情　報　学　基　礎　55－11
デジタル・ドキュメント　19－11
（１９９９．７．１６）

百科事典から動的に年表を生成するテキスト検索法のための年代情報の抽出法と表現法

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

「テーマ年表検索」というテキスト情報検索法を開発した．この検索法においては，あらかじめ文書集合から年代参照を抽出して年代インデクスを生成し，ユーザ入力があると年代インデクスと文単位の全文インデクスとから年代参照と入力された語の出現場所をもとめて結果を年代順に組織化 (ソート) して表示する．検索結果は年代参照とそれをふくむ文，もとのテキストへのハイパーリンクをふくんでいる．この報告ではテーマ年表検索における年代情報抽出法について説明する．この方法を世界大百科事典に適用して評価した結果，大半のばあいに 99% 以上の抽出精度がえられた．また，年月日や世紀など，いくつかの単位がまざった年代表記を効率よくかつすくない誤差でデータ表現する方法について説明する．

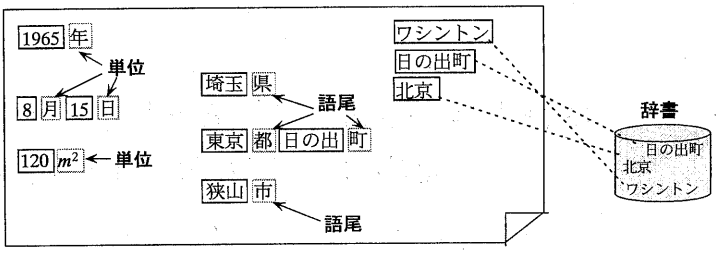
Methods of Extracting and Representing Year References from an Encyclopedia for Chronological-table-generating Text Searching

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

A method of textual information retrieval, which is called the thematic chronological-table search method, has been developed. In this method, an index is generated by extracting and collecting year references from a text collection, the index and a statement-by-statement full-text index are used for searching for year references and search words when the user inputs the words, and the result items are sorted by year and displayed. The result item contains a year reference, a sentence that contains the year, and a hyperlink to the original text. In this paper, the method of information extraction in the thematic chronological-table searching is explained. This method has been applied to a Japanese encyclopedia. An evaluation shows the precision of extraction is higher than 99% in most cases. An efficient and less error-prone data representation for year expression that may contain several units such as century, year, month, day, and so on, are also explained.

1. はじめに

CD-ROM やインターネットの普及にともなって，大量の文書のなかから単純な入力でほしい情報をさがしだすことができ，発見的な検索ができる，あたらしい検索法の開発がもとめられるであろう．このニーズにこたえるために**軸づけ検索法** [Kan 98] [Kan 98a] を開発した．軸づけ検索法においては，ユーザは通常の全文検索と同様にことばを指定するが，それとあわせて，用意されたメニューのなかから軸を選択する．すると，その軸にそって**整理された検索結果**がえられる．また，指定された軸に関して一文書中に複数の話題が記述されているとき，軸づけ検索法ではこれらを分離してとりだせる．すなわち，**細粒度の検索**を可能にしている．



(a) 年月日と数量の抽出　(b) 接尾辞にもとづく地名の抽出　(c) 語種別辞書にもとづく地名の抽出

図 1 軸の指定法 / 3 種の軸づけ検索法

年代軸による軸づけ検索を**テーマ年表検索** [Kan 99] [Kan 99a]，地理軸による軸づけ検索法を**テーマ地図検索**という．日立デジタル平凡社においては，会員制ネットワーク・サービス「ネットで百科」[1] のなかにこれらの検索法をとりいれている．これは，さまざまな軸で世界大百科事典 [HDH 98] を検索できるようにするための第 1 歩だということができる．この報告では，第 2 章でテーマ年表検索の概要を説明し，第 3 章でそのなかの年代情報抽出法について説明する．第 4 章では年代情報抽出における誤抽出の例をしめし，第 5 章で精度を評価する．また，第 6 章で年代のソートに適した内部表現をしめす．

2. テーマ年表検索の概要

テーマ年表検索は，軸づけ検索という，より汎用の検索法の一部として開発した．この章では軸づけ検索法について概説し，テーマ年表検索の機能と実現法をしめす．

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．クラスタリングをもちいた従来の検索結果組織化法においては整理の基準をシステムが選択するが，軸づけ検索においては軸つまり整理の基準はユーザがあたえるので，ユーザが意図したかたちの整理が実現できる．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．テーマ年表検索においては「テーマ年表検索」というメニューが選択された時点で「年代」という軸がさだまる．軸だけでなく軸上の値の範囲もユーザが指定できる．範囲を指定すると，範囲外の結果は削除される．

