

科学技術文献検索システムにおける異表記対応について

甲田 彰

独立行政法人 科学技術振興機構 情報提供部

〒102-8666 千代田区四番町 5-32

Tel: 03-5214-7992, Fax: 03-5214-7514

概要

独立行政法人科学技術振興機構(JST)では、2006年4月から科学技術文献検索システム「JDreamII」の提供を開始した。JDreamIIで扱うのは約3,900万件という大量の科学技術文献であり、こうした文献のデータには、同じ単語でありながら表記が異なるため、単純に検索するとヒットしない「異表記語」が多数存在している。JDreamIIでは文献中に登場する異表記語に対し、漏れなく高速に検索するための仕組みを搭載したので、本稿でその概要を報告する。

キーワード

科学技術, 文献検索, 異表記語, JDreamII

Search method of variant notations on a science and technology document retrieval system

Akira KOUDA

Department of Service, Japan Science and Technology Agency(JST)

5-3, Yonbancho, Chiyoda-ku, Tokyo, 102-8666, JAPAN

Phone: +81-3-5214-7992, Fax: +81-3-5214-7514

Abstract

Japan Science and Technology Agency(JST) has released a new science and technology document retrieval system named JDream II for JST database on April 2006. The database of JDreamII contains more than 39 million records and there are many kind of variant notations that we can't search exhaustively because of variant notations. This paper shows a variety of implementation to search variant notations speedy and comprehensively on JDreamII.

Keywords

science and technology, document retrieval system, variant notations, JDreamII

※本論文は第31回デジタル図書館ワークショップの論文です。

1. はじめに

独立行政法人科学技術振興機構（JST）では、2006 年 4 月から科学技術文献検索システム「JDreamII」の提供を開始した。JDreamII では約 3,900 万件という大量の科学技術文献を提供している。こうした文献のデータには、同じ単語でありながら表記が異なるため、単純に検索するとヒットしない「異表記語」が多数存在している。JDreamII では、ユーザがこうした異表記語を意識しなくとも漏れなく高速に検索することができるよう、各種対応を行ったのでその実現方式を報告する。

2. 異表記の種類と対応方法

2.1 異表記の種類

文献中には、単純な文字レベルから単語レベルまで様々な異表記語が存在する。以下にそれぞれの例を示す。

(1) 文字レベルの 異表記

「ソフトウェア」と「ソフトウエア」（拗音）や「Japan」と「JAPAN」（英語の小文字大文字）等の異表記を指す。

(2) 単語レベルの異表記

「タンパク質」と「蛋白質」等の単純な文字変換の規則では対応できない異表記を指す。特に科学技術用語、医学用語においては、日々新たな異表記が生まれている。

(3) 英単語の表記揺れ

英単語の場合は日本語の異表記とは異なり、名詞の単数・複数、動詞の現在・過去・進行形、英語・米語という表記の揺れが存在する。JDreamII では、英語標題等の英語フィールドに対しては、単語単位の完全一致方式を採っているため、システム側でこれらの表記の揺れに対応した。

2.2 JDreamII における対応

JDreamII では、2.1 のような様々な異表記を、ユーザが意識することなく網羅的に検索できるよう、以下の三つの対策を講じた。3 章以降でそれぞれの概要を説明する。

(1) 検索用データと表示用データの併用

(2) 日本語異表記展開辞書の搭載

(3) 英語異表記展開辞書の搭載

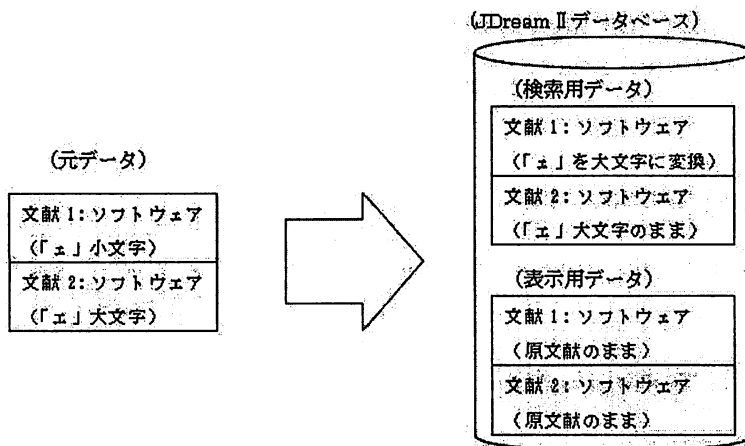
3. 検索用データと表示用データの併用

文字レベルの 異表記に対応するために、JDreamII では、予めルールを決め、データベースにデータを搭載する段階で文字変換を行っている。

3.1 データ構造

データベースに文献データを格納する際は、「表示用データ」と「検索用データ」という2種類のエリアにデータを格納する。表示用データには元データの表記のまま格納するのに対して、検索用データに格納する際は、以下の文字変換を実施する(図1)。

