

Javaを用いたインターネット上での索引支援環境の構築¹

理学部 情報科学科 後藤 智範

インターネットの著しい利用拡大は、オンライン情報検索システムにも影響を与え、DIALOG, JOISを始めとしてWEBブラウザによるGUI画面での利用が普及しつつある。一方、データベースを作成する過程—索引作業も、電子的な編集システム、例えば科学技術振興事業団のNAISSが開発・利用されている。

我々は、索引作業のインターネット上での電子的支援環境の構築を意図し、JAVAを用いた索引作業支援システムを昨年度から開発してきた。本システムは、クライアント—サーバーシステムとして稼動し、索引作業を効率的に支援する。本発表では、機能拡張が行われたシステムの機能、その構造について説明し、さらに利用例、問題点についても具体的に報告する。

1. はじめに

近年、インターネットの飛躍的普及は、オンライン情報検索環境にも大きな影響を与えつつある。オンライン情報検索システムは、従来からコマンドライン入力による利用が主流であったが、1996年以来、DIALOG, JOISを始めとしてWEBブラウザによる利用形態に移行しつつある。

データベースの作成過程も、印刷媒体のシソーラス・索引語入力シートという従来方式から、科学技術振興事業団が開発し、1994年度から運用が行われているNAISS(New Abstracting and Indexing Support System)のような統合的な電子的編集システムに移行している^[1]。Internet/Intranet上でのデータの送受信、およびプラットフォーム非依存、GUIによる簡便な操作性を有するソフトウェアによる主題分析作業は、図1に示されるような形態で行われると考えられる。

主題分析作業に必要なデータは、利用者—主題分析者の要求によりサーバーから転送され、その結果、すなわち、統制索引語・非統制索引語・分類コード、その他の主題指標(物質索引などデータベース固有の主題指標)が付与されたデータも、ネットワークを介してサーバーに転送される。シソーラスなど主題分析に必須であるが容量の大きなデータは操作速度に影響を与えるため、intranet環境ではサーバーからクライアントに要求に応じて送られ、internet環境ではクライアント自身で持つような形態が望ましいと考えられる。

¹ 本報告は下記の発表内容およびその後の発展も含めて書き直したものである。

下村央人、若林千晴、後藤智範. "Javaによる索引支援ブラウザの試作 その2". 第35回情報科学技術研究会発表予稿集. p.133-137(1998).

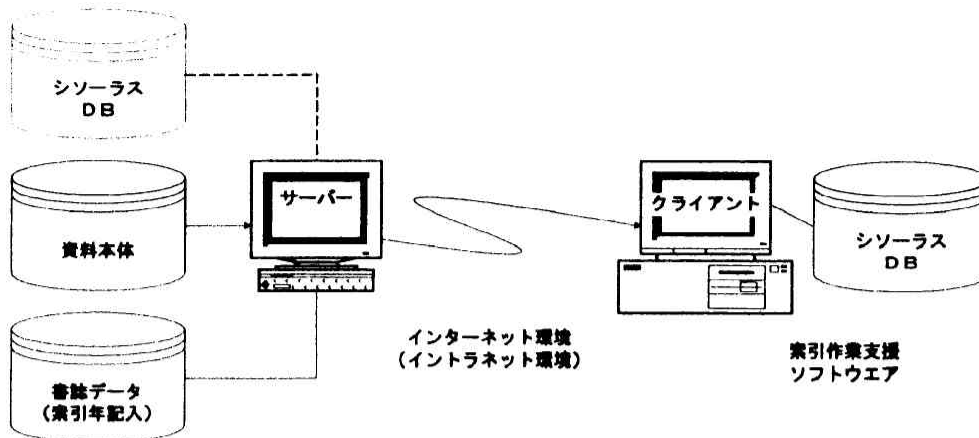


図1 Internet/Intranet 上での索引作業支援環境

2. Internet/Intranet 環境で稼動する索引作業支援ソフトウェア

筆者らは図1に示される形態で稼動する索引作業支援ソフトウェアをJAVAを用いて開発してきた[2]。当システムは以下のような特徴を有する。

- (1) クライアント-サーバーシステム
- (2) プラットホーム非依存
- (3) 主題指標(索引語などの用語構造)のグラフィカル表示
- (4) マウス主体オペレーション
- (5) オブジェクト指向設計