軸の指定法としては，大別して数量の単位系を指定する方法，語尾を指定する方法，語の種類を指定する方法の 3 つがある (図 1)．単位系を指定する検索法を**数量検索**とよぶ．テーマ年表検索は数量検索の一種である．テーマ年表検索によって，ユーザが希望するテーマに関する年表を動的につくることができる．世界大百科事典のテーマ年表検索においては，約 84,000 項目，SGML タグをあわせて 160 MB という (書誌情報だけでなく) テキスト

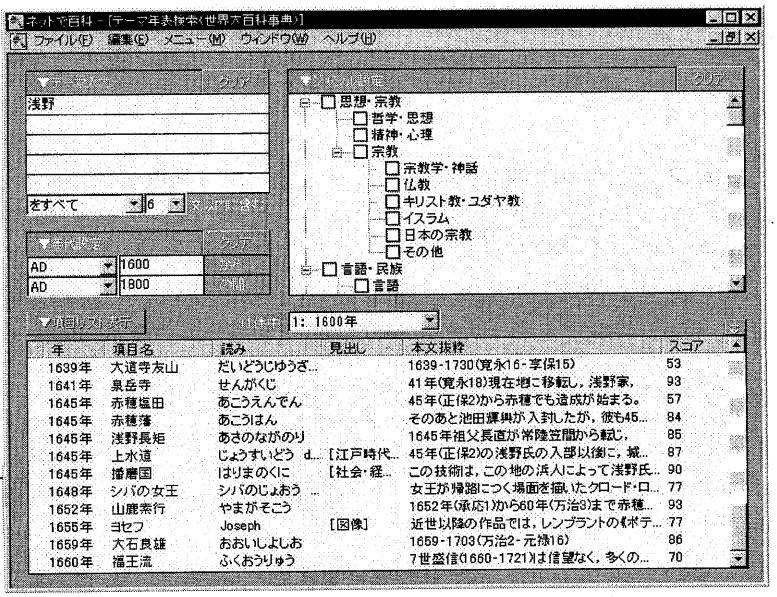


図 2 テーマ年表検索のユーザ・インタフェース例 (「浅野」の検索例)[2]

[1] http://www.hdh.co.jp/information/-net.html

[2] 専用クライアントは日立デジタル平凡社において開発された．Windows 95, 98, NT において動作する．

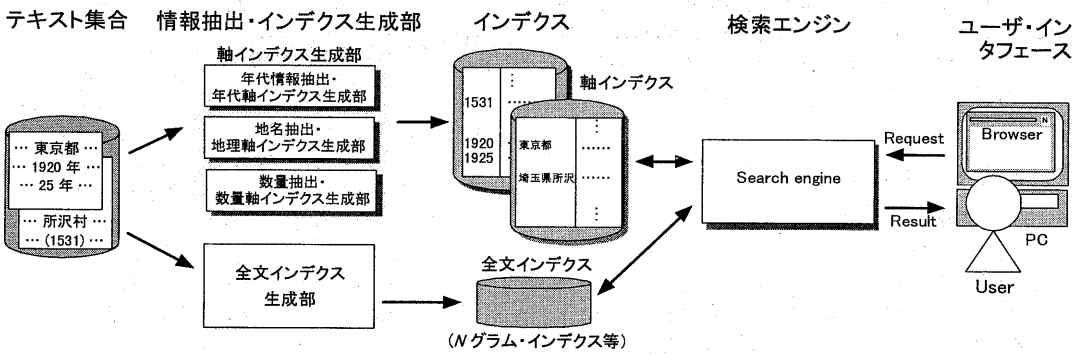


図 3 軸づけ検索のためのシステムの概略構成

全文から，年代表記と検索語とが近接して出現する箇所を検索し，それを年代順にソートして年表の形式で出力する．ユーザ・インタフェースには専用クライアントを使用している(図 2 参照)．

テーマ年表検索においては，基本的に検索質問はつぎの 2 つのくみあわせ (and) で指定される: **(1) 検索語** (and/or 指定可)，**(2) 年代範囲** (西暦/和暦で入力)．(1) だけを指定すれば検索語に関する全年代の情報があつめられれ，(2) だけを指定すればその範囲の全情報があつめられる．これらをくみあわせれば，検索結果をよりよくしぼりこめる．年代の範囲は西暦で指定するのが基本だが，図 2 の「年代設定」のメニューからべつのウィンドウをひらいて「天平 *n* 年」，「平成 *n* 年」などの和暦年で指定することもできる．「ネットで百科」においては，さらにジャンルによるしぼりこみもできる．

各出力項目はテキストから抜粋した文とテキスト原文へのハイパーリンクをふくんでいる．オプション指定によって，抜粋として年代と検索語のどちらの出現をふくむ文を出力するかを指定し (図 2 では年代を表示)，検索する年代の単位として「年」，「世紀」またはその両方が指定することができる．「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，現在は検索対象としていない．図 2 の例においては，ユーザは赤穂浪士周辺の情報を検索するために，「浅野」という語を検索している．

