軸づけ検索法
― 文書からの抜粋を抽出・整列して出力する全文検索法*

金田 泰

日立製作所中央研究所
E-mail: kanada@crl.hitachi.co.jp

軸づけ検索法という，文書集合からの抜粋情報を抽出し整理する機能をもつ全文検索法を開発した．この方法では，ユーザが検索語を入力すると同時に年代，地域，数量などの軸をメニューから選択する．すると，軸と検索語に関連する部分の抜粋とその原文へのハイパーリンクがその軸にそって整列出力される．この方法をつかえば，検索結果が膨大でもユーザ要求にあわせて整理されているので，ユーザはそれをサーベイすることができる．この方法を百科事典と新聞記事に応用した結果，分散された関連情報がうまく収集でき，検索結果のあいだの関係を発見することができることがわかった．たとえば，検索文字列として「流域」，軸として「面積」をあたえると，百科事典から世界の川の記述を収集し流域面積でソートした結果がえられる．

Axis-specified Search: A Full-text Search Method for Extracting and Ordering Excerpts from Documents*

Yasusi Kanada

Central Research Laboratory, Hitachi Ltd.
E-mail: kanada@crl.hitachi.co.jp

A full-text search method, which is called an axis-specified search method, is proposed. Excerpts are extracted from documents and ordered by using this method. The user selects an axis, such as year, area or quantity, from a menu, in addition to typing strings to be searched. Then, excerpts related to the axis and strings, and hyperlinks to the original sentences are ordered along the axis and displayed. Even if the number of results is very large, the user can easily survey them, because they are well structured. This method has been applied to an encyclopedia and a newspaper articles. In these applications, distributed descriptions that were related to each other could be gathered, and the user could discover their relationships from the results. For example, by specifying “basin” for a search string and “area” (m^2) for an axis, descriptions of the world’s largest rivers were extracted from the encyclopedia and sorted according to their basin areas.

* この報告の内容は金田 [Kan 98] をもとにしている．

1. はじめに

インターネット, CD-ROM などのメディアが普及してきているが, 今後さらに DVD-ROM などのあたらしいメディアが普及するであろう. このようななかで, 従来はプロのサーチャを介しておこなっていた情報検索をエンド・ユーザが直接おこなう機会がふえている. しかも, 従来よりはるかに大量のテキストの全文を検索するようになってきている. Alta Vista, goo などの WWW 検索エンジンにおいてそれはすでにある程度実現されているが, 今後さらにさまざまな電子テキストが普及し, ユーザによる大量のテキスト検索がおこなわれるようになるとかんがえられる. そして大量のテキストを検索すれば, 当然, 大量の検索結果がえられる.

このような背景のもとで, 情報検索にはつぎのようなことが要請されるであろう. 第 1 の要請は,「文書」よりこまかく意味的にまとまりのある単位での検索ができることである. このような検索を**細粒度の検索**とよぶ. 従来の情報検索は基本的に文書を単位としている. しかし, ひとつの文書には複数の話題がふくまれているから, それらをくべつして検索できるべきだとかんがえられる. しかも, 分割単位のなかに複数の話題がふくまれたり, ひとつの話題が複数の単位にまたがったりするため, 章, 節, 段落などの構文的な単位で文書を分割するパッセージ検索では適切でないとかんがえられる.

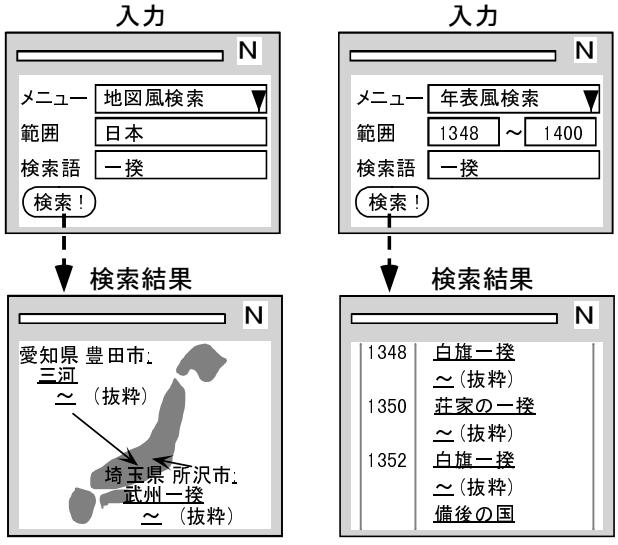
第 2 の要請は, **整理された検索結果がえられること**である. 大量の検索結果がえられると, それが雑然としたかたちであればもちろん, また要求にあわないかたちでまとめられていても, ユーザはそれを読みこなすことができない. しかし, それが要求にあうかたちに整理されていれば, ユーザはそれを概観し取捨選択して, ほしい情報をとりだして読めるであろう. 従来の検索法においては, 単純な検索条件をあたえると大量の結果がえられて読めなくなる一方で, 結果をうまくしぼりこむのはむずかしく, しぼりこみによって必要な情報まですてられたり, 関連情報を発見する機会がうばわれたりする可能性があった.

