知的ファイリングモデルシステムの開発（その２）
4Y-9　－自由語検索における異表記、異表現解消法－

畠山　敦　藤澤浩道　藤縄雅章
（株）日立製作所　中央研究所

１．はじめに

文書をファイルに登録するとき、文書名や文書の作成者あるいは抄録等の書誌的事項を入力する必要がある。しかしながら書誌的事項として記述される単語のなかには同一の意味を表わすものでも微妙に異なった書き方をするものがある。これが検索の際に障害となって所望の文書が探し出せないという問題が発生する。単純な例では、“MOTOR”のことを“モータ”とカタカナで記述する場合がある。こういった書誌的事項をどのように記述するかという問題は個人差があって、文書の登録者と検索者が異なる人物であった場合に特に問題となる。検索者の思い付いた単語から、同一の概念を示す他の単語をも検索することをここでは自由語検索と呼ぶことにする。この検索方式は知的ファイリングシステム〔１〕の上で高度なマンマシンインタフェースを提供している。本稿では自由語検索を行うための一つの手法について述べる。

２．自由語検索の課題

概念を言語、ひいては文字列で表わす際の問題点を以下にまとめる。表１には、その例を示す。

（１）言語：同じ意味を示すのに、日本語、英語といった記述方法がある。
（２）同義語・略称：同一の意味を示す他の単語が存在する。
（３）表現：日本語の場合はカタカナ、ローマ字、アルファベット表現の外来語、漢字といった書き表し方がある。
（４）音節表記：日本語のカタカナ、ローマ字には音節単位で異なる書き表わし方がある。
（５）送り仮名表記：漢字の送り仮名には異なる書き表わし方がある。
（６）文字コード系：カタカナやアルファベットの文字コードは現在１バイトコード系と２バイトコード系が併用されている。

表１　問題点の具体的な例

表現レベル	例
言語	計算機、コンピュータ、Ｃｏｍｐｕｔｅｒ
同義語・略称	ワードプロセッサ、ワープロ
表現	ケイサンキ、ＫＥＩＳＡＮＫＩ、計算機
表記	カンジョウ、カンジヨオ ＫＡＮＺＹＯＵ、ＫＡＮＪＯＨ
送り仮名	読み書き、読書き
文字コード系	ﾓｼﾞｺｰﾄﾞ、モジコード

３．自由語検索方式の概要

自由語検索方式の概要を図１に示す。本方式は、検索者が入力した文字列に対し、それを計算機で処理して同一の概念を示す他の文字列を網羅的に自動発生して検索するところに特徴がある。図１では、"ｺﾝﾋﾟｭｰﾀ"と検索者が入力したときの例を示している。

計算機は入力された文字列に対し、他の表現や他の表記を自動的に発生して、検索すべき文字列の候補を複数個発生する。発生した複数個の文字列に対し、書誌事項ファイルに該当する文字列を探す。この検索方法としては、複数個の文字列の探索を１回のスキャンで実行する多重ストリングサーチアルゴリズム〔２〕を用いる。

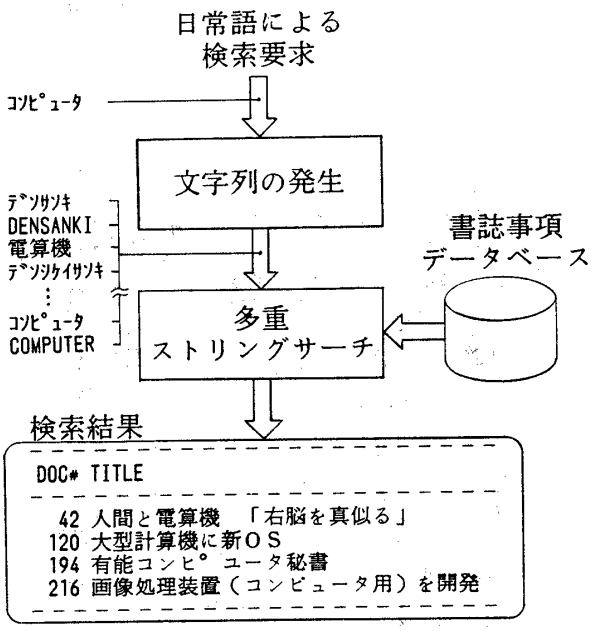

図１　自由語検索方式の概略

Development of a Knowledge-Based Filing Model System(2) Management of Irregular Expressions for Free Term Search
Atsushi HATAKEYAMA, Hiromichi FUJISAWA, Masaaki FUJINAWA
HITACHI,LTD.

