木谷 強. "固有名詞の特定機能を有する形態素解析処理" 情報処理学会研究報告, 92-NL-90, pp. 73-80, 1992.

[5] Salton, G. and McGill, M. J. "Introduction to Modern Information Retrieval." McGrawHill, New York, 1993.

[6] Buckley, C., Salton, G., Allan, J. and Singhal, A. "Automatic Query Expansion Using SMART." Proceedings of The 3rd Text Retrieval Conference (TREC-3), pp. 69-80, 1995.

[7] "Fourth Message Understanding Conference (MUC-4)." Morgan Kaufmann, San Meteo, 1992.

[8] "Fifth Message Understanding Conference (MUC-5)." Morgan Kaufmann, San Meteo, 1993.