軸づけ検索法は, 上記の 2 つの要請に部分的にこたえられる, あたらしいテキスト情報検索法である. 第 2 章でその機能を説明し, 第 3 章でその実現法の概要とそれに関するいくつかの話題をとりあげる. 第 4 章では軸づけ検索の百科事典と新聞の検索への応用についてのべ, 最後に関連研究と結論をしめす.

2. 軸づけ検索の機能

軸づけ検索は細粒度の全文検索法である. 軸づけ検索においては, ユーザは軸を選択し, 検索語を入力する. 検索語は検索主題をあらわし, 軸は結果を整理するための汎用の方法を指定する. 全文検索結果は軸にそって整列される. 抽象的にいえば, 結果は軸によって指定される空間上に配置される. クラスタリングによる検索結果組織化法 [Cut 92] [Cut 93] [Mor 95] とはちがって, 軸つまり整理の基準はユーザがあたえる. ただし, 軸の候補は検索システムによってあらかじめきめられている. 軸によって指定される空間を**特徴空間** (feature space) とよび, 軸上の値を**特徴値** (feature value) とよぶ. これらの用語をつかって定義するならば, 軸づけ検索とは, 検索対象の文書集合から抽出された文字列があらわす特徴値にしたがってその文字列をふくむテキスト (へのハイパーリンク) を, ユーザが選択した特徴空間上に配置する検索・組織化法だといえる. 軸だけでなく特徴値の範囲もユーザが指定できる. このばあいは, 範囲外の結果は削除されることになる.

例を図 1 (a) にしめす. ユーザは「一揆」という語を指定して百科事典を検索する. ユーザは「地域」を軸として選択し, さらに「日本」を特徴値の範囲として指定している. すると, 事典項目からの抜粋が地理的な位置によってソートされる. 抜粋は特徴値である日本の地名をふくむ. また, 検索語は抜粋にふくまれるか, または抜粋されたテキストの近傍にある. 検索結果は表のかたちで表示することも可能であり, また図 1 におけるように地図上に表示することも可能である. この例においては,「三河」と「武州一揆」とが事典項目名である. 項目名につづく文字列が抜粋である.



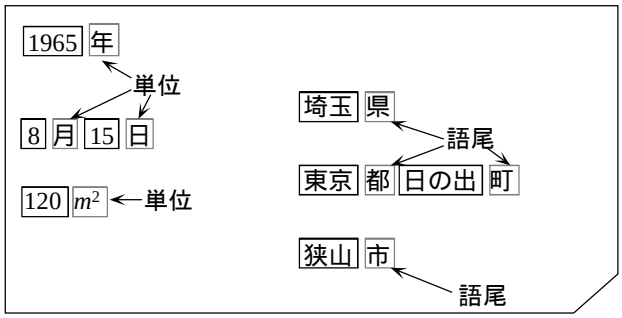
(a)「地域」を軸とする検索　(b)「年代」を軸とする検索

図 1. 軸づけ検索の機能 — 2 つの例

見出しだけでなく本文から抽出された地名によって, その周辺から抜粋されたテキストとその原文へのハイパーリンクが整列される. 抜粋されたテキストまたはハイ

パーリンクが指示するテキストによって, ユーザはその地名をふくむ話題の全体を知ることができる. 図 1 においては下線部に事典本文へのハイパーリンクがうめこまれている. とくに, 抜粋にうめこまれたハイパーリンクは抜粋の原文 (文書の途中) へのリンクである.

軸の指定法としては, おおきくわけて数量の単位系を指定する方法と語尾 (または語頭) を指定する方法の 2 つがある. これらの例を**図 2** にしめす.



(a) 年月日と数量の抽出　　(b) 地名の抽出

図 2. テキストからの特徴値の抽出例

単位系を指定する方法を (広義の) **数量検索**とよぶ. この方法では軸は検索対象のテキストにあらわれるべき数量の単位によって指定される. たとえば, もし軸が年代, 日付, あるいは時間であれば,「年」,「日」,「時」やほかの時間単位がつかわれる (図 2 (a)). これらの単位をもつ数量が文書集合から抽出され, 整列される. 時間はさまざまな単位で表現されるが, それらは換算可能である. したがって, それらはすくなくとも理論的にはおなじ軸のうえにおかれる. 検索語に対しては全文検索がおこなわれて, 数量検索の結果とマッチングがとられる. つまり, 数量があらわれる位置の近傍に検索語がなければ, その検索結果はすてられる (詳細は 3.3 節参照). この検索法によってえられる結果の例を図 1 (b) にしめす. この図は年代を軸として「一揆」という語を検索する例である. 年代の範囲として「1348 年から 1400 年」を指定している. 結果は年表のかたちでえられる.