年表の各行にはハイパーリンクがうめこまれている．したがって，各行をマウスでクリックすれば，Web ブラウザによって，抜粋元の文を先頭にして事典項目が表示される．スクロールすれば，抜粋された文の周辺 (その文をふくむ話題の全体) や事典項目全体がみられる．

テーマ年表検索サーバはインデクス生成部と検索エンジンとで構成される (図 3)．インデクス生成部はユーザ要求の発生前に文書集合から年代インデクスと全文インデクスとを生成する．年代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事典全体から抽出し，正規化して年代インデクスに登録する．年代インデクスのための情報抽出法は 3 章で説明する．年代インデクスの使用により検索時間を劇的にへらすことができる．全文インデクス生成部は従来の *N* グラム全文検索と同様の構造のインデクスを生成する．全文検索は文を単位とし，長文は適当にコンマの位置で分割している．「文」の数は約 270 万である．

検索エンジンはユーザによって起動され，年代軸インデクスから指定範囲の年代参照をもつ文を検索し，検索語を全文検索して年代軸検索の結果とマージする．そして，検索結果項目を年代軸にそって整列し出力する．項目ごとにスコアを計算し，スコアがひくすぎる項目はすてる．

検索時に整列が必要な理由はつぎのとおりである．軸が年代だけであれば，全文インデクス生成時に年代順に整列することによって検索時に整列する必要はなくなる．しかし，複数の軸があり，かつ 1 個の全文インデクスをすべての軸に関する検索に使用するばあいは，全文インデクス内部と軸インデクス内部の整列順序をともに原文の順序 (項目順) にしておき，全文検索と軸づけ検索の結果のマージの終了後にソートするほうがよい．

軸と検索語の両方を指定して検索エンジンがよびだされたときは，つぎのようにしてスコアをもとめる (図 4 参照)．検索対象のテキストにおける検索語の出現文番号を全文インデクスからもとめ，年代参照の出現文番号を年代軸イ

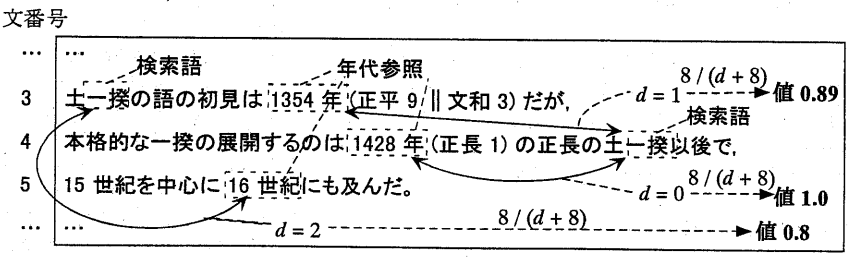


図 4 検索結果の評価

ンデクスからもとめ，それらの差として距離 d をもとめる．検索結果のスコアは d に関する単調減少関数をふくむ．現在使用している関数形は $8 / (d + 8)$ である．検索語や年代参照が複数回あらわれるときは，もっともちかいものを評価につかう．

3. 年代情報抽出法

3.1 年代情報抽出の概要

軸インデクス生成部において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．マッチングは文字レベルでおこない，形態素解析などの自然言語処理はおこなっていない．その理由は，第 1 に数値の抽出においてはあらかじめ形態素解析をおこなうことの利点がすくないこと，第 2 に比較的短期の開発をめざしたため，より軽量の方法をとったことである．しかし，あらかじめ形態素解析をおこなっていればより容易に解析できる部分もある．抽出精度の向上は，百科事典の検索において抽出不正が確認されるごとにパターンの追加，修正をおこなって対処した．

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年代参照もあるが，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数値の正規化法は検索対象テキストの性質にあわせる必要がある．

テーマ年表検索においては，つぎの形式の年代参照を抽出する．

1. 「年」がついた 1 ～ 4 桁の西暦年 (「前」がつくものもふくむ)．たとえば「1989 年」．
2. 「年」がついた西暦年の下 2 桁．たとえば「89 年」．
3. 「年」がついた 1 ～ 2 桁の和暦年．たとえば「平成 10 年」．
4. 「...000 年前」，「... 万年前」，「... 億年前」など．
5. 括弧つきの西暦年．たとえば「ロシア革命 (1917)」．
6. 人名項目における生没年．たとえば「アインシュタイン Albert Einstein 1879-1955」．
7. 「... 世紀」または「前 ... 世紀」．

これらのうち 1～4，7 は世界大百科事典以外のテキストにもほぼそのまま適用できる汎用性がある．しかし，5，6 をそのまま他のテキストに適用すると，ごみをひろう可能性がたかいとかんがえられる．

年代は西暦数値に正規化する．たとえば第 2 の形式において西暦の最初の 2 桁は先行する 4 桁の西暦年を使用しておぎなう．紀元前の年代に関しては，検索システム内部では負号をつけた数値を使用している．紀元 1 年の前年は紀元前 1 年なので 0 は使用しない．

