3V－1

キー概念に基づく情報検索方式の高度化(3)
—キーワードに基づく方式との比較—

藤崎 博也　大野 澄雄　阿部 賢司　鈴木 匡芳　飯島岐勇　片見 憲次
東京理科大学

1. はじめに

キーワードを利用する通常の情報検索システムでは、キーワードの異表記同義・同表記異義の存在により、検索もれ・不要な検索が生じる。筆者らはこれらを軽減し、かつユーザの検索要求に適合する文書をより多く抽出するため、キー概念[1]に基づく情報検索システムを提案している[2]～[4]。

本報では、このシステムの基本的な考え方を説明し、キーワードを利用する通常の方法と比較するために行った小規模なシミュレーションの実験結果について述べる。

2. 異表記同義・同表記異義への対処

着目するキーワード(Kw_0)が異表記同義をもつ場合の対処の方法は、(1)Kw_0の概念Cを表記/概念対応辞書に基づいて取り出す。(2)概念Cに基づいてKw_0と同義語の関係にあるキーワード$Kw_1 \sim Kw_m$を取り出す。この様子を図1(a)に示す。(3)Kw_0と$Kw_1 \sim Kw_m$を用いて以下の検索式を構成する。

$$Kw_0 + Kw_1 + \cdots + Kw_m \qquad \text{(I)}$$

ただしここで'+'は論理和を意味する。この式を用いることにより、検索もれを軽減することができる。

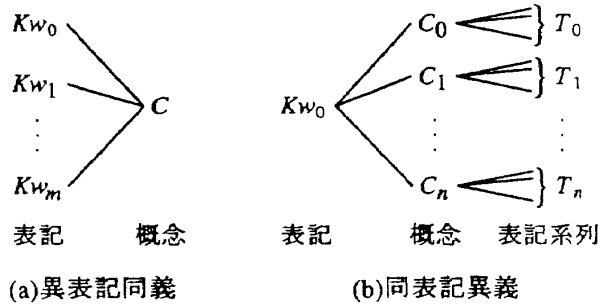


図1. キー概念による処理

一方、Kw_0が同表記異義をもつ場合の対処の方法は、(1)Kw_0のもつ複数の概念$C_0 \sim C_n$を表記/概念対応辞書に基づいて取り出す。ここでC_0は着目する概念を表すものとする。(2)概念$C_1 \sim C_n$に対応するキーワードの集合$T_1 \sim T_n$を表記/概念対応辞書に基づいて取り出す。この様子を図1(b)に示す。(3)C_0以外の概念$C_1 \sim C_n$にそれぞれ対応するキーワードの集合$T_1 \sim T_n$を検索の対象から除外する。この場合の検索式は次の式(II)で表される。

$$T_0 \cdot \overline{T_1} \cdot \overline{T_2} \cdot \cdots \cdot \overline{T_n} \qquad \text{(II)}$$

ただしここで'・'は論理積、'$\overline{\ }$'は否定を意味する。またT_0は、さきの式(I)で表される。

3. キー概念を用いた情報検索システムの概要

筆者らはキー概念に基づく情報検索システムの構築を進めており、その概略を図2に示す。このシステムは、5つの要素から構成されており、キー概念による検索を実現する上で、表記/概念対応辞書とキー概念による検索式生成は重要な役割を果たしている。

インターフェースの機能は、(1)ユーザからのキーワード、概念の選択の入力を受け付ける。(2)検索エンジンから検索結果を受け取り、画面に出力する。表記/概念対応辞書の特徴は、表記と概念両方のアクセスが可能な点である。キー概念による検索式生成の機能は、(1)着目しているキーワードの概念を表記/概念対応辞書に基づいて取り出す。(2)この段階でキーワードの概念が複数存在する場合、ユーザが選択した概念をインターフェースから受け取り、概念を特定す

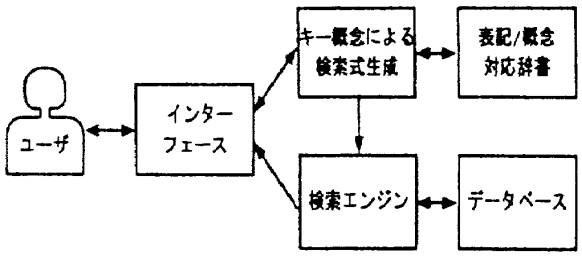


図2. キー概念を用いた情報検索システムの概略図

An advanced information retrieval system based on key concepts(3)
— Comparsion with a keyword-based system —
Hiroya Fujisaki, Sumio Ohno, Kenji Abe, Masayoshi Suzuki, Michio Iijima, and Kenji Katami
Science University of Tokyo, 2641 Yamazaki, Noda, 278-8510

