WWWによるJAPAN/MARCの提供実験

安齋 宏幸　山本毅雄
anzai@ulis.ac.jp, yamamoto@ulis.ac.jp
図書館情報大学
〒305 つくば市春日 1-2

JAPAN/MARC を UNIX-WS にコピーし提供実験を行った。本システムにおいて利用者は、WWW(World Wide Web) 環境によるクライアント ·サーバー方式により任意の端末から書誌情報の検索を行うことができる。利用者に分かりづらい分かち書き語による書名検索には、複数ハッシュふるい分けを用いたキーワード切り出しシステムを活用することにより、柔軟な検索を行うことを可能にした。また、DBMS には ADABAS を用いることで検索速度の向上と DB 管理の省力化をはかっている。本研究では日本語検索に対する複数ハッシュふるい分け法の有効性、ならびに、書誌情報提供において WWW を利用することの価値とその限界について検証する。

JAPAN/MARC EXPERIMENTAL SERVICE THROUGH WWW

Hiroyuki Anzai and Takeo Yamamoto
anzai@ulis.ac.jp, yamamoto@ulis.ac.jp
University of Library and Information Science
1-2 Kasuga, Tsukuba-city, Ibaraki, 305 Japan

An experimental JAPAN/MARC bibliographic information retrieval service through WWW(World Wide Web) is described. Since Japanese words are not separated by obvious delimiters, ensuring same segmentation between the query and the database is a problem. The present system solves the problem by using the multiple-hash screening technique for processing both book titles and query strings, based on the same dictionary and using similar algorithms. Database management is handled by ADABAS, reducing management chores and response time. The effectiveness of multiple hashing screening technique for a Japanese text-based inforamtion system is examined, and the limitation of WWW's hypertext enviroment for bibliographic information retrieval service is discussed.

1 はじめに

インターネットには無数のネットワーク情報資源が存在している。図書館などが提供している目録情報サーバーもまた、印刷出版物に関する情報を手に入れるための欠かせない情報資源となっている。しかしながら、その利用法は煩雑な場合が多く主題検索はおろか、目的とする書物のタイトルからでさえ満足に検索を行うことができないことがある。

書名検索を難しくしている原因のひとつが、分かち書きの存在である。日本語では1語という単位をはっきりと特定できない。そのため言葉をどこで区切って検索すればよいのか、果たして対象システムではトランケーションを使えるのか、キーワードはどういった基準で抽出されているのか、など、分かち書きに関連して利用者を悩ませる問題は多い。

今回我々が試作したシステムでは、中本が開発した複数ハッシュふるい分けによる日本語形態素抽出アルゴリズムを採用している。複数ハッシュふるい分けは、ICOT フリーソフトウェア No.33 形態素辞書をもとに作成するハッシュ表を使い、文字列中に存在している辞書該当語をすべて抽出する。データベースにおけるインデックス作成と利用者からの書名検索要求処理の双方に同抽出アルゴリズムを実装することで、日本語検索に伴う前述した欠点を克服することに成功した。

さらにWWWによるハイパーテキスト環境を広域ネットワーク（インターネット）への情報提供手段として用いることで、検索結果表示の著者名や件名、分類番号をクリックするだけで著者名・主題検索を行うことを可能にした。関心のある書誌を表示させ、その主題情報をもとに再び検索を行うことは一般によく用いられている。このような検索サイクルをハイパーテキスト環境で容易に行えるようにした。

2 システムの概要

試験的に国立国会図書館の許可を得て提供したのは、JAPAN/MARC 1年分の約73,549レコード (1993 vol.13 - 1994 vol.12) である。本システムにおいて、JAPAN/MARC データベースは ADABAS 上に構築される。