3.2 「年」がつく数値の抽出

数値に「年」がつづくパターンのうち，「年」につづいて「前」があらわれないものの抽出についてのべる．

- **数値のまえに元号がつくもの**
  数値のまえに「大化」から「平成」までの和暦元号がつくときは，これを和暦年とみなして西暦年に換算した値を抽出する．たとえば「昭和 60 年代まで 4 鉱山があった」[1] という表現から昭和 60 年 (1985 年) を抽出する．和暦以外の元号 (たとえば中国暦) も元号とみなすが，年代として抽出していない．[2] 元号とみなす文字列の例をしめす．[3]
  太始，嘉永，建元，元封，元朔，始元，元狩，…，
  顕宗，淳化，麟徳，延宝，天武，隆興，天鳳，康正
  皇紀
  没後，死後，戦後，前後，以後，前後，以来，続く，…，
  王の，周期，平均，炭素，樹齢，林齢，年齢，寿命
- **数値の直前に「後」がつくもの**
  数値のまえにつく「後」は，基本的には「紀元後」という意味でつかわれていると解釈して西暦年とみなし，その数値を抽出する．たとえば「後 70 年にローマ人が第 2 神殿を破壊する」という表現から 70 年を抽出する．なお，例外的に「その後」，「この後」，「帰国後」，「建国後」，「感染後」などがつくときは西暦年ではないものとみなして抽出しない．[4]
- **数値の直前に「前」がつくもの**
  数値のまえにつく「前」は，基本的には「紀元前」という意味でつかわれていると解釈して，数値の符号を反転させたものを西暦年とみなして抽出する．たとえば「前 184 年の〈大カトーのバシリカ〉」という表現から −184 年を抽出する．
- **「年」の直後に範囲を含意することばがつくもの**
  あきらかに期間を含意することばが「年」の直後についているときは，抽出しない．例として「10 年間」があげられる．このように期間を含意しているとみなすことばの例をしめす．
  前半，後半，半，来，目，に [も] わた (る)，に [も] 及 (ぶ)，に一度，に〈数字列〉回，前後の期間，の支配，のあいだ，期，周期，任期，の歴史，の伝統，…，
  単位，がかり，生，祭，加，減，強，弱
  ただし，このようなことばがつかなくても期間を意味していることもおおい．

---

[1] この報告全体にわたって，例題は世界大百科事典第 2 版 [HDH 98] から引用している．

[2] 2 桁の西暦年代の誤抽出をへらすためには，和暦以外の元号も識別する必要がある．

[3] ほとんどの和暦元号は 2 文字だが，例外的に「天平神護」，「神護景雲」など，4 文字のものがある．これらは，処理の都合上，末尾 2 文字を元号とみなしている．

[4] このようなケースを網羅してはいないので，例外的には西暦でないものを西暦とみなして誤抽出するばあいがありうる．「後」が接尾辞であるかどうかは，あらかじめ形態素解析をおこなっていればより系統的に正確に判定できるであろう．

- **数値の直前に年代でないことを含意することばがつくもの**
  このようなばあいは年代参照とはみなさず抽出しない．例として「独立 200 年」があげられる．つぎのような文字がその例である:「齢」(樹齢など),「没」,「ほぼ」,「生誕」．これらのおおくは元号とあわせて処理している (この節はじめの元号の項を参照).
- **数値が 3 ～ 4 桁のとき**
  数値に範囲を含意することばがつづかず，年代でないことを含意することばが先行せず，数値が 3 ～ 4 桁のときはそれを西暦とみなして，その値をそのまま抽出する．
- **数値が 2 桁のとき**
  数値が 3 ～ 4 桁の年表記が出現したのちに，数値が 2 桁の年表記があらわれたときは，先行する年表記の下 2 桁をおとしたものを後続の年表記の頭におぎなう．たとえば，「1960 年」が先行して「64 年」があらわれたときは，これを「1964 年」とみなす．世界大百科事典においてはこれで 99% 以上のばあいにただしい年代をもとめることができる．
- **世紀と年とのくみあわせ**
  「16 世紀の 80 年」というように世紀と 2 桁の年代とをくみあわせた表記は，そのあいだの文字列がかぎられたパターンにマッチするときだけ抽出する．
- **幅のある年代参照のあつかい**
  「1960 年から 80 年」,「60 ～ 80 年」のように幅のある表記においては，幅の開始年 (最初にあらわれる年代参照) をかならず抽出するようにしているが，終了年は抽出していないばあいがある．

