インターネット・タウンページの構築 (2) ～あいまい検索技術～*

7N－8

高橋 克巳† 三浦 信幸‡ 島 健一§

NTTソフトウェア研究所
E-mail: {takahasi, miura, kshima}@slab.ntt.co.jp

1 はじめに

インターネット・タウンページ[1]で使われている検索技術について述べる．NTTの職業別電話帳(タウンページ)情報をコンテンツとするWWWサーバを構築し，情報検索などいくつかの実験を行なっている[1]．

電話帳情報の問い合わせは多種多様であり，ユーザフレンドリーな検索を実現する必要がある．本論文では本サーバーで使われている検索技術について説明する．インターネット・タウンページの検索は，(1)文字コード／字種の自動判定，(2)シソーラス参照，(3)ストップワード除去，(4)文字列の正規化による近似文字列照合，(5)適合度評価，(6)キーインデックス作成(語幹抽出など)，などの機能を有している．本論文では(4)の近似文字列照合を中心に説明する．この近似照合により，人名／会社名や地名などの固有名詞の表記に存在する，漢字の置換え，読みのゆれ，発音の類似，カタカナ語の表記のゆれなどによる検索もれを減じることができる．

2 インターネットタウンページの検索の概要

図1に本システムのデータベース構成の概念図を示す．「職業名」「エリア」「会社名」の3つのフィールドを限定した検索が可能である．それぞれ文字列入力が可能で，前2者のフィールドには選択メニューも用意されている．また位置座標情報(緯度経度)は所在の位置から情報を検索するためのものである．位置座標は用意された地図をクリックすることで入力可能である．

図2に本システムの検索の概要を示す．サーバがユーザからの問い合わせを受けると，サーバはCGIプログラムを呼びだし以下の処理を行なう．(1)では日本語コード判定をするが，この時いわゆる半角のカタカナは全角に変換する．ひらがなとカタカナの区別は行なわない．(2)のシソーラス参照は「職業」と「エリア」に関して行なう．(5)の適合度評価では検索結果を問い合わせに近い順に並び替える[2]．(6)のキーインデックス作成で

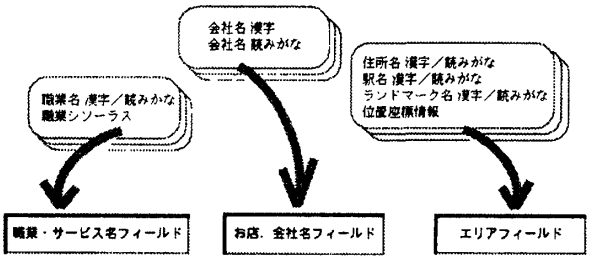

図1: データベース構成の概念図

は，部分文字列による検索要求に対応するために，「レストラン-○○」といった接頭語や「○○-商店」といった接尾語を取り除くなどの処理を行なっている．(4)の正規化は後の章で説明する．

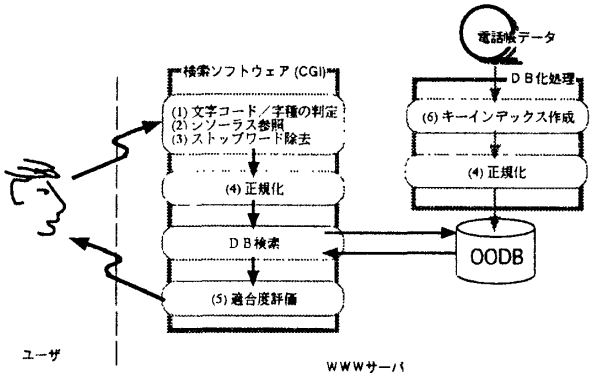

図2: 検索の概要

3 あいまい検索

あいまい検索という言葉は，以下のようないくつものの異なった検索方法を包含している[3]．本稿では主に1をターゲットにした検索方法について説明する．

1. ユーザの表記方法にゆれがある検索
   「やまざき」と「やまさき」の濁音清音のゆれや，「斎藤」と「齋藤」の異体字の関係など．

---

*Internet Townpage (2) - Advanced search methods -

†Katsumi Takahashi, NTT Software Laboratories

‡Nobuyuki Miura

§Ken'ichi Shima

[1] http://townpage.isp.ntt.co.jp/

[2] この機能は英語版TOWNPAGEでのみ実装されている．

[3] "完全一致" 以外の全ての検索を指して使われているようでもある．

2. ユーザの記憶／知識が不確かである検索
「○○デパートの隣」のように条件を正確に表現できないことなど.

3. 意思決定を伴う検索
「おいしいレストランに行きたい」のように，対象が具体的でないもの.