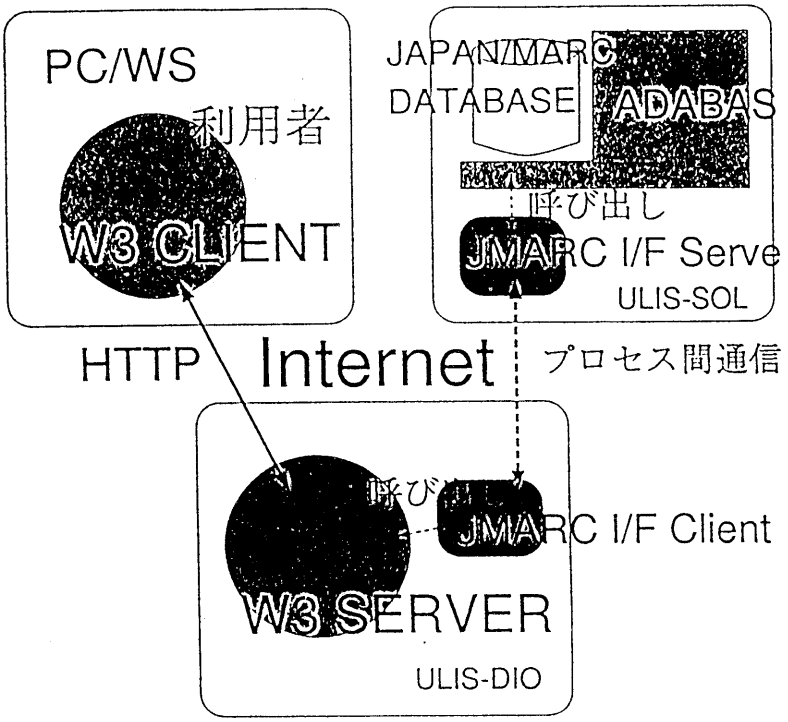
利用者は任意の W3 クライアントから本システムの URL(Uniform Resource Locator) を指定することで、検索フォームや検索例、データベースの情報などを受け取ることができる。W3 サーバーとクライアント間は、NCSA httpd による HTTP(HyperText Transfer Protocol) で通信を行う。検索フォームは WWW の Fill-Out Form 機能で実現されており、利用者が入力した検索式は送信ボタンを押すことによって W3 サーバー (httpd) に送られる。

httpd は cgi(Common Gateway Interface) を通じてインタフェースクライアントプログラムを呼びだし、検索要求をプロセス間通信でインタフェースサーバーに連絡する。ADABAS への検索要求は C 言語に埋め込まれた ADABAS コマンドによって行われる。JAPAN/MARC インタフェースサーバはクライアントから受け取った検索式の評価を行い、検索要求を ADABAS に送る。受け取った結果は HTML(Hyper-Text Markup Language) 形式に変換され、クライアントに送信される。インタフェースサーバは書名検索文字列への複数ハッシュふるい分けや、ADABAS コマンドに用意されていないトランケーション処理なども行う。

検索結果をあらわすHTMLはそのままW3クライアントに送られ、利用者は自分の好みのW3クライアントソフトウェアで結果をブラウジングすることができる。また、詳細な書誌情報の取り寄せや著者名、主題などからの再検索をクリック動作で行うことができる。

システムの全体構成図を下に示す。

システム全体構成図



2.1 JAPAN/MARCの利用

JAPAN/MARC MTを、HITAC M660を経由してSUN WS上にコピーし、これを標準的なテキストファイルに変換した。変換処理は国立国会図書館から出版されているJAPAN/MARCマニュアル[1]をもとに行った。

テキストファイルとして処理できるようになったMARCに対して、複数ハッシュふるい分けを用いた書名インデックス生成を行う[2]。

ここでいうふるい分けとは、あるキーが辞書中に存在しているかどうかを調べることである。複数のハッシング関数を用いて、辞書中に存在する単語を複数のハッシュ表の1ビットのフラグのエントリに変換する。ふるい分け対象のキーについてそのすべてのハッシュ表エントリに1ビットのフラグが認められれば、そのキーは辞書に存在しているとみなされる。

中本による研究ではICOT(Institute for New Generation Computer Technology)フリーソフトウェアの形態素辞書を使用している。本システムの作成においては中本による複数ハッシュふるい分けシステムを、書名検索におけるインデックス作成と照合処理の中枢とした。

変換したレコード例、与えられた書名に対するインデックス例を下に示す。このようなインデックスは対象書誌のシリーズ名についても作成され、MARC レコードに追加される。