数量検索で有用な他の単位の例をあげる.

- 「円」と「銭」— 日本の通貨単位. 換算可能.
- 「ドル」と「セント」— 米国の通貨単位. 換算可能. これらは円に対しても換算可能だが, レートが変動するので「円」,「銭」とはわけてあつかっている.
- 「人」と「名」— 人員の単位. たがいに等価.
- 「頭」,「羽」,「尾」,「匹」など — 動物数の単位.
- 「個」,「冊」,「種」など — ものの数の単位.
- m, km, mm, 光年, および km/h など — ながさや距離の単位と速度の単位. 速度の単位以外はたがいに換算可能. 速度の単位と他とは換算できないが,「… km」という表記が距離と速度のどちらをあらわすのかはくべつしにくいので, あえてまとめてある.
- m^2, mm^2, エーカー, ヘクタールなど — 面積の単位. 換算可能.
- m^3, mm^3, *l*, m*l* など — 体積の単位. 換算可能.
- Hz, kHz, MHz など — 周波数の単位. 換算可能.
- ℃, K — 温度の単位.

第 2 の方法は**部分語検索**である. この方法では軸は語尾 (または語頭) で指定される. たとえば, 日本では地名には通常, 県, 郡, 市などの語尾がつく (図 2 (b)). 語にこれらの語尾がついているときは, それが「地域」という軸にのる特徴値であることがわかるし, 構文的な情報をつかうだけで地名をよい精度で抽出することができる. 数量のばあいとはちがって, これらの語尾は数理空間を直接指定しているわけではないが, 地域軸にかぎっていえば地名は特定の経緯度 (の範囲) にむすびつけられるので, 間接的に 2 次元空間を指定していることになる. 検索の方法はおよそ数量検索と同様である. 図 1 (a) は部分語検索の例である.

部分語検索で有用な他の語尾の例をあげる.

- 「川」,「湖」,「海」,「山」— 地理的特徴をあらわす語尾.
- 「主義」,「教」.
- 「派」.
- 「大統領」,「首相」.

これらの語尾は, それがつく語がある概念空間 (カテゴリー) 上にあることを指定しているとかんがえられる. その概念空間が特徴空間である. ただし,「富士山」と「高見山」のように, ひとつの語尾 (「山」) が複数の空間のいずれかを指定できるばあいがあり, 構文情報だけでひとつの空間だけをとりだすことはできない. 概念空間には全順序は定義されていないが, 語を辞書式にソートすることによって全順序をつけることができる.

数量検索, 部分語検索のいずれにおいても, システムは指定された部分文字列をふくむ文字列を検索する. 数量検索においては全体文字列は単位つき数量であり, 部分文字列は単位である. 部分語検索においては全体文字列は語であり, 部分文字列は語尾 (または語頭) である.

軸づけ検索において検索語と軸は基本的には “and” 関係の検索条件である. しかし, 上記の例のような単位や語尾は特定の分野だけでつかわれるのではなく, 汎用性がある. したがって, 軸づけ検索においては個々

