3P-9　「ネットで百科」における「テーマ地図検索」の機能と実現法*

金田　泰[1]　山崎　幹夫[2]　澤田　瑞穂[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] のテキストに適用し，年代を軸とする**テーマ年表検索** [Kan 99a] とともに，地域を軸とする検索である **テーマ地図検索**の機能を会員制ネットワーク・サービス「ネットで百科」のなかにとりいれた．ここではテーマ地図検索の機能と実現法について報告する．

2. テーマ地図検索の機能

テーマ地図検索は，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，地名と検索語とが近接して出現する文を検索し，それを地域によってソートして表形式で出力する (**図 1**).

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) 地域範囲**. ジャンルを限定しなければ全ジャンルの情報をあつめることができ，地域範囲を限定しなければ全地域の情報をあつめることができる．ジャンルや地域範囲を限定することによって，検索結果をしぼりこむことができる．

図 1 の例においては，ユーザは世界のチーズに関する情報を検索している．この検索によって，世界のどこでどういう種類のチーズがつくられ，たべられているかを把握することができ，「チーズ」という項目をみるだけではわからないさまざまな情報をえることができる．

各出力項目はテキストから抜粋した文とテキスト原文と地図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，抜粋として地名と検索語のどちらの出現をふくむ文を出力するかが指定できる (図 1 では地名を表示).

検索結果の表の行をマウスでクリックすれば，検索された文を先頭にして事典項目が Web ブラウザによって表示される．ブラウザでスクロールすれば，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．また，表の左端の地球マークをクリックすれば，当該の地域をふくむ地図をひらくことができる．

3. テーマ地図検索サーバの実現法

3.1 システム構成

テーマ地図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**).

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地名を抽出して地名インデクスを生成するととも

図 1.「チーズ」の検索 — テーマ地図検索 (日立デジタル平凡社) の例

* The functions and implementation method of "thematic mapping 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).

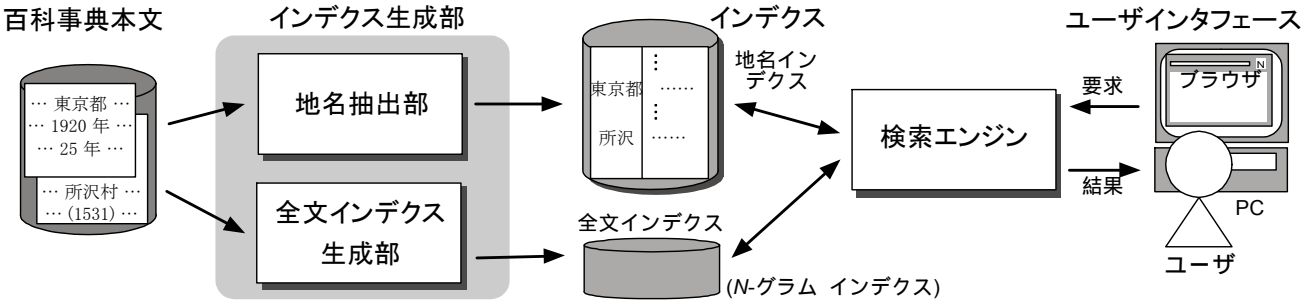

図 2. テーマ年表検索のためのシステムの概略構成

に, 全文インデクスを生成する. 地名抽出部は地名データベースにふくまれる地名にマッチする文字列を事典全体から抽出し [Kan 99a], 正規化して地名インデクスに登録する. 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, 地名インデクスから指定範囲の地名が出現する文を検索し, 検索語の全文検索をおこなって地名検索の結果とマッチングをとる. そして地域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

地名抽出部は, 百科事典の全項目を入力し, 地名データベースに登録された地名とのマッチングをとって地名を抽出し, 抽出地名を同定し正規化して, 地名インデクスに登録する [Kan 99b]. 抽出地名数は日本地名が約 13 万, 世界地名が約 34 万 (重複あり) である.

地名マッチングについて説明する. 既知地名だけを抽出するおもな理由は, テーマ地図検索において未知地名は検索結果の整理にも地図表示にもつかえないので価値がひくいことである. 地名データベースは世界大百科事典 [HDH 98] の地図のために開発されたものである. マッチングは最長一致法による. マッチした文字列の前後のテキストをしらべて, それを地名として抽出するべきかどうかを判定している. この局所的な文脈マッチは文字列単位でおこない, 形態素解析は使用しない.

つぎに, 抽出地名の同定・正規化について説明する. たとえば「コロンバス」(全米で 6 箇所以上) のように同名の地名が複数あれば, 文脈をみて地名を同定する. 部分的な地名や別名は標準の地名に正規化する. たとえば「プエルトリコ」は「米領プエルト・リコ」に変換する.

3.3 検索

地域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, テーマ年表検索 [Kan 99a] と同様の方法で検索結果にスコアづけする. すなわち, まず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, これは容易にもとめられる. また, 地名の出現位置を地名インデクスからもとめる. これらから検索語出現と地名出現との距離 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. スコアがひくすぎるときはその検索結果はすてる. 検索語が複数回出現するときは, 地名出現からもっともちかいものを評価につかっている.

検索結果は地名の登録順にしたがってソートする. すなわち, 登録順序をきめることでソート順をきめている. 日本の地名に関しては県単位にはほぼ北から順に整列し, 海外の地名に関しては国単位に 50 音順に整列している. また, 県内, 国内は 50 音順に整列している.

4. まとめ

テーマ地図検索をつかうことによって, 文書中にあらわれる地名情報をつかって細粒度テキスト検索結果を整理した表形式の検索結果がえられる. 一文書中に複数の地域に関する情報が記述されていれば, それらを分離してとりだせる. 今後は金田 [Kan 98] で実験した地名, 年代以外の軸による軸づけ検索を実用化していきたい.

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.

[Kan 99a] 金田, 山崎, 澤田, 平野, 藤井: 「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-3, 1999.

[Kan 99b] 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, 情報処理学会自然言語処理研究会報告, 99-NL-132-2, 1999.