- 英語小文字→英語大文字
- 英語半角文字→英語全角文字
- カタカナ半角文字→カタカナ全角文字
- 半角記号→全角記号
- 拗音、促音→大文字



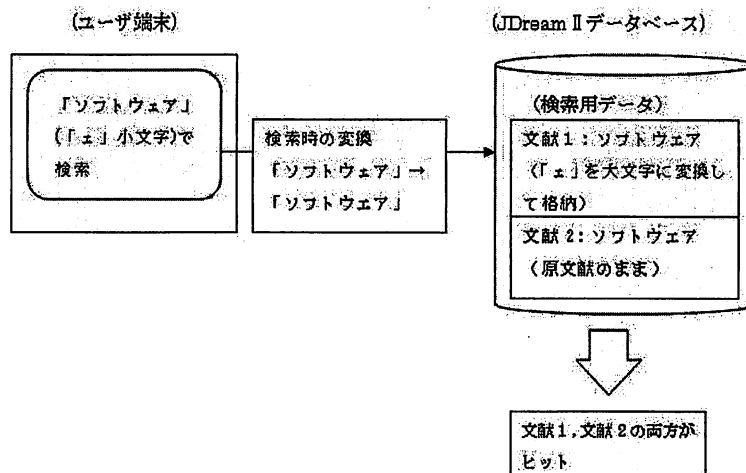
(図1) 検索用データと表示用データの格納

3.2 検索の実行

検索を実行する際は、図1の検索用データに対して行う。この場合、ユーザが入力した検索語をそのまま使用するのではなく、検索用データを格納した際と同様の文字の変換を行い、変換後の文字を使用して検索を実行する。こうすることにより、ユーザは大文字・小文字、全角・半角、拗音、促音といった様々な異表記を意識することなく漏れの無い検索を実行することが可能である。また、変換後の文字列を用いて1回だけ検索を実行するため、検索処理の高速化が可能である(図2)。

3.3 検索結果の表示

検索結果を表示する際は、原文献の表記をそのまま格納した表示用データを使用する(図3)。



「ソフトウェア」(「ェ」小文字)で検索すると原文献が「ソフトウェア」(「ェ」大文字)の文献と「ソフトウェア」(「ェ」小文字)の文献のいずれもヒット

(図 2) 検索時の文字変換と検索の実行

4. 日本語異表記展開辞書の搭載

単語レベルの異表記に対応するために、JDreamII では、「日本語異表記展開辞書」を搭載し、日々拡充を行っている。

4.1 日本語異表記展開辞書の必要性

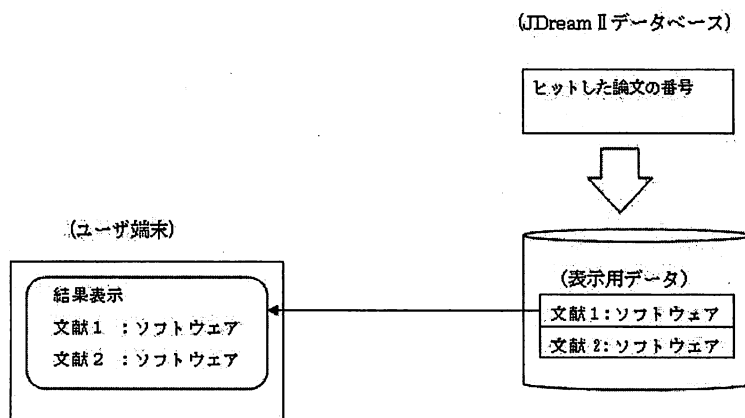
検索用データを用いた異表記対応方式では、新たに異表記を追加する場合、すでに搭載した検索用データを再作成、再搭載する必要があり、タイムリーに対応することは困難である。

一方、科学技術文献の世界では、単純な文字変換では対応できない異表記もあり、これらは日々新たに発生している。新たに発生する異表記に対応するため、JDreamII では専用の日本語異表記展開辞書を搭載した。この辞書に登録された内容に基づき、JDreamII は検索時に内部的に「or 検索」を実行する(図 4)。

こうした異表記辞書による自動展開機能は、ユーザが検索時に「展開する」「しない」を指定することが可能である。

4.2 日本語異表記展開辞書のメンテナンス

日本語異表記展開辞書には、JST が文献データに索引付けを行う過程で気がついた異表記語を随時追加する。これとは別に、JDreamII では、ユーザからの指摘を受け付けるための「日本語異表記展開辞書投書箱」も用意した(図 5)。