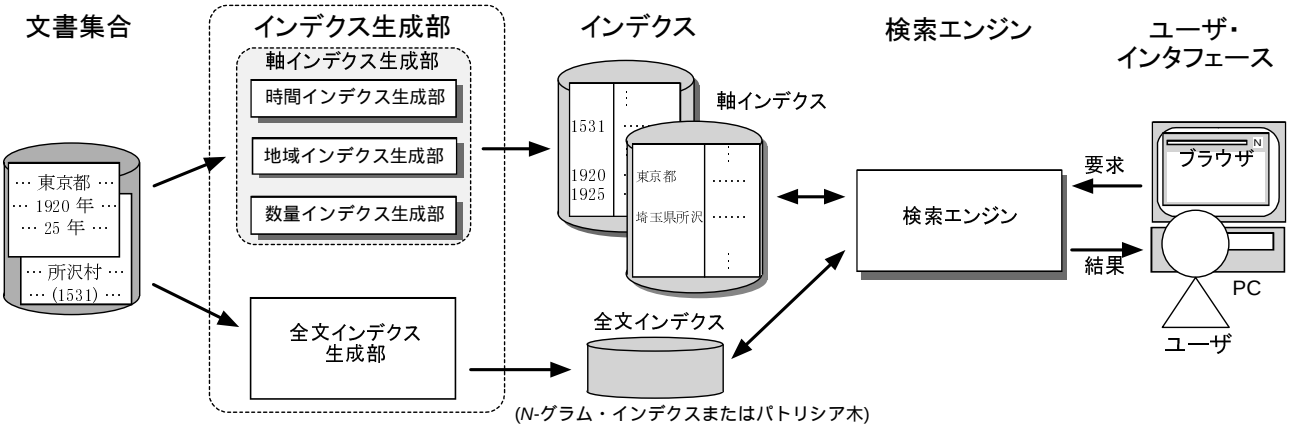


図 3. 軸づけ検索のためのシステムの概略構成

の検索に特徴的な検索条件が検索語としてあたえられ，検索結果の整列法が汎用的な軸によってあたえられるとかんがえることができる．

前章でのべたテキスト検索への要請は，軸づけ検索をつかえばつぎのようにみたされる．第 1 の要請は「文書」よりこまかく意味的にまとまりのある単位での検索ができることだった．軸づけ検索においては文書単位でなく特徴値をふくむテキストへのハイパーリンクを単位として検索するので，ユーザはちょうど必要な部分を読むことができる．したがってこの要請をみたしている．第 2 の要請は整理された検索結果がえられることだった．「整列」は整理法としては単純なものではあるがその基本であり，要請を部分的にみたしている．

3. 実現法

軸づけ検索の実現に関するいくつかの重要な話題をとりあげる．

3.1 概要

軸づけ検索システムは 2 つの主要な部分によって構成される (**図 3**)．それらはひとくみのインデクス生成部と検索エンジンとである．インデクス生成部は，ユーザ要求が発生する以前に文書集合から軸インデクスと全文インデクスとを生成する．軸インデクス生成部は既定のパターンにマッチする文字列を文書から抽出し，抽出情報を正規化し，軸インデクスに登録する．情報抽出の方法は 3.2 節で説明する．軸インデクスを使用する目的は，それを使用しないばあいにかかる検索時間を劇的にへらすことにある．軸インデクスは軸の型ごとに生成する．それは，型ごとにことなる構造のインデクスが必要だからである．全文インデクス生成部は従来の全文検索と同様の構造をした全文インデクスを生成する．

検索エンジンはユーザによって起動される．ユーザは軸を選択することによって特徴空間を指定し，特徴値の範囲を指定し，さらに検索語を指定する．検索エンジンは選択された軸に対応する軸インデクスから指定範囲の特徴値をもつ情報を検索し，検索語の全文検索をおこなって軸づけ検索の結果とマッチングをとる．そして，結果を軸にそって整列する．検索エンジンは全文インデクスをさきにひいてもよいし，軸インデクスをさきにひいてもよい．どちらがより効率的かは，ばあいによる．

軸インデクスはユーザ要求にさきだって生成されるため，使用可能な軸はあらかじめきめられていて，ユーザが自由に入力することはできない．検索結果はスコアづけされ，スコアがひくすぎる項目はおとされる．

3.2 情報抽出とインデクス生成

軸インデクス生成部は検索対象の全文書を入力し，既定の文字列パターンにマッチする文字列を抽出する．マッチング・パターンの組は軸ごとに定義される．抽出されたテキストは正規化され，軸インデクスに登録される．文脈独立な規則によって抽出される特徴値もあるが，省略された西暦年のように文脈依存のものもある．マッチング・パターンとそれにマッチした特徴値の正規化の方法とは，検索対象テキストの性質にあわせる必要がある．ことなる種類のテキストにあらわれるおなじ文字列はことなる特徴値に正規化されうる．この情報抽出の方法について 2 つの例をつかって説明する．

最初の例は年代を軸とする検索であり，検索対象としては世界大百科事典 [NEC 95] [HDH 98] を想定している．この例ではつぎの形式の年代を抽出する．

- 「年」がついた 1 ～ 4 桁の西暦年．たとえば「1989 年」．
- 「年」がついた西暦年の下 2 桁．たとえば「89 年」．
- 「年」がついた 1 ～ 2 桁の和暦年．たとえば「平成 10 年」．
- 「…000 年前」または「… 万年前」．
- 括弧つきの西暦年．たとえば「ロシア革命 (1917)」．

- 「… 世紀」または「前 … 世紀」.

年代は西暦数値に正規化される．たとえば第 2 の例においては西暦の最初の 2 桁は先行する 4 桁の西暦年を使用しておぎなうことができる．世界大百科事典においては 99% 以上の 2 桁の西暦年はこの方法でただしくおぎなうことができる．また，毎日新聞 [Mai 95] においてはそもそも西暦年を 2 桁で記述しているばあいはすくない．しかし，他の文脈では正規化はかならずしも容易でないであろう．

第 2 の例は地域を軸とする検索のための情報抽出法である．この方法は日本語で記述されたテキストの検索法において，ひろくつかうことができる．つぎのような地名を抽出することができる．

- 「… 県」のかたち．「北海道」,「東京都」,「大阪府」,「京都府」もあわせて抽出する．
- 「… 郡」,「… 市」．東京都においては「区」も郡や市とおなじレベルとしてあつかう．
- 「… 町」と「… 村」.

