漢字の異形字表記に対応した検索システム

○阪口哲男、赤穂義範

A Full-text Retrieval System with a Function of Unifying Kanji Variants

○Tetsuo SAKAGUCHI, Yoshinori AKO

Abstract

The Japanese character set consists of Hiragana, Katakana, and Kanji. Some Kanji characters have the variants that have the same meaning and pronunciations. The users of full-text retrieval systems sometimes confuse some Kanji characters with the variants and get insufficient result for their needs. The authors considered that a function that unifies Kanji variants is need for full-text retrieval systems. This paper describes a full-text retrieval system that has functions of unifying Kanji variants. The system has a thesaurus of Japanese Kanji variants and uses it for indexing and unifying Kanji characters. The authors built two example retrieval systems based on the system. They have the database of ULIS-DL (http://lib.ulis.ac.jp/) metadata and Japan-MARC.

1. はじめに

インターネットの発展に伴い World Wide Web (WWW)のサーチエンジンなどの全文検索システムが普及している。このような全文検索システムではWWWページの文中に含まれる文字に基づいて索引付けを行うことが一般的である。そのため、利用者によって指定された文字と同じ文字が含まれているか否かが検索の際の基本的な条件となる。

日本語で用いられている漢字には、字形は異なるが元々の意味と読みが同じであり、同一と見なしても良いとされる組合せが存在する。例えば、「阪」と「坂」はその組合せの一つであり、「大阪」という地名は時代によっては「大坂」と表記されていた。このような組合せに含まれる一つの文字は同じ組に含まれる他の文字の「異形字」と呼ばれる[1]。このような異形字の関係にある文字は互いに字形も似ていることが多く、日常においても混同されることが多い。例えば、「阪口」という名字はしばしば「坂口」と取り違えられることがある。

このような異形字の存在は全文検索システムの利用にも影響を与える。例えば、Google (http://www.google.com/)において「阪口哲男」という条件で検索すると約51件の結果が得られるが、「坂口哲男」という条件では1件しか結果が得られず、しかもその内容は全く異なるということが生じる。従来からの検索システムには、索引付けの際に統制語による付与キーワードを与えることや、読みデータを付加することでこの問題に対応することが可能なものもある。しかしながら、それらの手法は多量のデータを扱う全文検索システムでは手間がかかり過ぎるため、何らかの自動化された手法が必要となる。

本研究では、このような異形字表記の問題に対応した検索システムの構築を行った。本システムでは異形字シソーラスを組み込み、データの索引付けや検索の際に参照する。本システムでは検索の洩れをより少なくすることを目的として、利用者が検索条件に指定した文字のみではなく、その異形字関係にある文字も含めた検索を行う。

2. 検索システムにおける異形字表記への対応

日本語の文章では表記に様々な揺れが生じることがある。例えば、表記に用いる字種（漢字、平仮名、片仮名）の違い、外来語の音訳の違い（「インタネット」と「インターネット」など)、送り仮名の違い（「行う」と「行なう」など）などがある。これらの問題に対処するため、例えば外来語の音訳などの片仮名の異表記に対応する研究などがこれまでも行われてきている[2]。本研究では漢字に関する異形字表記への対応手法を考察し、その機構を取り入れた検索システムを構築する。

JIS X0208 などの文字コード体系では基本的に字の形が違えば異なる文字コードが割り振られる。例えば前出の「阪」と「坂」の例の場合はそれぞれ区点コードで 2669 と 2668 が割り振られている。通常、検索システムにおける文字の比較は文字コードを用いているため、文字コードが異なるものは原則として異なる文字として扱われる。アルファベットの大文字と小文字のような場合には対応表による変換を行うなど、同一の文字と見なすための処理によって対応している。アルファベットのような場合は文字そのものも少なく、また機械的な処理が容易になるような文字コードの割り振り方がされていたので、広く対応されている。一方、漢字に関しては異形字関係の対応付けを行う何らかの辞書が必要となる。この辞書を異形字シソーラスと呼ぶ。

