中間部分一致検索における単語区切り位置情報の検索精度に対する効果

奥 雅博、永井良史、野田良輔
NTT情報通信研究所
〒249-0847 横須賀市光の丘1-1
{oku, nagai, noda}@isl.ntt.co.jp

1．はじめに

我々はプッシュボタン(PB)信号送出可能な電話機(PB電話機)を入力端末として利用できるPB入力型電話番号検索システムの開発を進めている(東田 1997)(東田他 1998)。このシステムは、家庭やオフィスに普及しているPB電話機を用いて住所や名前の入力を可能とする日本語入力方式(以下、PB入力方式)(佐藤他 1997)をもとに、入力情報を用いてデータベースを検索する技術(奥他 1997b)、HMIに関連する対話誘導技術(奥他 1997a)、および音声応答技術などから構成されている。

PB入力方式は、図1に示すように、1つのPBボタンに複数のかな文字を対応させ、1ストロークで1かな文字を入力させる方式である。例えば、「トウキョウ」と入力する場合には、“41281”と順に押下する。従って、1ストロークごとを見るとかな文字レベルで複数の候補が存在することになる(“1”の押下は“あ”～“お”の5つの文字のどれかを入力したことになる)。この曖昧さの解消を住所や名前の実在性によって、すなわちDB検索を行うことによって解消していく。

図1：PBボタンへの文字の割り当て例

一方、電話番号検索では利用者からの不完全な入力に対しても何らかの解を求めるため(再現率を高めるため)に、利用者が入力した検索キーと一致するあるいは検索キーを一部に含む情報を効率よくしかも高速に検索する部分一致検索手法が要求される。

我々は、PB入力方式によって入力された住所や名義などの検索条件をもとに電話番号を検索するPB入力型電話番号検索システムにおいて、部分一致検索を効率よく高速に実行する手法について提案し、その速度面および性能面から見た評価を実施した(奥他 1997b)。この手法は、全文検索の分野で使用されているPAT木における半無限文字列(sistring)(Frakes 1992)(菊田 1996)に似たインデックスを派生させ、それを市販のDBMSによって管理することによって、DBMSの持つインデックスを利用した検索の高速性を損なうことなく、部分一致検索を行うことができる。

しかし、部分一致検索では、入力の一部を含む候補を検索するので、その検索結果には検索ノイズが含まれる。我々は、この検索ノイズを低減するために、検索結果である文字列の初めと終わりとがそれぞれ単語区切り位置に一致するものだけを候補とする手法を提案した(永井他 1997)。本稿ではこの単語区切り位置チェックが検索精度に及ぼす影響を定量的に評価する。

2．用語の定義

○検索キー

データベース検索を行う際の照合文字列。

○インデックス、インデックスレコード

インデックスレコードの集まりをインデックスと呼ぶ。インデックスレコードは名義部とポインタ部とから構成される。名義部はデータベース検索を行う際に検索キーと照合される被照合文字列であり、ポインタ部はデータベース本体情報を指し示すポインタである。以下、インデックスレコードの名義部を単にインデックスレコードと呼ぶこととする。

○部分一致

検索キーの一部がインデックスレコードに含まれている場合、そのインデックスレコードに“部分一致”したという。例えば、検索キー＝「電話」は、インデックスレコード「日本電話」、「電話会社」や「電信電話会社」に部分一致する。また、検索キー＝「京都」は、インデックスレコード「京都／府」の他に「東京／都」のように単語区切り位置をまたぐものにも部分一致する。このような部分一致は検索の適合率の低下を招く1つの要因となるが、検索キーと一致している部分が単語区切り位置に一致しているか否かをチェックすることによって、「東京／都」を検索結果から削除することができ、検索ノイズを低減することができる。

○前方部分一致

部分一致のうち、検索キーがインデックスレコードの先頭から一致しているものを“前方部分一致”と呼ぶ。検索キー＝「電話」は、インデックスレコード「電話会社」に対して前方部分一致する。

○中間部分一致

部分一致のうち、検索キーがインデックスレコードの先頭ではない部分から一致しているものを“中間部分一致”と呼ぶ。検索キー=「電話」は、インデックスレコード「日本電話」や「電信電話会社」に対して中間部分一致する。

○適合率

情報検索の指標の１つで、ある検索条件によって得られた結果の中に含まれる正しい結果の割合を適合率という。

○再現率

情報検索の指標の１つで、ある検索条件によって得られるべき結果（正しい結果）のうち、実際に得られた結果の割合を再現率という。すべての情報を取得すれば再現率は１００％であるが、適合率が非常に小さくなる。

３．DB検索手法の概要

中間部分一致検索を高速に行うために、元のインデックスレコードからPAT木で用いられている半無限文字列と同様な文字列から派生インデックスをあらかじめ生成しておく。そして、この派生インデックスに対して、入力された検索キーで前方部分一致検索による照合を行い、検索候補を取得する。次に、検索候補に対して単語区切り位置チェックを行い、不要な検索候補を削除したものを検索結果とする。