「県」,「道」,「都」,「府」はそれをふくむ地名が都道府県であることをあらわす．「郡」と「市」とはそのつぎのレベルであり,「町」,「村」はさらに下位のレベルである．したがって日本の地名に関しては，上記の 3 つのレベルからなる，あらいモデルをつくることができる．地名は省略された部分をおぎなうことによって正規化される．たとえば,「中野区弥生町」という地名は「東京都 中野区 弥生町」と正規化される．

軸インデクスの例を**図 4** にしめす．図で「キー」とあるのが正規化された特徴値であり，このインデクスをひくためのキーである．換算可能な単位をもつ数量やことなる形式の等価な数量が検索できるためには，キーは正規化されている必要がある．原文における抜粋テキストの位置を指定するために文書 ID とその先頭からの位置とが使用される．「文書 ID」は特徴値があらわれる文書 (またはその一部) の識別子である．「位置」は特徴値をふくむテキストの位置である．位置は文書先頭からの変位によってあらわす．

「抜粋」は特徴値があらわれるテキスト部分をぬきだしたものである． 検索性能を向上させるために，抜粋をインデクス中に格納しておくことができる．あるいは，抜粋はそれを表示する直前に原文から抽出するようにして，インデクス・サイズをおさえることもできる． ひとつの特徴値に対して複数の項目が登録されるばあいもある．

3.3 検索

軸と検索語の両方が指定されて検索エンジンがよびだされたときには，全文インデクスと軸インデクスの両方から検索結果がえられる．つぎのようにしてこれらのマッチングがとられ，ソートされる (**図 5**)．検索対象のテキストにおける検索語の出現位置が全文インデクスからもとめられ，特徴値の出現位置が軸インデクスからもとめられ，さらに検索語と特徴値との距離 d がそれらの差としてもとめられる．検索結果のスコアは d に関して単調に減少する関数をふくむ．我々のプロトタイプで使用している関数形は $1 - 10^{-5}d^2$ である．d は文字数で評価している．スコアがひくすぎるときは，その検索結果はすてられる．検索語が複数回あらわれるときは，もっともちかいものが評価につかわれる．評価関数には文書内および文書集合内での検索語の出現頻度もふくまれる．

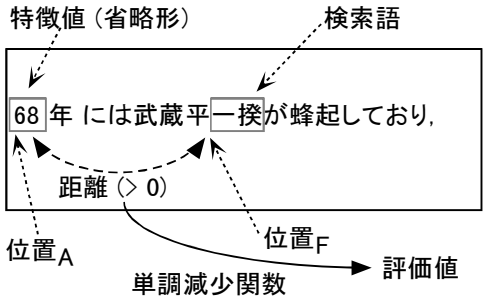


図 5. 検索結果の評価

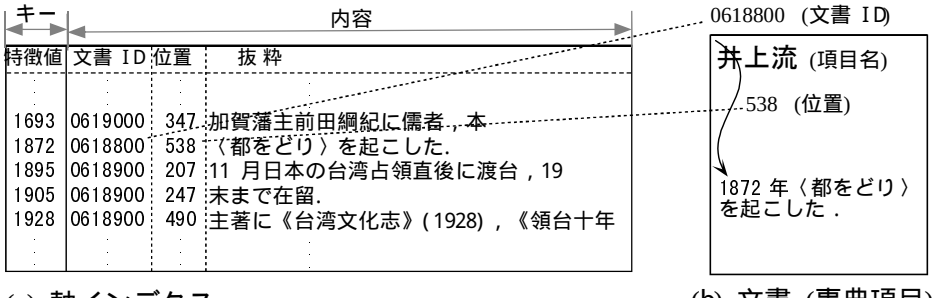
検索結果は，特徴値を第 1 のキーとし，スコアを第 2 のキーとしてマルチキー・ソートされる．したがって，結果は特徴値の昇順または降順に表示され，特徴値がおなじものはスコアの順に表示される．

4. 応用

軸づけ検索の 2 つの応用例についてのべる．

4.1 百科事典検索

軸づけ検索をつかって世界大百科事典 [NEC 95] [HDH 98] 本文全体を検索できるプロトタイプを開発し



(a) 軸インデクス　　(b) 文書 (事典項目)