3.3 「年前」がつく数値の抽出

数値に「年前」がつづくパターンについてのべる．

- **数値が 1 万未満の端数をふくむばあい**
  2000 からその数値をひいたものを西暦として抽出する．これは，現在の西暦年がほぼ 2000 年だからである．たとえば，「1 万 5000 年前」という表現は –13000 年 (紀元前 13000 年) と解釈する．
- **数値が 1 万未満の端数をふくまないばあい**
  符号を反転させた数値を西暦として抽出する．2000 を加算しない理由は，千の桁は有効でないとかんがえられるからである．たとえば「1 万年前」という表現は –10000 年 (紀元前 10000 年) と解釈する．[1]
- **範囲をふくむばあい**
  「約 1 万～ 1 万 5000 年前」というような範囲をふくむ表現のばあい，ふるいほうの数値を抽出している．この例においては –13000 年を抽出する．

3.4 括弧にふくまれる年代の抽出

括弧にふくまれる年代のパターンについてのべる．括弧内にあっても「... 年前」または「... 年」とおなじパターンにしたがうときは前記のように抽出するので，ここでは「年」がつかない数値の抽出についてのべる．

世界大百科事典のテーマ年表検索においては 57 から 2100 の範囲の括弧内の数値を年代として抽出している．たとえば「ブハラ革命 (1920)」においては 1920 年という西暦年を抽出する．56 以下および 2101 以上の数値を抽出しないのは，これらの範囲の大半の数値が年代以外のものを意味しているからである．とくに，「(1)」,「(2)」などの表現は箇条がきにおいて頻繁に使用される．[2]

3.5 生没年の抽出

個人を記述した項目における生没年の抽出についてのべる．世界大百科事典の SGML テキストにおいて，生没年は専用のタグでかこまれている．したがって，年代抽出時にこのタグでかこまれた部分だけを生没年として抽出する．生没年は不詳のばあいがしばしばあり，'?', 'ころ', 'か', '以前', '以後' などのことばをつかって表現されている．たとえば，つぎのような表現がある．

- 「行信　ぎょうしん　?– 752？(天平勝宝 4？)」→ 752 年を抽出している．
- 「ストラボン　Strabon　前 64 か 63 – 後 23 ころ」→ –64 年と 23 年を抽出し，–63 年は抽出していない．

生年は西暦の全桁が記述されていることを仮定しているが，没年は 2 桁のことがおおいので，生年をつかって補足する．また，数値のまえに「前」がついているときは，負号をつける．

3.6 「世紀」がつく数値の抽出

数値に「世紀」がつづくパターンについてのべる．「世紀」に 1 ～ 2 桁の数字が先行するとき，これを年代参照とみなして抽出する (例: 20 世紀)．数値の直前に「前」がつくものは紀元前の世紀とみなして負号をつける (例: 前 2 世紀)．ただし，年のばあいと同様に，期間を含意することばが世紀につづくときは抽出しない．

3.7 他の年代参照の抽出

前節までにのべてきた以外の年代参照の抽出についてのべる．「年」,「世紀」がつくもの以外にも年代参照が存在する．代表的なものとして「...時代」(たとえば「弥生時代」,「江戸時代」など) がある．これは現在は抽出していない．その理由はつぎのとおりである．[3]

---

[1] 「1 万年前」という表現が紀元前 8000 年ころを意味しているばあいには，あやまった解釈をすることになる (つぎの項目を参照)．また，「1 万年前」という表現と「1 万 1000 年前」という表現とでは後者のほうがふるいとかんがえられるが，抽出情報のうえではこれが逆転するという問題点がある．

[2] 57 ～ 2100 の範囲はいくつかの例外をのぞいて抽出している．この範囲でも年代以外のものがあるので，誤抽出されることがある．例外的に抽出しないのは，括弧の直前に '〉', '市', '町', '村', 'に', 'を', 'は' などがくるばあいである．

[3] これらの問題は解決不可能なものではない．「...時代」以外に

- 「...時代」はある年代範囲をあらわすが，その上下限の決定には高度な判断を必要とする．すなわち，それらの上下限はかならずしも明確でなく，またそれはなにかの知識源 (世界大百科事典じたいでもよい) にもとづいてきめられなければならない．
- 検索結果の表示において，「...時代」ということばと，数値であらわされた年代との関係を明確にしめすのがむずかしい．

4. 年代情報の誤抽出例

前章でのべたように，年代参照の抽出にあたっては誤抽出をさけるためにきめこまかい規則と一部の例外の補足をおこなっている．しかし，それでも誤抽出は完全にはさけられない．この章では誤抽出の例をしめす．なお，下記の例のなかには対策ずみのものもある．