文字列や発音のゆれという現象の解消は，文字列の誤り訂正と関連した問題であり，複数の方法が知られている [2]. しかし文字列検索で起こり得る誤りは，一般の脱落，置換えに比べ現象を限定しやすいので，言語情報などを使って訂正する範囲を限定し，かつ高速な近似照合を実現する方法が知られている. 例えば英語の類似発音語を検索するための Soundex[3] アルゴリズムが知られており，現在では市販の DB にも組み込まれている[4]. 筆者らは日本語の固有名詞の表記のゆれに対する検索の研究 [4] を行なっているが，その対象を拡張し，実システムへの適用を行なったのが次の章で述べる近似文字列照合である.

4 文字列の正規化による近似文字列照合

本章では，文字列の表記のゆれによる検索もれを解消するための正規化処理について説明する. ゆれの問題に対処するには，表記を統一して検索すること，すなわち，問い合わせの文字列とデータベースに登録する文字列に表記を統一するための同一の変換処理を適用して照合を行なうことが一般的である. この変換処理を正規化と呼ぶ. 正規化を使う第一の理由は，照合の前段階で変換を行なうのでデータベース処理に余計な時間をかけない点である. ここで注意すべきことは，統一すべき表記の選び方，および，検索する際に表記を統一する方法の 2 点である. 以下順に説明する.

4.1 日本語固有名詞のゆれ

日本語表記のゆれには以下のようなものがある.

(1) カタカナとひらがなの使い分け，(2) 漢字と読みがなの使い分け，(3) 拗音，促音などの小書きの文字 (ウエア と ウェア)，(4) 読みがなのゆれ (アメミヤ と アマミヤ)，(5) 漢字のゆれ (斉藤 と 齊藤)，(6) カタカナ語 (外来語) の表記 (ヴァイオリン と バイオリン)

文献 [4]，[5] および，電話帳に頻出する固有名詞の情報の解析などから統一すべき表記ゆれの組を 825 組選んだ. 表 1に例を示す.

4.2 正規化処理

前節で選んだゆれの組の同値関係から，正規化を行なうための正規化規則を作成する. しかしこの正規化規則を正しく作成する作業は，1 つの文字列が 2 つ以上の文字列に書き換えられる可能性があるので容易ではなく，機械的操作で求められることが必要である. 同値関係から変換規則を求めるアルゴリズムは Knuth-Bendix の完備化アルゴリズム [6] として知られている. このアルゴリズムは，1 つの文字列が異なる 2 つの文字列に書き換えられる場合，この 2 つの文字列が等しくなるような規則の追加を繰り返すことである. このアルゴリズムを使って正規化規則を作成した. 実際に 825 組のゆれの同値関係から 730 組の正規化規則が求められた. この規則に基づいた正規化を行ない，近似文字列照合を行なっている. この正規化による検索は，読みなどのゆれがあっても，もれのない検索を高速に実行すること可能としている. また，必要に応じてゆれの組を決めるだけで，近似照合ができるため実装が容易である.

漢字のゆれ	沢	澤
読みがなのゆれ	ズ	ヅ
	アマ	アメ
	サカ	ザカ
	トウ	トオ
カタカナ語の表記	リヤ	リア
	レーン	レイン

表 1: 統一すべき表記のゆれの組の例

5 おわりに

本手法を WWW のユーザログから評価する予定である. また今回実現した，地図で位置を指定し検索する方法のように，よりあいまいなあるいは要望の高い問い合わせに対する検索技術を開発する予定である.

インターネットは電子的な社会であると言われている. サーチエンジンなどでも実際の固有名詞の問い合わせが多いと言われており，電話帳のようなデータベースは，実社会とインターネットの架け橋になり得る存在である. 今後サーチエンジンを含めた各種検索への適用を含め検討を行なう予定である.

参考文献

[1] 島健一，高橋克巳，三浦信幸 インターネット・タウンページの構築 (1) ～概要～，情報処理学会 第 54 回 全国大会，(1997).

[2] K. Kukich (金岡恭訳)：単語文字列の自動訂正技術，コンピューターサイエンス，ｂｉｔ別冊，pp.121-174(1994)(日本語訳)

[3] M K.Odell, R C.Russell· U S.Pat, 1 261 167(1918). 1 435 663(1922)

[4] 高橋克巳，梅村恭司：人名のかな表記のゆれに基づく近似文字列照合法，情処論，vol.36, No 8,(1995).

[5] 国立国語研究所：現代表記のゆれ，国立国語研究所報告 75，秀英出版 (1983).

[6] D E Knuth, P G.Bendix Simple Word Problem in Universal Algebras, Computational problems in abstract algebra, pp 263-297, Pergamon Press(1970).

[4] 例えば Oracle ConText Option にその記述がある.