図 4. 軸インデクスの構造

た．この事典は約 83,000 項目をふくみ，SGML タグをあわせて 160 MB のテキストをふくんでいる．

このプロトタイプにおいては 3 種類の軸を指定できる．それらは，年代，地域，(年代より一般的な) 数量である．年代軸を他の数量軸からわけているのは，年代軸検索においてはそれ専用の情報抽出法をつかうとともに，専用のユーザ・インタフェースをつかうためである．ユーザは特徴値の範囲を西暦年だけでなく一部の和暦年 (平成，昭和など) で指定できるし，単位として「年」だけでなく「世紀」も指定できる．「年」と「世紀」とをわけて検索するのは，これらの検索結果をあわせて表示するとわかりにくいからである．なお，「月」，「日」，「時間」などの単位は百科事典においては「年」，「世紀」ほど重要ではないとかんがえられるので，現在は指定できない．第 2 章で例示したすべての単位とそのほかの単位とがこのプロトタイプのメニューにふくまれる．

地域軸検索のインタフェースと「一揆」を検索語として検索した結果の例を**図 6** にしめす．ここでユーザは入力フレームにおいて軸と検索語を入力する．検索結果は結果リストフレームに表示される．リストフレーム内でユーザがハイパーリンクをクリックすると，事典本文が項目フレームに表示される．抜粋中のハイパーリンクをクリックすれば，本文のその部分が項目フレーム先頭に表示される．

プロトタイプのサーバは Pentium PC の Linux OS 上で動作している．インデクス生成部と検索エンジンはともに Jperl5 によって記述している．クライアントとしては Netscape Navigator や Microsoft Internet Explorer などの Web ブラウザを使用する．検索は CGI (common gateway interface) をつうじておこなう．GNU データベース・マネジャ (GDBM．不揮発性ハッシュ表管理プログラム) をインデクス生成・検索に使用している．

開発に Perl を使用することによって，2 つの利点がある．第 1 はテキストからパターン・マッチで情報をとりだすのが容易に記述できることである．第 2 は GDBM や他のハッシュ表管理プログラムが Perl の連想配列によって容易につかえることである．これによって，軸インデクスと全文インデクスを容易に実装できた．しかし，Perl プログラムはインタプリタで実行されるので，検索エンジンはおそい．したがって，Perl による開発はプロトタイプむきではあるが，実用システムには適さない．

各インデクスの量と登録数を**表 1** にまとめる．各軸イ

表 1．事典検索のための軸づけインデクスのサイズ

インデクスの型	インデクス量 (MB)	登録数
年代 (時間)	12.0[*1] (4.5[*2])	220,536 (年) 6,579 (世紀)
数量	32.3[*1] (14.5[*2])	611,889
地域	1.9 (1.3[*2])	17,224

