データベースへの知的検索インターフェイスのための基礎実験

7F－5

丸田 浩二　黒岩 眞吾　浅見 徹
国際電信電話株式会社 上福岡研究所

1．はじめに

筆者らは、知識ベース技術および自然言語処理技術の一つの応用として、テキスト文書の知的検索システムの検討を行なっている。具体的には、知的検索インターフェイス技術をテキストデータのキーワード検索システムへの適用することを目標とする。

通常のキーワード検索システムでは、柔軟な検索を行なわせるためシソーラスが利用されることが多いが、検索システムで利用するシソーラスは必ずしも一般的な分類体系に基づくシソーラスは必要であるとは限らない。つまり、キーワード検索用という目的に特化したコンパクトなシソーラスが構築できれば、検索者の検索意図に合う柔軟な検索システムを効率的に開発できる。

そこで本稿では、この実験システム開発の手掛かりとするための基礎データの収集実験を行なったので報告する。

2．実験の目的

目的とする知的検索実験システムは、海外の国際通信に関する記事情報を検索対象としている。ここで知的検索とは、検索者の自然言語による検索要求に対して文章を解析しシソーラスおよび対象世界モデルを参照して得たキーワードで検索を行なうことを言う。例えば「ＫＤＤの専用線による電話サービスについて知りたい。」という要求文に対して「ルートＫＤＤ」（ＫＤＤの専用線による電話サービスの名称）というキーワードで検索する。

対象とする記事情報は、記事のID番号、見だし、本文、キーワード、その他の分類用の統制キーワードからなるテキスト文書であり、それぞれの検索キーは、汎用リレーショナルデータベースのフィールドとして登録する。またキーワードは、記事の作成者によって与えられた自由キーワードである。本システムは、このような文書検索の際に検索の漏れを以下の方式により減少させる。

（1）キーワード入力に対してはキーワード間のシソーラス情報、対象世界知識等を利用する。

（2）また自然言語による検索要求に対しては情報検索に特化した自然言語処理を行なう。

知的な検索を以上の方式で行なうことにより検索利用者の柔軟な検索支援を行なうことが目的である。本実験では、シソーラス等の構築の指針および自然言語インターフェイスにふさわしい自然言語処理の方法を探るため、このテキスト文書の検索要求に現われるキーワードおよび自然言語による検索要求文を収集した。

3．実験方法

検索対象となる記事は国際通信に関する専門的な内容であるが、本実験の被験者は専門知識を持たない者（８名）とした。専門知識を持たないものを被験者とすることで専門家の付加したキーワードとの差異を求める等、データベース作成者と検索者の概念のずれについての情報を収集する。

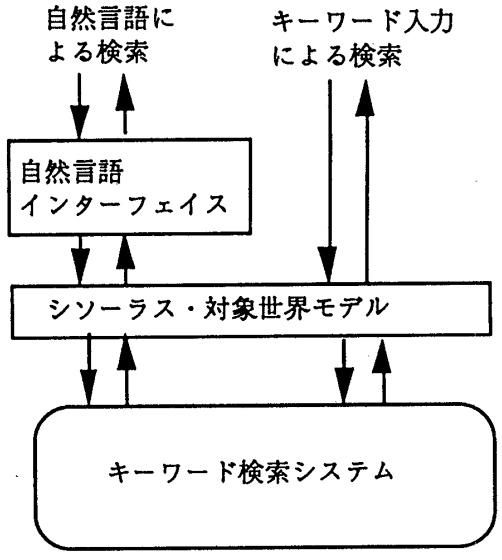

図１．インターフェイスの概念図

A Study of Intellectual Interface to Document Database
Koji Maruta, Shingo Karoiwa, Tohru Asami
KDD Kamifukuoka R&D Laboratories.