変換したレコード例

```
010A4-17-268005-4
020AJP
020B93065055
100A19931117 1993        H1JPN 1312
251A 地域経済レポート
251D 平成５年
251F 経済企画庁調査局／著
270A 東京
270B 大蔵省印刷局
270D １９９３．５
275A ２６２ｐ
275B ２１ｃｍ
350A 平成５年の副書名：地域経済の構造変化と調整局面
360B １８００円
551A チイキ　ケイザイ　レポート
551X Ｔｉｉｋｉ　ｋｅｉｚａｉ　ｒｅｐｏｏｔｏ
551B 地域経済レポート
551D １９９３
658A ニホン／ケイザイ
658X Ｎｉｈｏｎ／Ｋｅｉｚａｉ
658B 日本／経済
677A ３３２．１０７
685A ＤＣ５５
751A ケイザイ　キカクチョウ　チョウサキョク
751X Ｋｅｉｚａｉ　ｋｉｋａｋｕｔｙｏｏ　ｔｙｏｏｓａｋｙｏｋｕ
751B 経済企画庁調査局
905A ＤＣ５５－Ｅ５１０
```

書名インデックス例

公務員試験労働法
公/公務/公務員/務/員/試/試験/験/労/労働/働/法/

最新データ比較政治ハンドブック
最/最新/新/データ/比較/政/政治/治/ハンド/ハンドブック/ブック/

精選不動産登記法商業登記法先例書式６０問
精/精選/選/不/不動/不動産/動/動産/産/登/登記/記/記法/法/商/商業/業/登/登記/記/記法/法/先/先例/例/書/書式/式/問/

お願い！かなえて !! トランプ恋占い
お/願/願い/い/か/かな/かなえ/な/なえ/え/えて/て/ト/トランプ/ランプ/恋/占/占い/い/

コンピュータでビジネスチャンスをつかむ本
コンピュータ/で/ビジネス/チャンス/を/つ/つか/つかむ/か/かむ/む/本/

オブジェクト指向分析手法入門
オブジェクト/オブジェクト指向/指/指向/向/分析/手/手法/法/入/入門/門/

ファイリング革命最前線
ファイリング/革/革命/命/最/最前/最前線/前/前線/線/

2.2　検索文字列へのキーワード抽出

実際に利用者が入力した書名検索語は、同様に複数ハッシュふるい分けを用いて形態素が抽出される。抽出されたキーワードは検索での処理効率を上げるために、他の文字列の一部として存在している部分文字列の除去を行う。

検索文字列の処理の一例

1. 公務員試験労働法
2. 公/公務/公務員/務/員/試/試験/験/労/労働/働/法/
3. 公務員/試験/労働/法/

2.3 書名による検索

同じ辞書を用いてふるい分けられた書名インデックスとキーワードは、もとの文字列が同じ場合には完全に検索を行うことができる。一見当たり前のように感じられる話ではあるが、利用者にとって分かち書きという存在がこの容易な検索を難しくしてきたという事実は前述のとおりである。さらに、一般に行われる「経済」「入門」などといったワード単位での検索の論理積作業が、事実上,一度の検索で自動的に行われるなどのメリットも存在する。直感的に入力された書名文字列は、その文字列に含まれる部分ワードの論理積検索に分解されることにより、さらに再現率の高い検索へと変化する。

一方、固有名詞や比較的新しい現代用語など、そもそも辞書に登録されていない単語は当然のことながら抽出されない。しかし、それら辞書に登録されていない未知語は、登録されているもっと短い文字列の単語の集合としてインデックス処理されることが多い。特に漢字やひらがなは 1 文字で意味をなす形態素として辞書に登録されている場合が多く、これらの漢字やひらがなから構成される未知語に対しても本システムにおける検索はかなりの柔軟性を示している[1]。ただし、検索速度と精度のさらなる向上のためには将来的に改善を要求される部分であり、辞書そのものの管理を行うトゥール群の開発が必要である。