本システムは、書誌データに対して主題分析を支援するクライアント、クライアントに転送し、主題分析結果を受信する機能をもつサーバーからなる。サーバーもJAVAで記述されているため、本システムは、サーバー/クライアント、両者ともUNIX(Solaris2.6)およびWindowsNT、Windows95上で稼動する。

本システムはJAVA言語の特性を生かし、オブジェクト指向で設計されているため、対象となる書誌データベース固有の機能を容易に開発・実装できる。物質索引、数値索引など、特定の書誌データベース固有の主題指標については、索引語を表示するコンポーネント(window)に対する抽象クラスを作成することにより、機能の異なった部分だけをその派生クラスとして作成することができる。

3. 改善項目・新機能

現在、上述の機能に加えさらに実使用を意図し、下記に挙げる機能を開発した。

3.1 表示速度、操作性の向上

Java言語で書かれたコードは、C、C++などの言語と比較して実行速度が様々な理由により遅いことは従来から指摘されている。本ソフトウェアにおいては、JICSTシソーラスのように数万語を含む場合、用語リストWindowおよび、用語の階層構造表示についてこの影響が大きい。一昨年度XToolkitを用いて作成したWindow^[3]と同等の表示速度に近づくよう表示アルゴリズムを改善した。

- (1) 用語Windowにおける当該用語の表示高速化
- (2) 木構造表示 Windowの高速化

(1) 用語Windowにおける用語表示の高速化

従来は、Windowに表示された行の頭文字（あ行～わ行）が選択されてから、イベントが生じる度に、その行に含まれる全用語に対するリストコンポーネントをその都度作成していた(図3.1参照)。新方式は、全用語を50音及び、「きゃ」「しゃ」などの2文字の濁音を含む、総計93文字をハッシュキーとし各々の用語の配列をハッシュ表に格納しておくため、利用者が探したい用語の頭文字のボタンクリックによりWindowコンポーネントを生成し表示する方式に変更した(図3.2参照)。このため、従来の方式と比較するとWindowコンポーネント内に含まれる用語の数が減少し、利用者がスクロールにより当該用語を見つける時間、負担が大幅に軽減された。

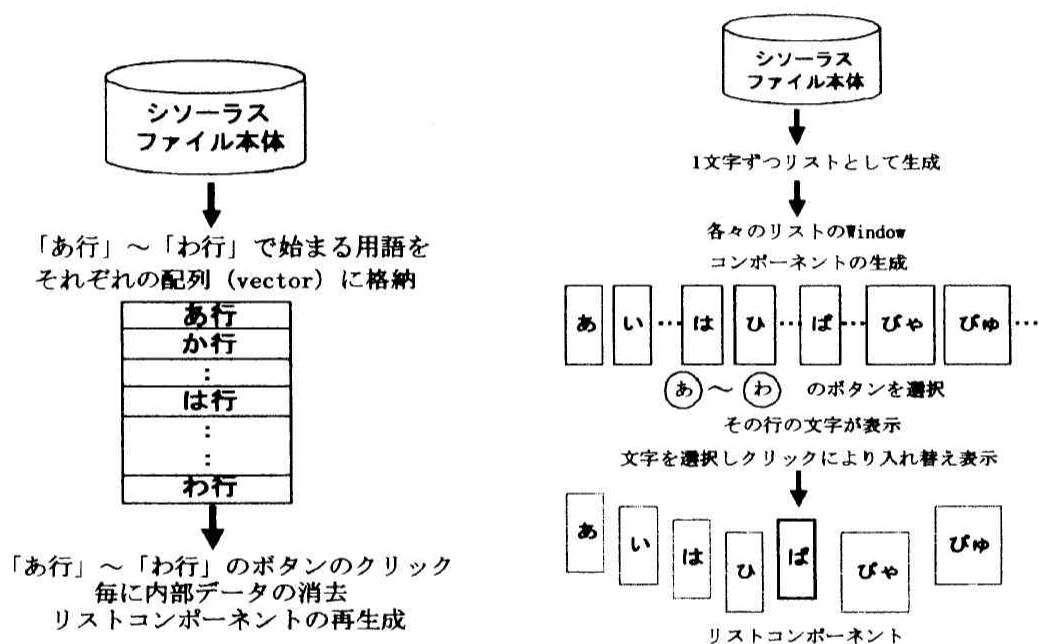


図3.1 用語Windowのアルゴリズム(旧方式) 図3.2 用語Windowのアルゴリズム(新方式)