実験は、被験者に記事を事前に読ませたのちにキーワードおよび自然言語による検索要求文を紙に書かせた。

(1) 記事の例

一記事は、見出し、本文を含めて2～３０００文字程度のテキスト文である。記事の見出しの例を示す。

「欧米主要9か国における国際電話料金比較」、「ＡＴ＆Ｔ、大口顧客別の電話料金制度をＦＣＣに申請」

また、キーワードは、各記事に３～１０個程度与えられている。キーワードの例は「ＡＴ&Ｔ、ＡＯＳ、電話料金割引サービス、ＢＴ、国際ＶＡＮ、国際回線交換サービス・・」等である。

(2) 検索者キーワードの収集

キーワード作成者の付与するキーワード集合と検索者の入力するキーワード集合がずれている場合、再現率を上げるためにはこれらのずれをシソーラス情報等で吸収する必要がある。この情報を集めるために検索者の選択したキーワードを収集した。実験の方法は以下のとおりである。

・まず被験者に一つの記事を読ませて理解させる。わからない部分は、説明を与える。このとき、記事に付与されているキーワードは被験者には見せない。

・この記事について重要と思われるキーワードを7、8個程度選択させ紙に書き出させた。

・各被験者は、６つの記事に対して上記のことを行なわせた。

(3) 自然言語文による記事の検索要求

・この理解した記事を検索したいときの問い合わせ文を自然言語で書かせた。数の指定はせずに複数個書かせた。

・また、この同じ検索要求をキーワードの組み合わせで表現させた。

3．実験結果

(1) 検索者のキーワード

各被験者に6つの記事のキーワードを選択させた。被験者の選択したキーワード数は、述べ337個、一つの記事に対し平均7.0個である。これらのキーワードを、記事作成者のキーワードとの関係で、以下のように分類した。

完全一致：作成者のキーワードと完全に一致しているもの

下位：このキーワードの上位概念が作成者キーワードとしてあるもの

上位：このキーワードの下位概念が作成者キーワードとしてあるもの

類義語：同じ意味を持つキーワード

関係語：上位・下位・類義の関係の組み合わせをたどると作成者キーワードとしてあるもの

その他：上の関係以外のもの

シソーラスを用いない場合、完全に一致したもの以外のキーワードは検索時には無効なものとなってしまう。これらの分類結果の頻度グラフを図2に示す。

4．結果

被験者によるキーワードの選択は記事を見せて選択させたため、大部分のキーワードが、本文中に出現する言葉から選ばれた。しかし、通常の検索者は、漠然とした検索意図よりキーワードが選ばれるため、検索用キーワードとは異なる可能性が高い。また、キーワードの分類において、「その他」に含まれているものの中にはキーワードが結合した複合語的なものが多い。例えば、長距離電話料金割引制度、国内専用線料金など。このような複合語的なキーワードについては、シソーラスで表すと膨大な数になると予想される。この問題を克服するためにはキーワードを分解して意味を解析する必要がある。

5．おわりに

テキストデータベースへの知的検索インターフェイスの構築の指針を得るための実験について報告した。今後本実験結果を元にさらに詳細な実験を行なう必要がある。

［参考文献］

[1]拝原：”日本語文献データベースへの知的アクセス”、電子情報通信学会誌Vol.72 No.7 ,pp797-806(1989)

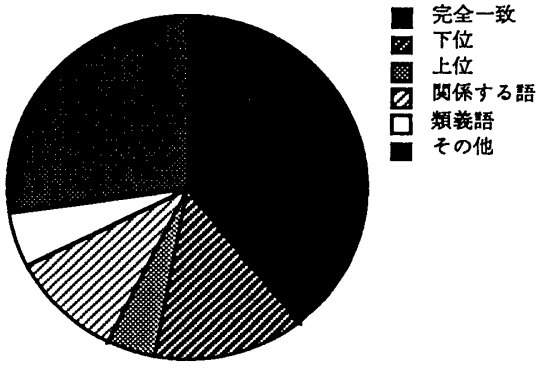

図２．収集したキーワードの分類