2.4 ADABAS でのデータベース構築

加工した MARC レコードをデータベースとして運用するために ADABAS[3] を用いた。ADABAS はインバーテッド型 DBMS である。ディスクリプタ指定をおこなったフィールド値からの高速検索が可能で、検索集合同士のブール演算なども高速に行うことができる。同 DBMS を選択したことによって、検索から検索結果の抽出までの時間を十分満足できる範囲内に押さえ、複雑な検索処理も効率よく行うことができた。マシンの負荷状態や質問内容にもよるが、検索処理自体は数秒で行われる。検索要求と検索結果のネットワークでの転送にかかる時間や、検索フォームを W3 クライアントが描写するのにかかる時間のほうが一般に長くなる。

2.5 WWW での提供

インタフェースプログラムに対して WWW サーバーである NCSA httpd が Common Gateway Interface(CGI)[4] を使って検索を行えるようにした。インターフェースプログラムは検索要求を ADABAS に伝え、得られた検索結果を HTML 形式に変換して WWW サーバーに返す。主題検索を実現するためのアンカーや、検索結果が 0 件の場合の情報ページへのリンクも、この際に HTML のなかに折り込まれる。

[1]一方、カタカナの新語はその部分文字列も含めてまったく抽出できないことが多い。幸いにもカタカナ語は前後のひらがなや漢字との区切りが明確であるため、複数ハッシュふるい分けとは別にこういったカタカナの新語を抽出する仕組みを実装している。

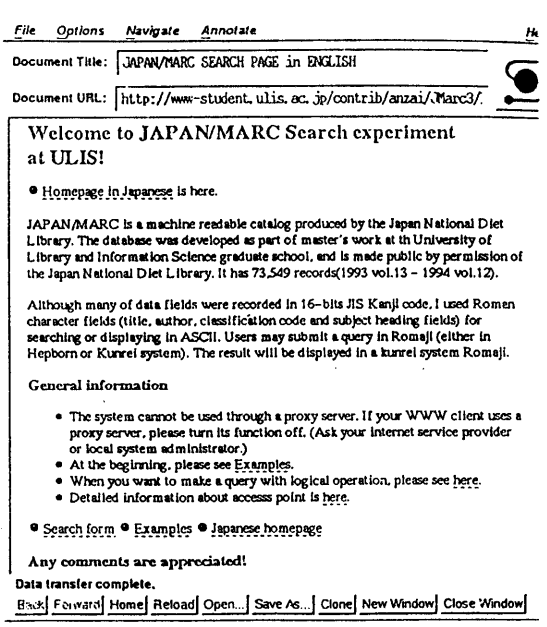
2.6 海外への対応

本システムには書名、著者名、分類記号など英数文字で記述が用意されている情報を海外に提供するための半角英数文字による検索・表示モードが用意されている。外国からの利用者が、日本語を使えない環境でも日本における書誌情報を手に入れることのできるための配慮である。インターネットでの情報提供における標準語が英語である以上、情報提供における非日本語環境への対応は無視できない。

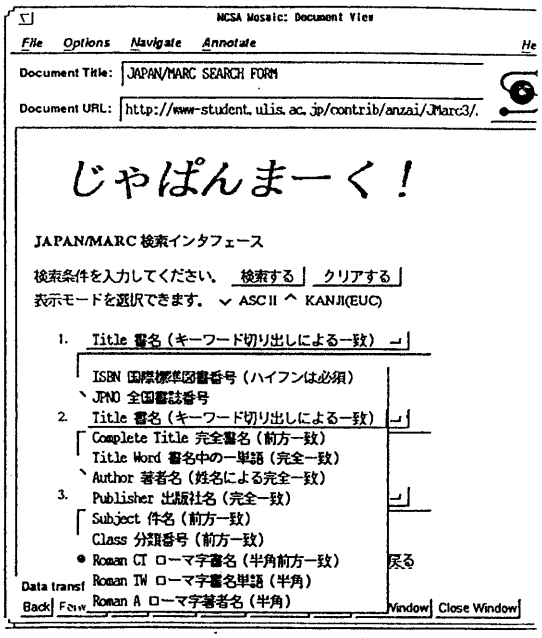
3 システム実行例

実際の検索例を以下に示す。

1. ホームページからはリンクをたどることによって、望むページを見ることができる。
2. 検索フォームでは、プルダウンメニューの形式で検索対象のフィールドを指定できる。結果表示は漢字とASCIIから選択することができる。