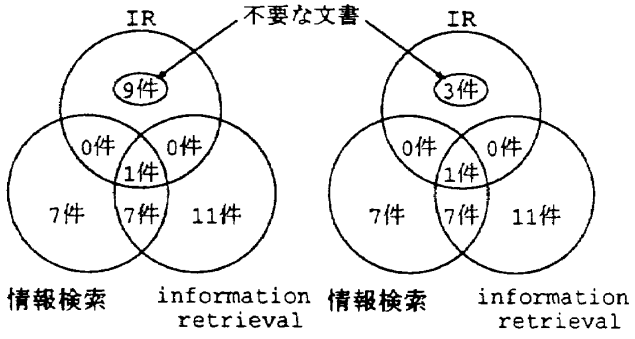
る。(3)キーワードの概念が特定された後、表記/概念対応辞書に基づいて、着目しているキーワードと同義の関係にあるキーワードを取り出す。(4)着目しているキーワードと同義の関係にあるキーワードから式(I)を用いて検索式を構成する。検索エンジンの機能は、(1)キー概念による検索式生成プログラムから検索式を受け取る。(2)データベースを検索する。(3)検索結果をインターフェースに送る。ここで行う実験では、データベースとして学術情報センター電子図書館サービスの1998年1月現在のものを使用した。

4. キーワードを用いる通常のシステムとの比較

キー概念を用いた本システムとキーワードを用いる通常のシステムの性能を比較するため、小規模な検索のシミュレーション実験を行った。

キー概念に基づく検索結果の一例として「情報検索」の略語として用いられる「IR」をあげる。(1)異表記同義に対処することによって、「IR」と異表記同義の関係にある「情報検索」と「information retrieval」に関する検索結果が得られ、抽出件数が10件から35件に増加した。これを図3(a)に示す。しかし「IR」から得られた検索結果には、不要な文書が9件含まれていた。(2)次に同表記異義に対処することによって、「IR」の不要な検索を減少させた結果を図3(b)に示す。上記の処理により、不要な文書は9件から3件に減少し、適合率は74.3%から89.7%に上昇した。

キー概念による検索と通常の検索とを比較するため、情報処理、人工知能、通信、半導体、電磁気、生物の6分野、60例に関する小規模な検索実験を行った。両者をキーワード比、抽出比、適合率によって比較評価したものを表1に示す。ここでキーワード比と



(a)異表記同義への対処　(b)同表記異義への対処

図3. キー概念検索方式の検索例「IR」について

表1 検索方式の差による評価結果

検索方法	キーワード比	抽出比	適合率(%)
通常	1.00	1.00	84.1
キー概念	3.07	2.00	85.4

は、通常の検索システムの場合のキーワード数を基準として、キー概念を用いたシステムの場合のキー概念により拡張された後のキーワードの個数を相対値として表したものである。抽出比および適合率は、それぞれ以下の式で与えられる。

$$\text{抽出比} = \text{抽出件数}/\text{通常の検索の抽出件数} \quad \text{(III)}$$

$$\text{適合率} = \text{適合件数}/\text{抽出件数} \quad \text{(IV)}$$

表1の結果は、キー概念を用いることにより、適合率には顕著な変化が見られなかったが、キーワード比は約3倍になり、抽出比は2倍になること、換言すれば、その結果、適合文書がほぼ2倍多く抽出できることを示している。

5. おわりに

検索もれ・不要な検索を軽減するため、キー概念を媒介としてキーワードの検索式を構成し、キーワードを用いる通常のシステムと性能を比較した結果、本システムの有効性が示された。同表記異義への対処に関しては、より有効な方法に関して検討を行っている。

参考文献

[1] 藤崎博也, 亀田弘之, 河井恒：“新聞記事情報の階層構造に基づく記事分類・検索システム,” 情報処理学会「自然言語処理」研究会資料 44 - 4 (1984).

[2] 藤崎博也, 亀田弘之, 大野澄雄, 阿部賢司, 伊東卓哉, 佐久間聖仁：“キー概念の抽出と未知語の処理に基づく情報検索方式の高度化,” 情報処理学会第54回全国大会講演論文集, vol. 3, pp. 23-24 (1997).

[3] H. Fujisaki, H. Kameda, S. Ohno, T. Ito, K. Tajima and K. Abe : “An intelligent system for information retrieval over the Internet through spoken dialogue,” *Proceedings of Eurospeech'97*, vol. 3, pp. 1675-1678 (1997).

[4] 藤崎博也, 亀田弘之, 大野澄雄, 阿部賢司, 劉軼, 戸井田和重, 八杉大輔：“キー概念に基づく情報検索方式の高度化 (1) −キーワードの異表記同義の処理−,” 情報処理学会第56回全国大会講演論文集, vol. 3, pp. 128-129 (1998).