(2) 用語木 Window表示の高速化

従来の木の表示の仕方は、用語表示Windowから用語が選択された後、その用語の木の最上位語を検出し、その最上位語から子を検索するために、順探索を行っていた。次に、根を基準に、木に含まれるそれぞれの用語ノードの表示座標を一時的に決定し、これを初期状態としていた。この方式では、用語ノードで重なりが生じてしまい、個々の用語ノードの座標はその木のレベル回の再計算が必要であった。新方式では用語は、語番号と語の両者をキーとするハッシュ表に格納する。最初に、リンク情報をもとに、木の論理構造を生成する。次に、木の最大レベルをもとに表示座標の大きさを求め、後順巡回より座標を一意に決定し描画する(図3.3参照)。順探索および座標の再計算は必要ないため、結果として、用語木の表示の高速化をもたらした。

旧方式と新方式の表示速度の相違を調べるために、用語木 Window が表示されるまでの時間を測定した。INSPEC、JICST シソーラスよりそれぞれ「ZOOLOGY」(レベル: 6, 用語総数: 256)、「改変」(レベル: 3, 用語総数: 159)を実験対象用語としたときの、そ

それぞれの表示速度を表 3.1 に示す。

表3.1 用語Windowの表示速度例

用語	データベース	従来方式(ms)	新方式(ms)
"ZOOLOGY"	INSPEC	4 2 8 1	2 0 3 1
「改変」	JICST	1 5 6 7 2	6 4 0

この表から「ZOOLOGY」では 2.1 倍、「改変」では 24.1 倍の速さになり、用語木の表示の高速化につながったことがわかる。従来方式は順探索を行なっているため、INSPEC は約 6000 語、JICST は約 45000 語であり用語総数が多く時間を要する。新方式はハッシュ表に保管しておくため、用語総数が多いほどその効果が大きい。上の表のような結果になったと考えられる。

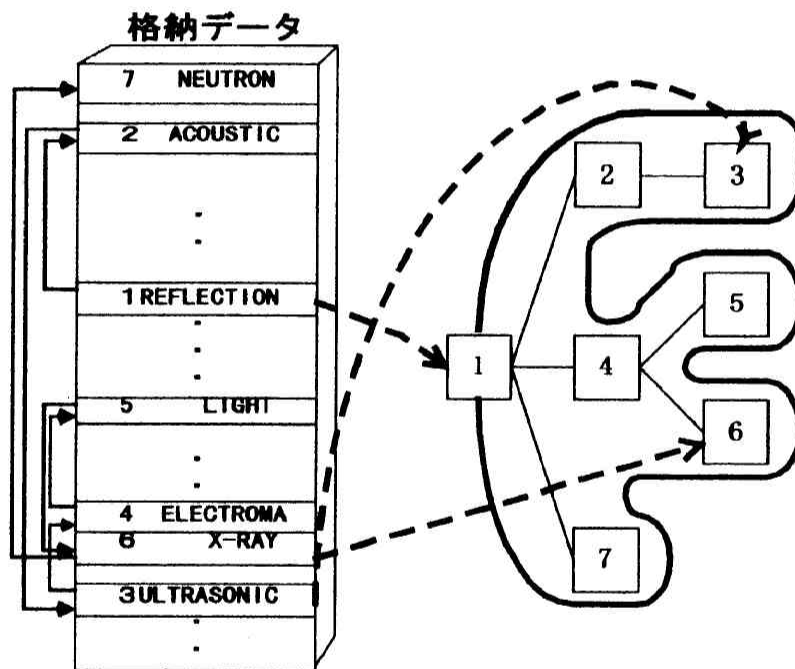


図3.3 用語木Window表示のアルゴリズム(新方式)

(3) 操作性の向上

操作性については、新たに2種類のWindowを導入し、改善を試みた。

- (a) PopUp メニューの導入による用語Windowからの統制索引語の入力
- (b) 過去に使用した用語の一覧表示Window

従来、統制索引語を登録するためには、用語木Windowが表示されないと登録できないように実装されていた。用語Windowに表示される用語が統制索引語として妥当な場合、その時点で登録できることが望ましい。今回の改良では、用語Windowに表示される用語

上で右クリックすると、PopUpメニューを表示し、ここで登録ボタンを押すことに登録できるように実装した。

通常、索引者は特定の主題分野の資料に対して索引作業を行う。したがって、シソーラス中の当該領域の分野の索引語を頻繁に用いることになる。このような使用形態を考慮し、用語Windowに表示される任意の索引語をファイルに保存し、システムの次回の起動時にメインのWindowに表示される機能をもたせた。この機能により、メインのWindowに表示される索引語に対して、用語木(上位一下位語)および関連語の表示、さらに登録も可能となった。