３．１　DBにおけるインデックスの派生

本節では派生インデックスの生成手法について述べる。図２にインデックス派生の例を示す。

元のインデックスレコード「546*04*0304*0026*322*138（日本電信電話株式会社）」を形態素解析すると、「546*0（日本）／4*030（電信）／4*00（電話）／26*32（株式）／2*138（会社）」（／は単語区切り位置）であるので、元のインデックスレコードの先頭から１単語ずつ削除した「4*0304*0026*322*138（電信電話株式会社）」、「4*0026*322*138（電話株式会社）」、「26*322*1

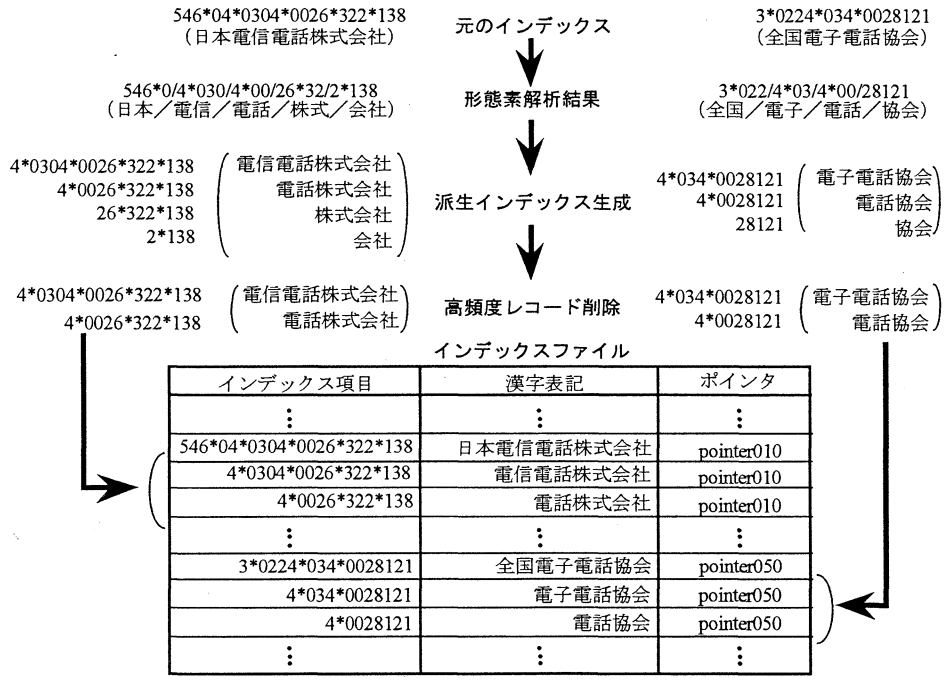

図２：　派生インデックスの生成

38（株式会社）」、「2*138（会社）」の４つが派生対象となる。ここで、「26*322*138（株式会社）」、「2*138（会社）」の２つは高頻度に現れるため、インデックスレコードとしては派生させないものとする。従って、「4*0304*0026*322*138（電信電話株式会社）」、「4*0026*322*138（電話株式会社）」の２つをインデックスレコードとして派生させる。「3*0224*034*0028121（全国電子電話協会）」の場合も同様にして「4*034*0028121（電子電話協会）」、「4*0028121（電話協会）」の２つをインデックスレコードとして派生させる（「28121（協会）」は高頻度とする）。

これにより、元のインデックスを構成する単語の先頭から成る文字列を派生インデックスとして生成することができ、これに対する前方一致検索は元のインデックスに対する中間部分一致検索と同等の結果を与える。

３．２　DB検索の流れ

図２のように生成した派生インデックスに対して以下の２段階の処理によって検索キーに対するDB検索結果を取得する：

a．派生インデックスに対して検索キーによる前方一致検索を行い、検索候補を取得する。

b．検索キーの最終文字位置が検索候補の文字列中の単語区切り位置と一致しているか否かをチェックし、一致しているもののみを検索結果とする。

以下にいくつかの検索例を示す。

（１）「4*00（電話）」での検索

a．「4*00%（電話%）」として前方一致検索することによって、派生インデックスレコード「4*0026*322*138（電話株式会社）」と「4*0028121（電話協会）」の２つが検索候補として得られる。

b．次に単語区切り位置をチェックする。検索候補の単語区切り位置は、それぞれ、「4*00/26*32/2*138（電話／株式／会社）」（／は単語区切り位置）と「4*00/28121（電話／協会）」であり、検索キー「4*00（電話）」の終わりはそれぞれの単語区切り位置に一致している。従って、検索キー「4*00（電話）」に対する検索結果としては、それぞれの派生インデックスの元である「546*04*0304*0026*322*138（日本電信電話株式会社）」と「3*0224*034*0028121（全国電子電話協会）」が得られる。