4.1 「年」がつく数値の抽出

- **数値のまえに元号がつくもの**
  「昭和 10 年から 25 年までの 15 年間はその首位を独占してきた」という表現において，25 年は西暦 25 年として抽出される．
- **数値のまえに「後」がつくもの**
  「アメリカで学んだ後 73 年からブッパータール舞踊団で活躍し」という表現において，73 年は 1973 年を意味しているが，西暦 73 年として抽出される．この例においては，意味を考慮しないかぎり「後 73 年」を年代として解釈することが可能である．
- **数値のまえに「前」がつくもの**
  「変動所得については前 2 年の変動所得の平均額を超える額であり」という表現において，「前 2 年」は範囲をあらわすが，西暦 2 年として抽出される．
- **「年」のあとに範囲を含意することばがつくもの**
  「... 年の歴史」，「... 年の生涯」，「... 年の伝統」などの表現からは抽出しないようにしている．このように範囲を含意することばと「年」とのあいだに他のことばが挿入されているときは誤抽出されうる．「ほぼ 100 年のハイネ受容史」という表現においては「史」ということばが範囲を含意しているとかんがえられる．
- **数値が 3 ～ 4 桁のとき**
  「同様に 1000 年に 1 m 以下 10 cm 以上のものを B 級」という表現における 1000 年．
- **数値が 2 桁のとき**
  正規化あやまりの例:「やがて 16 世紀の 80 年代，上野国のほとんどが後北条氏の勢力下に入ったとき」という表現における 80 年は 1580 年を意味しているが，1500 年代の年代表記が先行していないため，ことなる世紀として抽出される．
  正規化以外のあやまりの例:「天武紀 (下) の 10 年 3 月丙戌 (17日) 条に」という表現における「10 年」は和暦でも西暦でもないが，西暦として抽出される．

2 桁の西暦年代参照に関するあやまりをなくすことには，3～4 桁の年代参照にくらべてつぎのような困難がある．後者においてはその年代を抽出しないようにすることで，再現率を犠牲にして適合率をあげることができるが，そうすることによってその年代表現に依存している 2 桁の年代参照の正規化あやまりを増加させ，適合率を低下させる危険がある．たとえば，「1890 年」という年代参照のあとに出現する「1900 年」という年代参照を抽出しないことによって，その後にあらわれる「20 年」は，1920 年を意味しているのに 1820 年に正規化される．

4.2 「年前」がつく数値の抽出

「年前」という文字列が数値につづいていても，その「前」は単語の先頭部分であるばあいがある．たとえば，「1866 年前橘に移封」，「33 年前衛美術家・建築家の集団〈ユニット・ワンUnit One〉を結成し」などの表現がある．形態素解析をおこなっていないときには，このような表現を誤抽出しないように注意する必要がある．

4.3 括弧にふくまれる年代の抽出

つぎのような括弧つき数値が誤抽出されうる．

- 「(011), (100), (101), (110), (111) の $2^3 = 8$ 通りとなる」という表現における 100, 101 などを西暦年として抽出する．
- 「社民党 (161), 左翼党 (22), 緑党 (18), キリスト教民主社会 (15), 自由党 (26), 中央党 (27), 保守党 (80)」という表現 (括弧内は議席数) における 161, 80 などを西暦年として抽出する．
- 「長寿を祝う年祝には，還暦 (61), 古希 (70), 喜寿 (77), 米寿 (88) などがある」という表現における年齢を西暦年として抽出する．

4.4 「世紀」がつく数値の抽出

つぎのような表現が誤抽出されうる．

- 「その近代建築が 1 世紀前後の耐久力しかないとすれば」という表現における「1 世紀」．これは範囲をあらわしているが，そう判定するのは容易でない．
- 「アルタシェス (アルタクス) 朝 (前 190-前 1 世紀，〈大アルメニア王国〉とも呼ぶ)」という表現において「前 190」は年をあらわすとかんがえられるが，「前 190 世紀」として抽出される．

5. 年代参照の抽出精度評価

年代抽出精度評価の結果を**表 1** にしめす．ここでは，限定数の例題について，特定の条件による検索結果の全

は，「...紀」(たとえば「カンブリア紀」)，「...代」(たとえば「古生代」) などがある．これらについても「...時代」と同様のことがいえる．ただし，これらの表現は通常の年代表現と併記されているために，抽出すると冗長になることがおおい．

件について人手で年代情報の正誤を判定している．この評価では大半のばあいに誤抽出は 1% 以下，すなわち精度は 0.99 以上である．しかし，サンプリングの方法によっては 1% をこえるあやまりがみつかる．

表 1 年代抽出精度の評価

検索語	ジャンル設定	年代範囲	年・世紀	文距離	検索結果件数	誤抽出件数	精度
アメリカ	哲学・思想 or 宗教	全域	年だけ	4	819	1	0.999
アメリカ	なし	全域	世紀だけ	0	632	0	1.000
徳川	なし	全域	年	0	653	0	1.000
生命	なし	全域	年だけ	5	832	4	0.995
なし	なし	～前5000 年	年だけ	-	984	0	1.000
なし	なし	2000 年～	年だけ	-	409	6	0.985
なし	なし	1 年～100 年	年だけ	-	276	27	0.902

6. 年代参照の順序づけと内部表現

年代参照をソートするためには，その大小関係を規定する必要がある．また，通常の数値とはことなる性質をもっている年代情報をコンピュータ内部で表現し，ソートのために必要な高速な比較を可能にするために，テーマ年表検索においてはやや特殊な表現をつかっている．この章ではこれらについて説明する．