[*1] これらのインデクスは文書 ID をキーとしてふくんでいる．すなわち，図 4 の構造とは部分的にことなる．
[*2] 抜粋をのぞいた量．

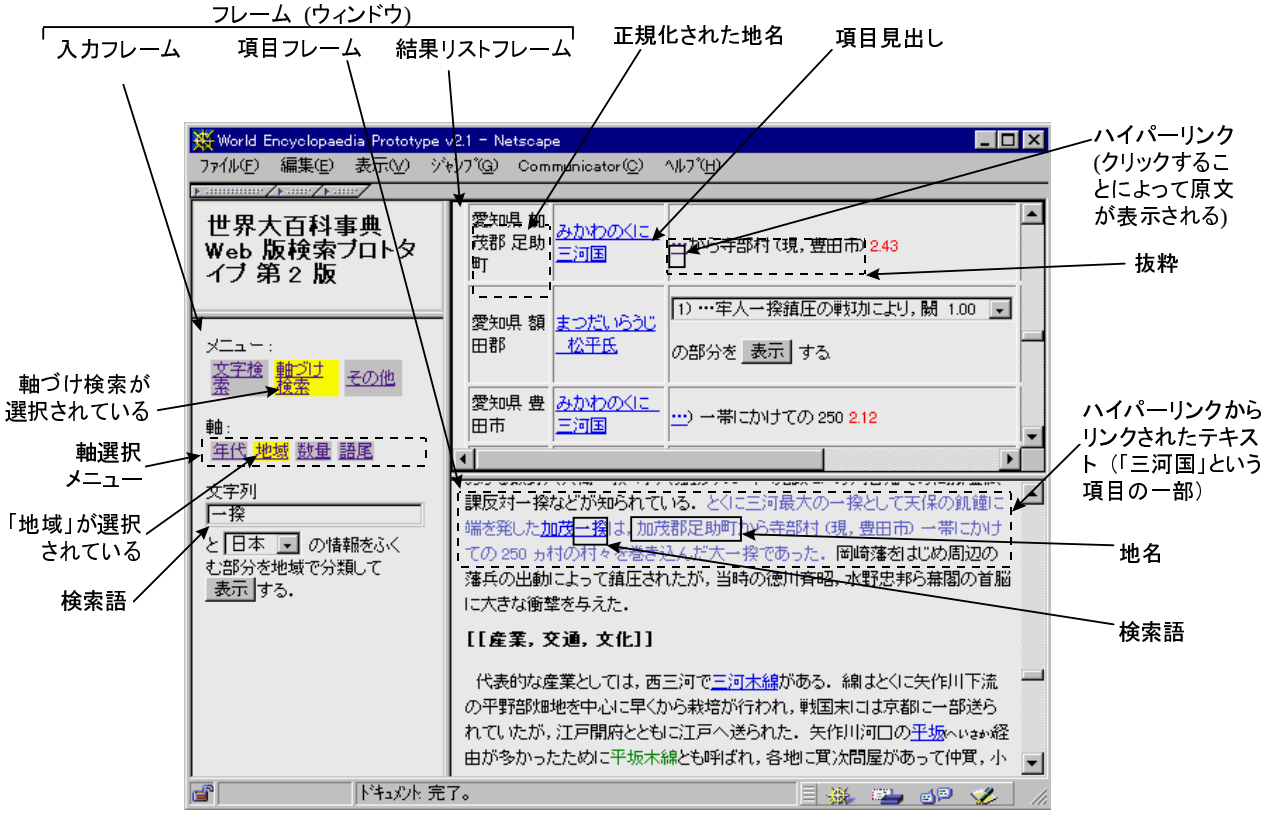


図 6．世界大百科事典の軸づけ検索インタフェースと地域軸検索結果の例 (Netscape Navigator 4 使用)

ンデクスは抜粋をふくんでいる．全文検索には 2 グラム・インデクス [Kik 92] を使用している．全文インデクスの量は，現在 420 MB である (差分圧縮はしていない)．

4.2 新聞記事検索

1995 年に発行された毎日新聞の記事を軸づけ検索するプロトタイプも開発した．記事数は 111,000，テキスト量はタグをふくめて 120 MB である．

このプロトタイプにおいて使用しているプログラムは事典検索で使用しているのと同一またはほぼ同一のプログラムである．軸のメニューも同一にちかく，年代，地域，数量を用意している．しかし，年代軸検索において「月」の情報をあつめている点はことなる．数量軸についていえば，メニューにふくまれる単位はほぼかさなっているが，新聞と百科事典とではおなじ単位でもことなる記法がつかわれている．たとえば，百科事典では「m」というように記号的にしるされているのに対して新聞では「メートル」というようにことばでしるされている．

単位に先行する数値は新聞においては漢字で記述されている (たとえば「一億九千万人」) が，百科事典においてはすべて算用数字で記述されたり，算用数字と漢字がまざった形式 (たとえば「1 億 9000 万人」) で記述されている．例として，半導体に関する年代軸検索のインタフェースと結果の例を**図 7** にしめす．検索語は「半導体」であり，年代範囲は指定していない．

各軸インデクスの量と登録数とを**表 2** にしめす．表 1 とくらべると，年代軸インデクスは百科事典と同程度だが，数量軸インデクスは 3 ～ 5 倍ほどあることがわかる．全文インデクスの量は 360 MB である．

表 2．新聞検索のための軸づけインデクスのサイズ

インデクスの型	インデクス量 (MB)	登録数
年代 (時間)	4.9 (4.4[*1])	64,092
数量	94.3 (73.1[*1])	2,034,330
地域	5.6 (4.0[*1])	184,419

[*1] 抜粋をのぞいた量．

第 2 の例として流域面積の数量検索をあげる．ユーザは世界のおおきな流域面積をもつ川について知りたいと仮定する．検索語として「流域」を入力し，軸として面積の単位 (「m^2」) を選択する．特徴値の範囲は指定しない．これによってつぎのような結果がえられる．

数量	項目見出し	抜粋とスコア
650 万 km^2 ($6.50e+12\ m^2$)	アマゾン[川] Rio Amazônas	…km^2ほどもあって，世界第 1 位である．0.66
	アマゾニア Amazonia	…km^2の広大な地域を占める．0.35
435 万 km^2 ($4.35e+12\ m^2$)	ラ・プラタ[川] Río de la Plata	…km^2で世界 4 位であり，流量はアマゾン川に 1.02
369 万 km^2 ($3.69e+12\ m^2$)	コンゴ[川] Congo River	…km^2におよび，アマゾン川に次いで世界第 2 0.57
324 万 8000 km^2 ($3.25e+12\ m^2$)	ミシシッピ[川] Mississippi River	…km^2に達し，アマゾン川，コンゴ川に次ぎ世界 1.33

アマゾン川については 2 個の記述がみつかり，他の川については各 1 個の記述がみつかっている．数値はもとの記述と正規化された値の両方で記述されている．正規化値はことなる単位をもつ値の比較にやくだつとかんがえられる．ラプラタ川をのぞけば，えられた結果における川の順序は流域面積の順序と一致する．

ここでは，検索結果が川の名称と流域面積を欄とする

図 7．毎日新聞の軸づけ検索インタフェースと年代軸検索結果の例 (Netscape Navigator 4 使用)

