1J-03　「ネットで百科」における「テーマ年表検索」の機能と実現法*

金田 泰[1]　澤田 瑞穂[2]　山崎 幹夫[2]　平野 義明[3]　藤井 泰文[4]

[1]日立製作所 中央研究所
[2]日立東北ソフトウェア
[3]日立製作所 情報システム事業部
[4]日立デジタル平凡社

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ているとかんがえられる．このニーズにこたえるために，我々は軸づけ検索法 [Kan 98] 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．

我々は軸づけ検索法を世界大百科事典 [HDH 98] のテキストに適用する第 1 歩として，会員制ネットワーク・サービス「ネットで百科」のなかに年代を軸とする検索である「テーマ年表検索」の機能をとりいれた．ここではその機能と実現法について報告する．

2. テーマ年表検索の機能

テーマ年表検索は，約 84,000 項目，SGML タグをあわせて 160 MB という世界大百科事典の (書誌情報だけでなく) テキスト全文から，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 (図 1)．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．

検索質問はつぎの 3 つのくみあわせ (and) で指定される: (1) 検索語 (and/or 指定可)，(2) ジャンル，(3) 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，(2) だけを指定すればそのジャンルに関する全年代の情報があつめられ，(3) 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．

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 1 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 1 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．

3. テーマ年表検索サーバの実現法

3.1 システム構成

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，Windows NT 上で動作する (図 2)．

図 1. テーマ年表検索の例:「浅野」の検索

* The functions and implementation method of "subject chronological-table search" in "Net-de-hyakka", by Yasusi Kanada and Yoshiaki Hirano (Hitachi Ltd., email: kanada@crl.hitachi.co.jp), Mizuho Sawada and Mikio Yamazaki (Hitachi Tohoku Software, Ltd.), and Yasufumi Fujii (Hitachi Digital Heibonsha).

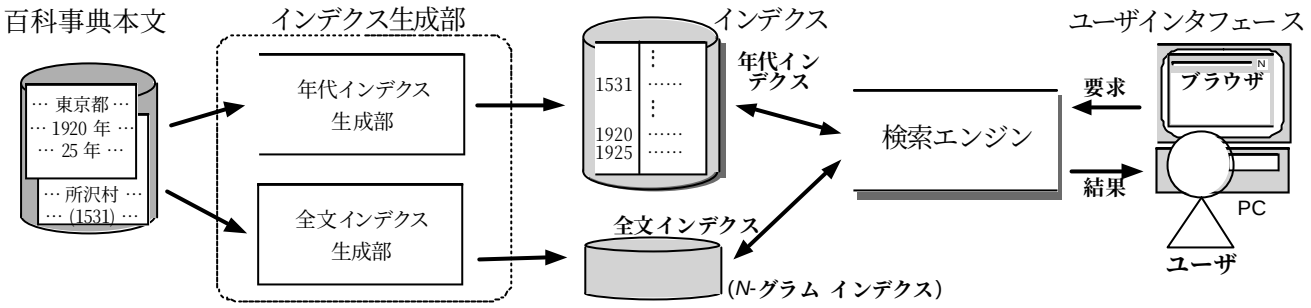

図 2. テーマ年表検索のためのシステムの概略構成

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. 年代インデクスは検索時間を劇的にへらすために使用する. 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ところで分割している. 「文」の数は約 270 万となっている.

検索エンジンはユーザ要求によって起動され, 年代インデクスから指定範囲の年代が出現する文を検索し, 検索語の全文検索をおこなって年代検索の結果とマッチングをとる. そして年代によって結果を整列・出力する.

3.2 情報抽出とインデクス生成

年代インデクス生成部は百科事典の全項目を入力し, 既定の文字列パターンにマッチする文字列を抽出・登録する. おもなパターンはつぎのとおりである.

1. 「年」がついた 1 〜 4 桁の西暦年. 例: 1989 年.
2. 「年」がついた西暦年の下 2 桁. 例: 89 年.
3. 「年」がついた 1 〜 2 桁の和暦年. 例: 平成 10 年.
4. 「..000 年前」, 「.. 万年前」または 「.. 億年前」.
5. 括弧つきの西暦年. 例: ロシア革命 (1917).
6. 人名項目における生没年. 例: 「アインシュタイン」という項目タイトルにつづく「1879 -1955」.
7. 「.. 世紀」または 「前 ... 世紀」.

マッチング・パターンとマッチした年代表記の正規化の方法とはテキストの性質にあわせる必要があるので, 事典用にチューニングして高精度の抽出を実現した. 抽出した年代は西暦数値に正規化し, 年代インデクスに登録する. 文脈独立な規則によって正規化されるものもあるが, 省略された西暦年のように文脈依存のものもある. たとえば 2. において西暦の上位桁は先行する無省略の西暦年を利用しておぎなう. 世界大百科事典では 99% 以上の 2 桁の西暦年はこの方法で正確におぎなえる.

3.3 検索

年代範囲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, 図 3 のようにして検索結果にスコアづけする. 検索対象のテキストにおける検索語の出現位置 (出現文) を全文インデクスからもとめる. 検索単位が文なので, これは容易にもとまる. また, 検索年代の出現位置を年代インデクスからもとめる. これらから語出現と年代出現との距離 x (文の数) をもとめる. 検索結果のスコア関数は x に関する単調減少関数 (現在使用のものは $8 / (x + 8)$) をふくむ. x が一定値以上のためスコアがひくすぎるときは, その検索結果はすてる (x の上限はオプションで指定). 検索語が複数回出現するときは, 年代出現からもっともちかいものを評価につかっている. 検索結果は, 年代をキーとしてソートする.

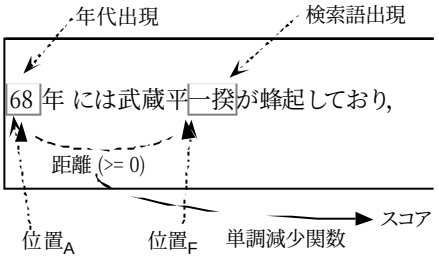

図 3. 検索結果のスコアづけ

4. まとめ

テーマ年表検索をつかうことによって, 文書中にあらわれる年代情報をつかって細粒度テキスト検索結果を整理した年表形式の検索結果がえられる. また, 一文書中に複数の年代に関する情報が記述されているときはこれらを分離してとりだすことができる. 今後, 年代以外の軸による軸づけ検索を実現していきたい.

謝辞

サーバの設置, 運用等で協力していただいた (株) 日立国際ビジネスの三村, 神庭両氏に感謝します.

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.

[Kan 98] 金田 泰: 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, *情報処理学会情報学基礎研究会報告* 98-FI-50-4, 1998.