6.1 年代の順序づけ

年代情報の大小関係はつぎのようにきめている．

- **大小関係は年代の数値にしたがう．**
  年の参照は，テーマ年表検索においては，それが特定の年をあらわすか，範囲をあらわすか，明確であるか，不明確であるか (「ころ」のようなぼかしの表記があるかどうか) をとわず，年の数値によってその大小をきめている．「1960 年代」のような表現においては，範囲の開始年 1960 年だけが抽出されるので，その値を使用する．世紀の表記どうしの大小関係も同様にきめる．
- **範囲の開始 (または終了) 年代も数値の昇順にソートする．**
  範囲の終了年代についても，それが抽出・検索されたばあいには，その数値によって大小関係をきめる．
- **年と世紀との関係**
  テーマ年表検索においては，検索対象を年の表記にかぎるか，世紀の表記にかぎるか，それら両方を検索するかをオプション選択ウィンドウにおいてユーザが選択することができる．ここで両方を検索することを選択すると，両方がまざって出力される．このとき，世紀に関する情報はその世紀の年に関する情報の直前にまとめて出力される．すなわち，世紀 c と年 y との大小関係は，つぎのようにきめている．
  - $y \leq 100(c-1)$ なら c は y よりおおきいとする．たとえば $y = 1900$ (年)，$c = 20$ (世紀) なら，c のほうがおおきい．
  - $100(c-1) < y$ なら y は c よりおおきいとする．たとえば $y = 1901$ (年)，$c = 20$ (世紀) なら，y のほうがおおきい．

上記のような大小関係にもとづいてソートをおこなっているため，つぎのような例においては，意味的な順序と出力順序とでくいちがいが生じるという問題点がある．

- 「1960 年代」に関する検索結果項目は，それが実際には 1968 年に関するものであっても 1960 年に関する結果項目とまざって出力され，1961 年に関する項目よりはまえに位置する．
- 「20 世紀後半」や「20 世紀末」に関する結果項目は「1901 年」に関する結果項目よりまえに位置する．

大小関係の定義をかえることによって，「20 世紀中ごろ」や「20 世紀後半」に関する結果項目をたとえば「1950 年」に関する項目のあとに位置づけることは可能であり，そうすることによって意味的にただしい順序にちかづけることはできる．しかし，そのようにしても正確な順序からはまだほどとおい．また，「20 世紀末」などに関する結果項目は「21 世紀」に関する項目の直前に位置づけることも可能である．しかし，これも，正確な順序を実現するわけではない．したがって，むしろ数値だけで順序をきめる方法のほうが単純でユーザにとってもわかりやすいと判断し，そのようにした．

6.2 年代の内部表現

百科事典に出現する年代・日時の参照としては，「200 億年前」というおおきな値から「1 月 10 日」というような値まである．時刻参照は百科事典の性質上まれにしか出現しないが，基本的にはおなじわくぐみで表現できるべきである．また，ソートの際には「年」による表現と「世紀」による表現とに対して全順序をあたえる必要がある．さらに，この表現で年代をインデクスやメモリに格納するため，表現がコンパクトで，できれば固定長であることがのぞましい．そこで，テーマ年表検索においては年代の内部表現に単精度の 2 進浮動小数の値を対応づけた UniTime という表現をとっている．浮動小数をつかう理由はつぎのとおりである．

- 年代表現を文字列にすると可変長になり，インデクスのあつかいが複雑になる．単精度浮動小数なら固定長であり，4 バイトという最小限の桁数で表現できる．[1]
- 年代表現としておおきな数がつかわれる際には日時のようなこまかい時間は通常は問題にされないため，浮動小数という上位桁だけが表現できる表現法でほぼ十分だとかんがえられる．

[1] 2 バイトではどのような表現にしても桁数が不足する．

UniTime における浮動小数の値と年代表現との対応はつぎのとおりである.

- 基本的に「日」を単位として表現する. たとえば, 1 日は 1.0 によって表現する.
- 1 ヵ月は 32 日 ($32.0 = 2^5$) によって表現する. 1 ヵ月が 31 日未満のときは, その月の 31 日めは日の表現としては使用しない. たとえば, 1 月 0 日 を起点としたとき「2 月 30 日」(62.0) や「2 月 31 日」(63.0) も表現可能だが, 使用しない.
- 1 年は 16 ヵ月 ($16 \times 32.0 = 2^9$) によって表現する. 1 年は 12 ヵ月なので, 各年の 13〜15 ヵ月めは月の表現としては使用しない.
- 各年の 15 ヵ月めを世紀の表現として使用する. たとえば, 「2000 年 15 月」 ($(2000 \times 16 + 15) \times 32 = 1024480.0$) を「21 世紀」の表現とする. これは, 「21 世紀」をソート順で「2000 年 12 月 31 日」と「2001 年 1 月 1 日」とのあいだにおくためである.