異形字シソーラスを検索システムで用いる場合には大きく分けて 2 つの手法が考えられる。一つは、異形字関係にある漢字のうち一つを代表字として、データを索引付けする際にその代表字に統制し、検索条件中の文字も代表字に変換してから検索を行う方法である。もう一つは、索引付けでは特に変換などは行わず、検索条件中の文字について、その異形字関係にある文字を組み合わせた条件を生成して検索するものである。前者は検索性能を低下させることなく異形字表記に対応できるが、異形字表記を含めずに検索することが出来なくなる。後者はいくつもの異形字による組合せを生成して複数の条件についてそれぞれ検索するため、検索性能は低下するが、どの異形字表記を検索条件として採用するかを利用者が選択することが可能になる。

3. 漢字の異形字表記に対応した検索システム

前節で述べたような考え方に基づき、検索システムの構築を行った。開発環境は以下の通りである。

- ワークステーション: Sun SPARCstation Ultra1 (CPU: 200MHz, RAM: 256MB)
- OS: Solaris2.5.1
- 開発言語: Java (Sun JDK1.1)
- 検索エンジン: OpenText Ver.5
- WWW サーバ: Netscape FastTrack Server Ver.2.0

検索エンジンに用いた OpenText は SGML に準拠した文書について索引付けし、高速に文字検索を行うことが出来る。本システムでは対象となるデータと異形字シソーラスの検索に OpenText を使用している。検索対象データと異形字シソーラスは SGML 形式のものを扱い、文書形式については DTD 記述を用いずに次のような事項を記述した設定ファイルを用いることで対応している。

・検索対象データ: 利用者へ提示する名称、利用者が指定するフィールドとするか否か、結果表示に含めるか否かについて個々のエレメント毎に記述する。

・異形字シソーラス: 見出し字のエレメント名、異形字集合のエレメント名、代表字のエレメント名、漢字情報表示に用いるエレメントとその名称を記述する。

設定ファイルに従って、検索対象データの索引ファイルと異形字シソーラスの索引ファイルを生成する。検索対象データの索引ファイルは 2 種類存在し、一方は、データに含まれている文字を異形字シソーラスに基づいて代表字へと統制したものであり、もう一方は統制していないものである。

検索の利用者インタフェースには一般的な WWW ブラウザを使用する。統制済み索引ファイルを用いた検索を異形字簡易検索と呼ぶ。もう一方の索引ファイルを用い、利用者がどの異形字の組合せを使用するかを選ぶことが可能な検索を異形字詳細検索と呼ぶ。異形字簡易検索の際は、利用者が入力した条件中の文字を代表字に変換し、統制済み索引ファイルを用いて検索を行う。異形字詳細検索の際は、利用者の入力中の文字に対応する異形字を組み合わせた候補を生成し、利用者に提示する。そして、利用者が選んだ候補のみを用いて統制されていない索引ファイルを用いて検索を行う。

4. 検索システム構築事例と分析結果

本システムの機能を確かめるために具体的なデータを用いて 2 種類の検索システムの構築を行った。一つは図書館情報大学ディジタル図書館(ULIS-DL, http://lib.ulis.ac.jp/)で作成・蓄積を行っているメタデータ(ULIS-DL メタデータ)である。ULIS-DL メタデータは図書館情報学とその関連領域に関するインターネット上の情報資源のメタデータである。もう一方は Japan-MARC である。ULIS-DL メタデータは SGML 形式であるので、若干の修正を加えたのみでほとんどそのままの形式で索引付けを行った。Japan-MARC は図書館情報大学総合情報処理センターで提供している変換ツールを用いて plain text 形式に変換したものから SGML 形式への変換を行った。ULIS-DL メタデータは約 2 万 4 千件、Japan-MARC は 1999 年分のうちの約 4 万 3 千件を用いた。異形字シソーラスには小熊らによる「あやのふひと」[1]を用いた。

構築した ULIS-DL メタデータ検索システムの画面を図 1 と図 2 に示す。図 1 は異形字簡易検索を行った結果の簡略表示画面、図 2 は異形字詳細検索時の異形字候補選択画面である。異形字候補を選ぶ際は、その文字コードなど異形字シソーラスに含まれている情報を見ることも可能である。