3.2 非統制索引語の候補語表示

非統制索引語は、もともと索引語集合に存在しない用語であるため、キーボード入力は必須であるように思われる。しかし、標題・抄録を対象に重要語抽出を行い、非統制索引語の候補語リストを表示し索引者がマウスにより選択する、という方式も可能である。この方法により、全ての主題指標に対してマウス操作を主体とする入力が可能となる。重要語同定・抽出ツールを開発し、これを本システムに組み込むことによりこの機能を実現した。

4. 稼働例

図4.1は、下記の状況を示した索引支援システムの実働画面を示したものである。

- ・索引支援システム(クライアント)を立ち上げ、サーバーから索引作業対象の書誌レコードを書誌データWindow読み込む(右側のWindow)。



- ・対象レコードに対し、関連する統制索引語を選定するために、Thesaurusボタンを押し、用語リストWindowを表示させる(左上のWindow)。



- ・「制御」(主題と思われる用語)を用語リストWindowから探し、用語木Windowによりこの統制索引語を含む木を表示させる(左下のWindow)。

図4.1に表わされる状態で、用語木Windowの中で当該レコードに対し、用語木Windowで該当する統制索引語に対し、マウスクリックを行う。図4.2は、この操作により、その結果、統制索引語フィールドに「制御」が登録された状態を示している。

書誌データ表示WindowのAnalysisボタンを押すことにより、標題および抄録に対して、重要語抽出が行われ、同定された重要語が(実際の画面では赤色で)表示される。この状態で、非統制索引語と想定される用語の上にマウスを移動し、左ボタンを押すと非統制索引語フィールドに押された用語が登録される。

図4.3は、「静電多重極」、「調和ポテンシャル」、「対象変換」が登録された状態を示している。統制索引語フィールドおよび非統制索引語フィールドにおいて、誤って入力した用語は削除することもでき、また、非統制索引語フィールドでは、標題・抄録中に現れない用語をキーボード入力により登録することも可能である。