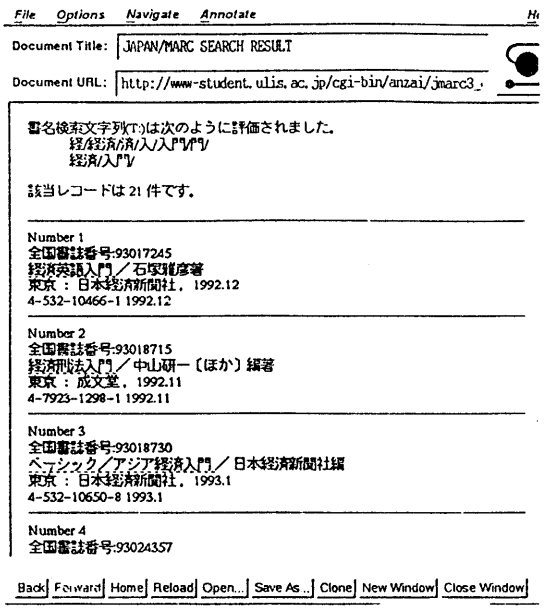


1

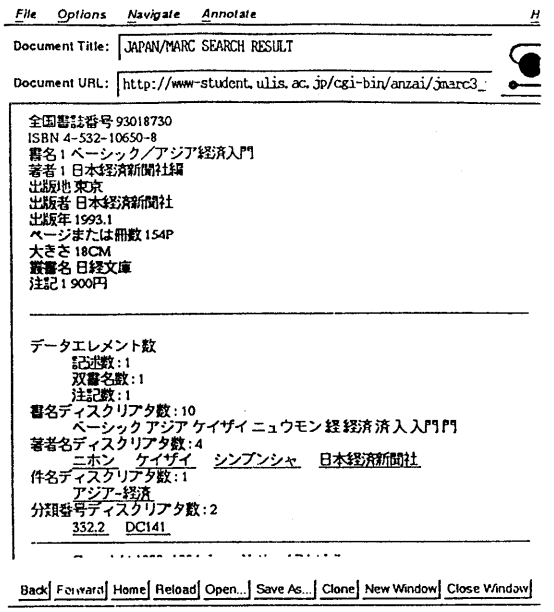


2

3. 検索結果は件数に応じて表示形態が変化する（簡略表示例）。書名をクリックすれば、書誌の詳細表示を見ることが出来る。
4. 書誌の詳細表示。著者名、件名、分類番号から再検索が可能。