4．同一の概念を示す文字列の発生

通常使用する文字列をコード系、言語、表現の違いによって分類することにより、同一の概念を示す文字列を発生することが可能となる。その分類により、図2の大きな楕円で示すように7種類の文字列の集合に分けられる。同一の概念を示す文字列を発生することは、どのような文字列が入力されても、図の7種類の文字列の集合に属する適切な文字列を抽出することに還元される。そしてそのためには図の矢印で示される文字列間を結ぶ変換プログラムが必要である。

発生の手順を説明する。計算機は、検索者が入力した一つの文字列に対し、それがどの集合に属するかをまず判定する。それは、入力した文字列がアルファベットかカタカナか、あるいは同義語辞書に外来語として登録されているか否か、というような区別で可能となる。次に表記の標準化を行う。標準化とは同義語辞書に登録されている標準とする文字列に変換することで、これにより同義語辞書を簡略化することができる。そしてその標準化された文字列に対し、図の実線で示される変換プログラムや、あるいは同義語抽出プログラムを使って他の言語や表現の文字列を発生する。さらに、発生された複数の文字列に対し、標準の文字列から表記の異なる他の文字列を発生する。このようにして、一つの概念に付随する文字列を網羅的に発生することができる。

今回開発したシステムでは同義語の問題に対処するために同義語辞書を用いた。さらに、表現の問題を解決するために、カナ、ローマ字変換プログラム及び仮名漢字変換プログラムを用意し、コード系の問題については、コード変換プログラムを用意した。また、図の点線の矢印で示される表記の標準化及び他の表記の発生には部分文字列の書き換え規則を使った。具体的には、部分文字列の書き換え規則から文字列を標準化したり、あるいは他表記を発生するオートマトンを生成し、これに文字列を入力することにより実行する（図3）。表記の問題のうち漢字仮名混じりの表現の送り仮名の問題については仮名漢字辞書により大部分をカバーできた。

5．おわりに

今迄、完全にキーワードを覚えていなければ探し出せなかった文書が探し出せる可能性が高くなった。現在は、新聞記事の切り抜きや、特許明細書等の文書データベースに適用して、有効性を確認しつつある。今後は検索者の指定する文字列以外の文字列を自動的に発生して検索することによる検索ノイズの程度を調べる必要がある。また、ユーザによる同義語辞書の拡張方法を考えていきたい。

6．謝辞

仮名漢字変換プログラムを提供していただいた当所上原徹三主任研究員に感謝いたします。

〔1〕藤澤他:"高度ファイリングの理念と要素技術",情処研報,日本語文書処理7-4(1986.7)

〔2〕Aho,A.V.and Corasick,M.J.:"Efficient String Matching",Communications of the ACM.Vol.18(1975)PP.333-340.

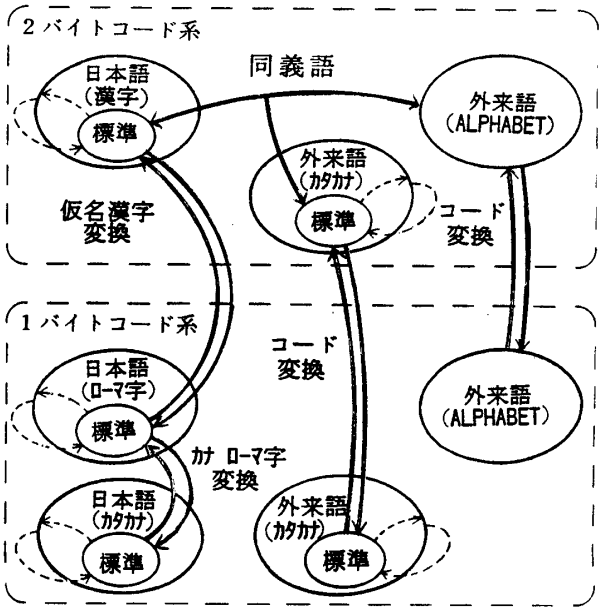

図2　文字列発生方法の概要

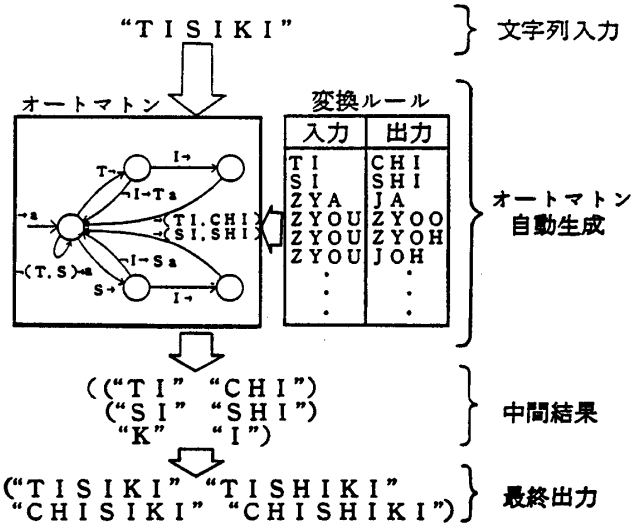

図3　異なる表記の発生方法