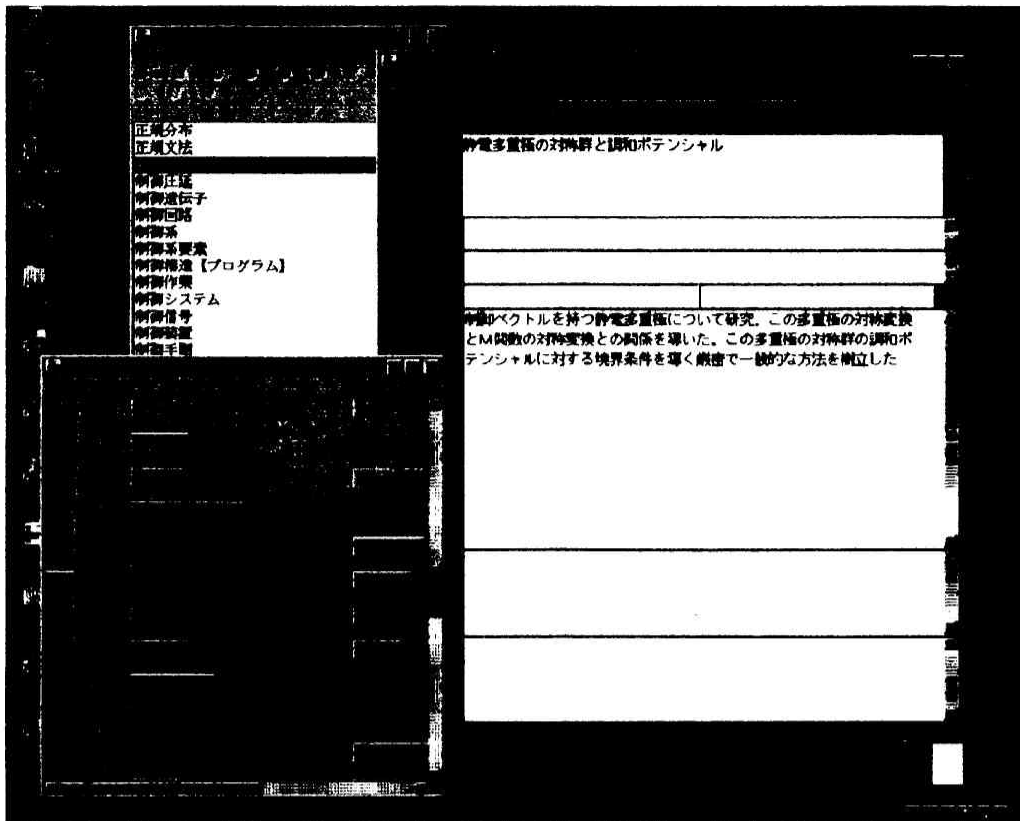


図4.1 索引支援システムの実働例

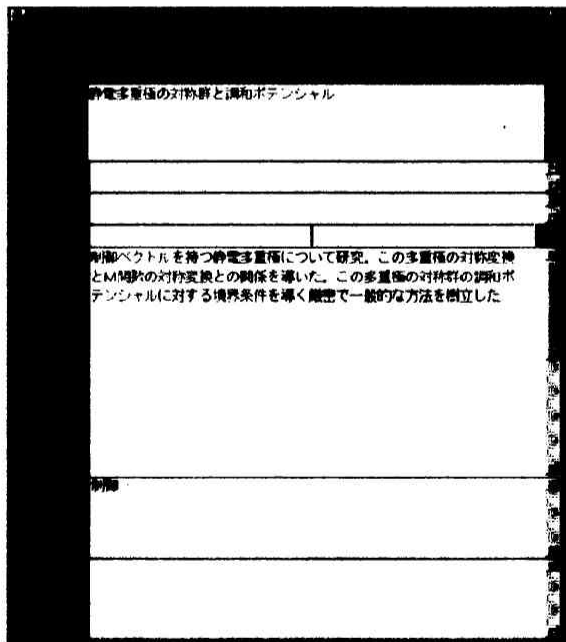


図4.2 統制索引語の登録例

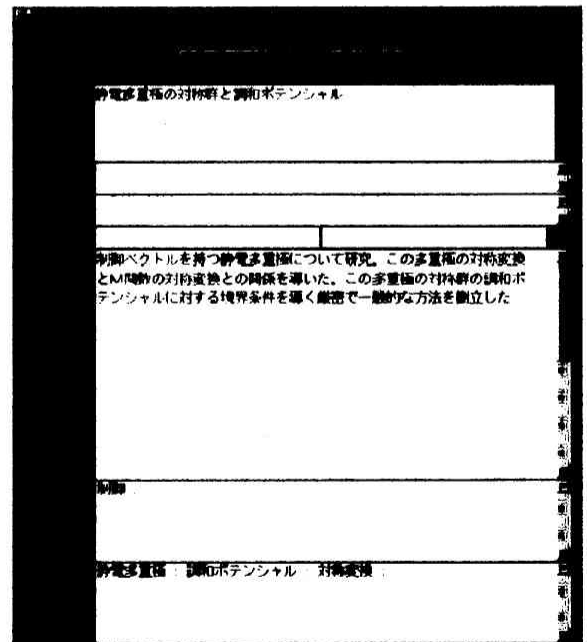


図4.3 非統制索引語の登録例

5. おわりに

本システムは書誌データベース作成において、最も中心的で、コスト・時間を要する主題分析作業(索引語・分類カテゴリー付与)の効率化を意図し、試作したものである。こ

の観点から、本システムに残された困難かつ重要な機能は、統制索引語の候補語提示機能であろう。索引者を全く必要としない自動索引システムを商用データベースに導入するのは、現時点では非現実的と考えられる。しかし、本システムにテキストの内容解析機能、および重要語と索引者が付与した統制索引語との関連の学習機能を備えることにより、当該資料に適合する候補として複数の統制索引語を提示することは可能であると考えられる。

参考文献

- [1] 河村昌也、高野文雄、佐原卓、原田英二. JICST抄録・索引支援システム「NAISS」-JICSTにおけるデータベース作成の効率化-. 第32回情報科学技術研究集会発表論文集. pp. 163-174(1995).
- [2] 下村央人、杉原直樹、屋島新平、豊島 隆之、後藤 智範. Javaによる索引支援ブラウザの試作. 第34回情報科学技術研究集会発表論文集. pp. 117-124(1998).
- [3] 下村央人、笹川泰秀、後藤智範. “汎用ソースブラウザの試作”. 第33回情報科学 技術研究集会発表論文集. p. 99-105(1997).