上位の単位を下位の単位の 2 のベキ乗で表現するのは, 2 進浮動小数によって表現したときに上位の単位による表現においても誤差が生じにくくするためである.[1] 現在は「日」よりこまかい時間単位 (時, 分, 秒など) はあつかっていないが, これらの単位も表現できる.

7. 関連研究

英語からの情報抽出に関しては多数の研究があるが, 日本語テキストからの数値やその他の情報の抽出に関しても斉藤ら [Sai 98], 山田ら [Yam 99], 佐藤ら [Sat 95], 高尾ら [Tak 99], 久光ら [His 97] などの研究がある. しかし, 年代情報について百科事典の検索に使用できるだけの正確な抽出を実現した研究はなかったとかんがえられる.

8. 結論

テーマ年表検索の年代抽出において, きめこまかい規則 (パターン) を用意し, 誤抽出への対策をほどこすことによって, 世界大百科事典において 99% 以上の精度を実現した. 年代抽出規則のなかにはこの百科事典固有のものもあるが, おおくは, 新聞など, 他のテキストにも適用可能である. また, この検索法においてはユーザ入力にもとづいてインデクスを検索し, えられた結果を年代順にソートする. 秒以下の単位から 100 億年単位までの広範囲の時間を精度よく表現でき, かつ高速なソートが可能な表現法を開発した.

今後の課題として, 年代抽出法に関しては, 年代抽出法にさらにくふうをかさねて百科事典からの抽出精度をたかめるとともに, 現在は精度上の問題などのために抽出していない年代表記を抽出可能にすること, 他の種類のテキストに適用すること (ただし毎日新聞については金田 [Kan 98][Kan 98a] が実験している) があげられる. また, 年代情報の整理に関しては, この報告でしめしたソート順が妥当かどうかという点, 1 次元的なソート以外の整理法などを検討することなどが課題である.

[1] たとえば 1 年を 366 日として表現すると 2 進数で 101101110_2 となるが, 1 年を 32×16 (512) 日として表現すると 1000000000_2 となって, 誤差の影響をはるかにうけにくくなる. 現在は 100 億年単位の年代も年を単位として表現しているが, 「万」= 16384 (2^{14}), 「億」= 2^{28} のような表現をとれば, 100 億年以上の年代表現まで誤差の影響をうけにくくなる.

謝辞

以下の方々に感謝する. 日立デジタル平凡社の藤井氏ほかの方には世界大百科事典のテキスト使用を許可し, 情報抽出法の改良に協力していただいた. 日立東北ソフトウェアの山崎, 澤田両氏には開発した専用クライアント等をつかわせていただいた. 日立製作所ソフトウェア事業部の星氏には全文検索エンジンを改良していただいた.

参考文献

[HDH 98] *CD-ROM 世界大百科事典 第2版*, 日立デジタル平凡社, 1998.

[His 97] 久光 徹, 丹羽 芳樹 : 辞書と共起情報を用いた新聞記事からの人名獲得, *情報処理学会 自然言語処理研究会*, 118-1, 1–6, 1997.

[Kan 98] Kanada, Y.: Axis-specified Search: A New Full-text Search Method for Gathering and Structuring Excerpts, *3rd Int'l ACM Conf. on Digital Libraries*, pp. 108–117.

[Kan 98a] 金田 泰 : 軸づけ検索法 — 文書からの抜粋を抽出・整理して出力する全文検索法, *情報処理学会 情報学基礎研究会*, 98-FI-50-4, pp. 25-32, 1998.

[Kan 99] 金田 泰, 澤田 瑞穂, 山崎 幹夫, 平野 義明, 藤井 泰文 : 「ネットで百科」における「テーマ年表検索」の機能と実現法, *情報処理学会 第58回全国大会* 1J-03, 1999.3.

[Kan 99a] 金田 泰 : 検索結果を地域で整理する百科事典テキスト検索のための地名情報抽出法, *情報処理学会 情報学基礎研究会*, 1999.7

[Yam 99] 山田洋志, 福島俊一 : 数値情報を用いたテキスト検索方式の評価, *情報処理学会 情報学基礎研究会*, 1999.3.

[Sai 98] 斉藤 公一, 迫田 昭人, 中江 富人, 岩井 禎広, 田村 直良, 中川 裕志 : 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*, 125-6, pp. 63–70, 1998.

[Sat 95] 佐藤 円, 佐藤 理史, 篠田 陽一 : 電子ニュースのダイジェスト自動生成, *情報処理学会論文誌*, Vol. 36, No. 10, 2371–2379, 1995.

[Tak 99] 高尾 宜之, 永井 秀利, 中村 貞吾, 野村 浩郷 : 複数製品の紹介記事からの製品情報抽出 — 製品記述パターンの分析 —, *情報処理学会 自然言語処理研究会*, 129-17, 117–124, 1999.