システム構築と同時に対象データに含まれる異形字の集計を行った。その結果、ULIS-DL、Japan-MARC 共に異形字を持つ漢字の割合は約 29%であり、代表字の割合は ULIS-DL が約 25%、Japan-MARC が約 21%であった。このことから、今回用いたデータでは異形字を持つ漢字の場合は代表字が多くを占めていることがわかる。つまり、利用者が入力した文字が代表字ではない場合の検索結果が少なくなるので、本システムの適用によってより多くの結果を得ることが可能になると考えられる。

5. おわりに

構築したシステムでは OpenText 検索エンジンとの通信にやや時間がかかるという問題が残るものの、異形字表記への対応機能を実現することが出来た。また、本システム構築を通じて次のような問題が残されていることがわかった。

(1) 読みは同じで日常生活では混同されるが意味や字の由来から別の文字とされる場合（例：斉藤の「斉」と「斎」）

(2) 一般的な環境における JIS X0212 補助漢字の表示

前者は使用した異形字シソーラスの内容の問題であるが、目的によって文字を同一として扱って良いかどうかが分かれる可能性も考えられる。後者については計算機環境の問題であり、Unicode などの普及に伴って文字フォント等の提供も進むことが期待される。なお、本システムでは補助漢字についてはビットイメージを用いて表示を可能にしたが、多言語 HTML ブラウザ(MHTML Browser, http://mhtml.ulis.ac.jp/)技術を利用することも考えられる。

参考文献

[1] 小熊善之, 永森光晴, 阪口哲男, 杉本重雄, 田畑孝一. 日本語漢字の異形字シソーラス. ディジタル図書館, No.17, p.37-45, 2000.

[2] 伍井啓恭, 清原良三, 鈴木克志, 太細孝. カタカナ異表記処理. 情報処理学会第 38 回(昭和 64 年前期)全国大会, p.351-352, 1989.

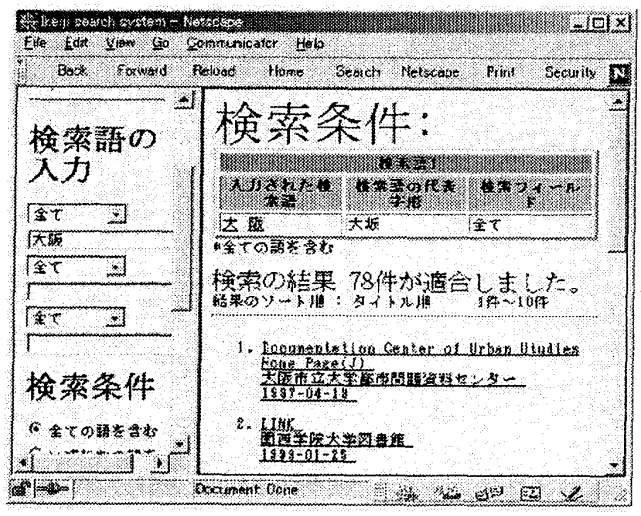

図 1　異形字簡易検索結果画面

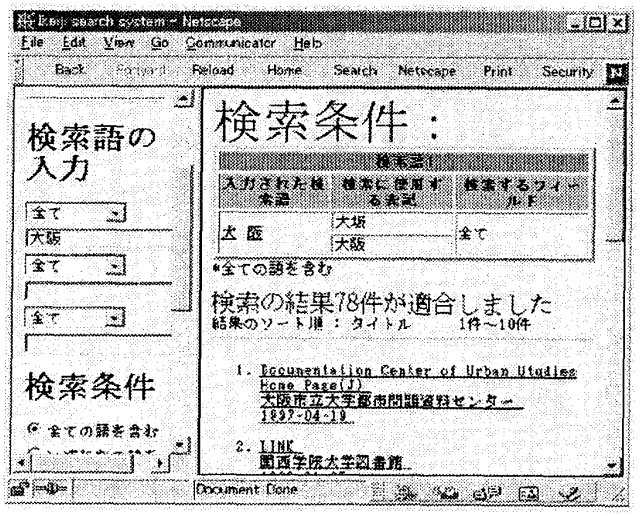

図 2　異形字詳細検索結果画面

阪口　哲男　図書館情報大学（〒305-8550 茨城県つくば市春日 1-2)

赤穂　義範　同上（2001 年 4 月より会社員）

Tetsuo SAKAGUCHI (saka@ulis.ac.jp) University of Library and Information Science

Yoshinori AKO　University of Library and Information Science