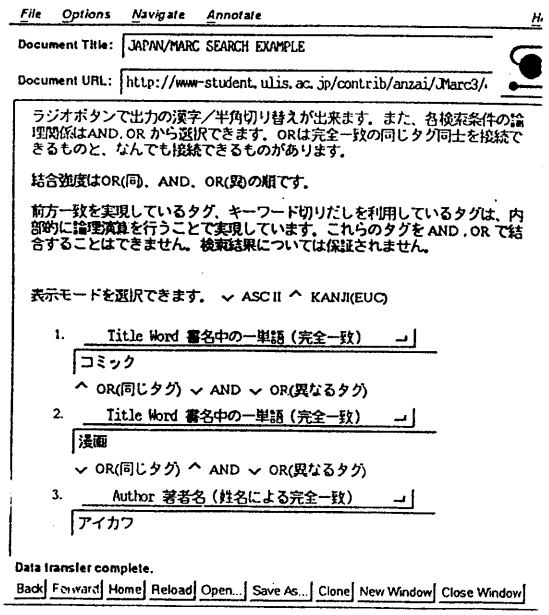
3



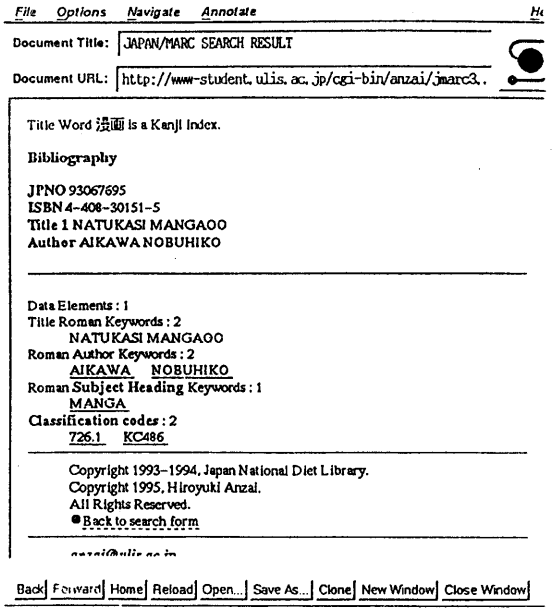
4

5. 検索ではブール演算が可能。

6. 半角英数での結果表示例。利用者が指定した「漫画」という書名単語が、複数ハッシュふるい分けによるインデックス対象となっているかを調べている。



5



6

4 システムの評価

この提供実験において、本システムは1日に100件前後の検索要求を受け続けている。情報システムである以上、蓄積されたデータがその価値を大きく左右する。残念ながら1年という収録年月では、利用者に本格的な検索のニーズを生じさせることができない。このため明らかに「通りすがり」的な利用のみが目立っており、検索要求の履歴からは本システムの利用者の明確な反応は得ることができなかった。

ハイパーテキストを用いた検索結果のブラウジングなどは、すでにたくさんのWWWを基盤に据えた情報システムで実現されている。しかし、次の点で本システムは評価できる。

- 複数ハッシュふるい分けを用いて日本語書名に対する柔軟な検索を実現した
- ADABASで検索処理速度の向上と豊富なアクセスポイントを実現した
- ハイパーテキスト(W3)を用いた主題や著者からの検索フィードバックを可能にした

本システムでは、検索結果集合を保存してブール演算を行うことができない。この課題は本システムで採用したWWW環境が、コネクションレス指向のクライアント・サーバー方式の情報システムであることから派生している。利用者には要求のたびごとに結果集合の書誌情報すべてが提供されるため、件数を絞り込んでから表示したり、結果集合を順に組み立てながら検索を行うことができない。

Z39.50[5][6]が、こういった書誌情報の提供には本来望ましい。Z39.50はANSI(American National Standards Institute)が定めた分散環境下におけるコンピュータ間のコネクション指向情報検索プロトコルである。このプロトコルを用いることによって、結果集合をサーバー側に作成し組み立てて検索を行うことができる。また、検索結果のソートや検索語のスキャニング、必要とするエレメントの絞り込みや検索結果上限などの設定等をサーバーに連絡することを可能にしている。

5 これからの展望

本システムではJAPAN/MARCを対象書誌データとして運用した。我々は同様の手法を取り入れて、図書館情報大学附属図書館OPACをインターネット上に提供する準備を始めた。この新しいシステムにはインターフェースプログラムにZ39.50のサーバーとしての特徴を取り入れることを目標にしている。

参照文献

[1] 国立国会図書館編. JAPAN/MARCマニュアル 図書編. 東京, 国立国会図書館, 1992,129p.

[2] 中本賢一、山本毅雄、長谷部紀元. 複数ハッシュふるい分け法の日本語情報システムへの応用. 情報システム 48-7. (1994.3)

[3] ソフトウェア・エージー. ADABAS(UNIX)マニュアル. 東京, ソフトウェア・エージー.

[4] National Cententer for Supercomputing Applications. Common Gateway Interface. http://hoohoo.ncsa.uiuc.edu/cgi/

[5] Z39.50-1995 Maintenance Agency Text. http://lcweb.loc.gov/z3950/agency/1995doce.html

[6] William Moen. ANSI/NISO Z39.50 Protocol : Information Retrieval in the Information Infrastrcture. http://www.cni.org/pub/NISO/docs/Z39.50-1992/www/50.brochure.toc.html