・ ・ 回答表示の際は、表示用データを用いて原文献の表記のまま表示

図 3) 検索結果の表示

5. 英語異表記展開辞書の搭載

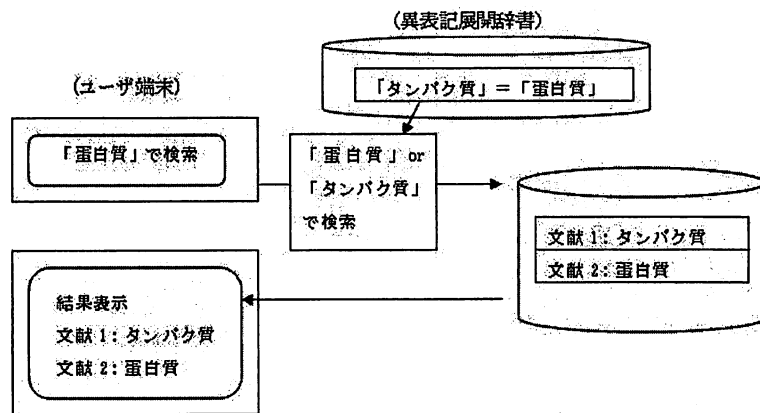
英単語についても、単数・複数、現在形・過去形・現在進行形、英語・米語という異表記が存在する。JDreamII ではこうした英語の異表記についても辞書を用意し、自動展開をサポートした。日本語異表記辞書と同様に、ユーザが検索時に「展開する」「しない」を指定することが可能である。

6. おわりに

JDreamII では、漏れの無い高精度の検索を実現するため、本稿で説明した「異表記対応」以外にも様々な工夫を行っている。また、利用者の要望に基づき、機能を追加改善しており、例えば、英単語について、「～ly」や「～ness」などの語尾変化に対して今年度中にシステム対応を行う予定である。今後とも利用者の声を踏まえ、網羅的で精度の高い検索システムを目指していきたい。

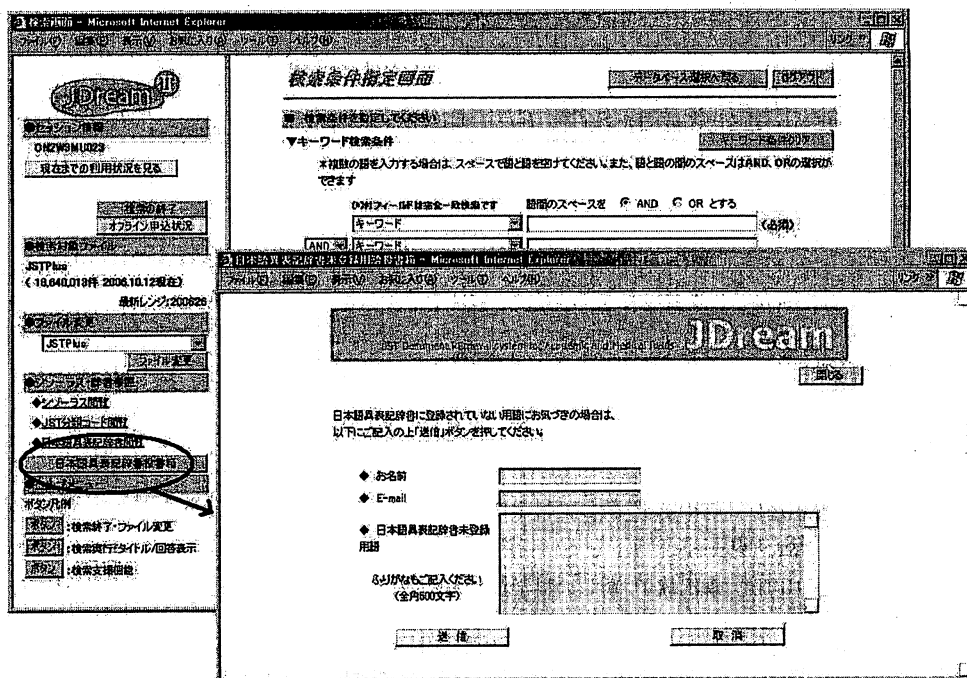
参考文献

- [1] 甲田彰, 平尾泰徳, 山田篤, 森田歌子. 新しい科学技術文献検索システム「JDream II」の技術解説 (I) . 情報管理, 49(4):182-190
- [2] 甲田彰, 平尾泰徳, 山田篤, 森田歌子. 新しい科学技術文献検索システム「JDream II」の技術解説 (II) . 情報管理, 49(5):266-273



・「蛋白質」で検索すると「タンパク質」の文献もヒット

(図 4) 検索時の日本語異表記の展開



(図 5) 日本語異表記展開辞書の登録画面