表ないし関係データベースのかたちに整理されているとかんがえることができる．しかし，とじた関係を抽出するのが軸づけ検索の機能ではない．このばあい，ユーザはハイパーリンクをたどることなしに 4 位までの川の順位を知ることができるが，ハイパーリンクをたどることによってその情報を確認することができる．情報抽出はつねにうまくいくとはいえないし，抜粋の前後に重要な関連記述があるかもしれないので，それが確認できることは非常に重要である．

5. 関連研究

関連研究としては，Hearst and Plaunt [Hea 93] によるパッセージ検索をつかったサブトピックの検索，武田ら [Mor 95] [Tak 97] による Information Outlining，Nowell ら [Now 96] による Envision という検索結果視覚化システム，Zhu ら [Zhu 97] による WWW 上での部品とサービスを検索する方法などがある．これらについては金田 [Kan 98] がややくわしく記述している．また，斉藤ら [Sai 98] は，軸づけ検索で使用した数値を中心とする情報抽出のより汎用性のある方法についてのべている．

6. 結論

この報告では，軸づけ検索法という，文書集合からの抜粋情報を抽出し整理する機能をもつ全文検索法についてのべた．軸づけ検索によって，ユーザ指定の軸にそって整理された検索結果をえることができるようになった．また，軸づけ検索においては文書中へのハイパーリンクを検索の単位とすることによって，文書よりこまかい意味的な単位でのテキスト検索が可能になった．検索結果にはテキストの抜粋がふくまれているので，それだけで概要を把握できることがおおいが，ハイパーリンクをたどることによってその情報を確認することができる．

軸づけ検索の実装に関しては，軸インデクスと全文インデクスの併用によって高速な検索が可能になった．

今後のおもな課題として，事典や新聞の検索のユーザによる評価，他の種類のテキスト，たとえば WWW やネット・ニュースへの適用があげられる．また，部分語検索に関してはまだ研究の余地がおおきい．

なお，この論文の誤記訂正や内容の改訂は http://www.st.rim.or.jp/~kanada/Papers/search-papers.html#Axis-FI に記述するようにしたい．

謝辞

世界大百科事典のテキストをつかわせていただいた日立デジタル平凡社と，言語処理学会をつうじて新聞記事テキストをつかわせていただいた毎日新聞社に感謝する．

参考文献

[Cut 92] Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W.: Scatter/Gather: a cluster-based approach to browsing large document collections, *15th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, 318–329, 1992.

[Cut 93] Cutting, D. R., Karger, D. R., Pedersen, J. O.: Constant interaction-time scatter/gather browsing of very large document collections, *16th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, 126-134, 1993.

[Fra 92] Frakes, W., and Baeza-Yates, R., ed.: *Information Retrieval: Data Structures & Algorithms*, Prentice Hall, 1992.

[HDH 98] *CD-ROM 世界大百科事典*, 日立デジタル平凡社, 1998.

[Hea 93] Hearst, M. A., and Plaunt, C.: Subtopic Structuring for Full-Length Document Access, *16th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, 59–68, 1993.

[Kan 98] Kanada, Y.: Axis-specified Search: A Fine-grained Full-text Search Method for Gathering and Structuring Excerpts, *3rd ACM Conference on Digital Libraries*, 1998.

[Kik 92] 菊地 他: 日本語文書用高速全文検索の一手法, *信学論* J75-D-I(9), 電子通信学会, 1992.

[Mai 95] *CD 毎日新聞 1995*, 毎日新聞社, 1996.

[Mor 95] Morohashi, M., and Takeda, K.: Information Outlining — Filling the Gap between Visualization and Navigation in Digital Libraries, *Int'l Symp. on Research, Development and Practice in Digital Libraries 1995*, pp. 151–158, Univ. of Library and Information Science, 1995.

[NEC 95] *世界大百科事典*, NEC ホームエレクトロニクス, 1993.

[Now 96] Nowell, L. T., France, R. K., Hix, D., Heath, L. S., and Fox, E. A.: Visualizing Search Results: Some Alternatives to Query-Document Similarity, *19th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, 67-75, 1996.

[Sai 98] 斉藤 公一 他: 数値情報をキーとした新聞記事からの情報抽出, *情報処理学会 自然言語処理研究会 報告* 125-6, 1998.

[Tak 97] Takeda, K., and Nomiyama, H.: Information Outlining and Site Outlining, *Int'l Symp. on Research, Development and Practice in Digital Libraries 1997*, 99–106, Univ. of Library and Information Science, 1997.

[Zhu 97] Zhu, Q., Hu, F., Yao, K., and Will, P.: Searching for Parts and Services on the Web, *Int'l Symp. on Research, Development and Practice in Digital Libraries 1997*, 123–130, Univ. of Library and Information Science, 1997.