（２）「4*0304*00（電信電話）」での検索

a．「4*0304*00%（電信電話%）」として前方一致検索することによって、派生インデックスレコード「4*0304*0026*322*138（電信電話株式会社）」が検索候補として得られる。

b．次に単語区切り位置をチェックする。検索候補の単語区切り位置は、「4*030/4*00/26*32/2*138（電信／電話／株式／会社）」であり、検索キー「4*0304*00（電信電話）」の終わりは単語区切り位置に一致している。従って、検索キー「4*0304*00（電信電話）」に対する検索結果としては、派生インデックスの元である「546*04*0304*0026*322*138（日本電信電話株式会社）」が得られる。検索キーの最終文字位置と検索候補の単語区切り位置とが一致しているか否かのみをチェックするので、この場合のように検索キーが複数の単語から構成されていても所望の検索結果を得ることができる。

（３）「4*03（電子）」での検索

a．「4*03%（電子%）」として前方一致検索することによって、派生インデックスレコード「4*0304*0026*322*138（電信電話株式会社）」と「4*034*0028121（電子電話協会）」の２つが検索候補として得られる。

b．次に単語区切り位置をチェックする。検索候補の単語区切り位置は、それぞれ、「4*030/4*00/26*32/2*138（電信／電話／株式／会社）」と「4*03/4*00/28121（電子電話協会）」であり、検索キー「4*03（電子）」の終わりと一致する単語区切り位置を持つのは後者だけである。従って、検索キー「4*03（電子）」に対する検索結果としては、後者の派生インデックスの元である「3*0224*034*0028121（全国電子電話協会）」のみが得られる。このように、単語区切り位置をチェックすることによって、単語の途中までが検索キーと前方一致している候補を検索ノイズとして排除することができる。

ここで述べた手法では、検索キーが、インデックス中の文字列のうち、単語の先頭から始まって単語（同じ単語とは限らない）の終わりまでの文字列と一致する場合のみを検索結果とする。本稿では、この手法の精度面から見た有効性を評価する。

４．精度評価実験

４．１　実験方法

（１）検索キーの準備

０～９の１０個の数字を使って、文字列長１０桁のランダムな数字列をユニークに１００個作成する。次に９桁の数字列を１００個作成する。以後同様にして４桁の数字列１００個までを順次作成し、合計７００個の数字列を作成する。これらを検索キーとする。

（２）検索対象DB

インデックス派生を行った後の北海道地区の企業名DB（インデックス総数約３１６万

件）を対象とする。なお、検索対象インデックスは清音化された企業名義とする。

（3）検索の種類

（3－a）%key%による中間部分一致検索（%はワイルドカードキャラクタ）。

（3－b）key%による派生インデックスに対する前方部分一致検索。

（3－c）（3－b）の結果のうち、検索キーの終わりがインデックス中の単語区切り位置と一致しているものだけを抽出。

（4）検索方法

（1）で作成した検索キー７００個のうち、数字列長の長いものから順に（2）のDBを（3－a）中間部分一致検索して検索候補を取得する。そして検索候補の累計が５０００件を越えたら、そのときに条件としている検索キーまでで検索を終了する。

次に、（3－b）派生インデックスの前方部分一致検索を実施する。検索キーは（3－a）中間部分一致検索で使用したものと同一のものとする。最後に（3－c）の検索も同様に実行する。

（5）再現率、適合率の計算

まず、（3－a）～（3－c）の検索結果に対して３人によって正解集合を作成する。なお、検索キーによってその検索結果が得られてもおかしくないものを正解とし、３人の意見が分かれたときは合議によって判断する。次に、（3－a）の検索結果を再現率１００％（（3－a）の結果の中に検索すべきものがすべて入っている）と仮定して（3－a）～（3－c）の再現率、適合率を計算する。

4．2　評価実験結果

表１に実験結果を示す。表１から明らかなように、適合率が１０％以下から９０％弱にまで大幅に増加している。このときの再現率の低下は２％強にとどまっており、精度面から見て、単語区切りチェックを行うことによって、必要な候補をあまり落とさずに検索ノイズを大幅に削減できることがわかる。

表１：　評価実験結果

	再現率［%］	適合率［%］
中間部分一致	１００	８．２
前方部分一致	９９．３	３６．５
区切り位置チェック	９７．６	８８．８

4．3　考察

若干ではあるが再現率が低下している要因は、単語区切り位置に一致しなくとも検索結果として抽出すべきものがあることを示している。実例を見てみると、「英語／ゼミナール」に対して検索キーが「英語／ゼミ」や、「××／コスメティック」に対して検索キーが「××／コスメ」などのように、単語の先頭数文字が略語として通用しているものがある。このような例は、全文検索等のように単語を単位として検索する場合にも検索もれとなってしまう。

このような例をもらさずに検索するためには、略語としての登録が必要である。3節で述べたインデックス派生の場合には、略語として通用する部分に単語区切りと同様な働きをする略語区切りを設けて、略語区切り位置からもインデックスを派生させる。これによって再現率の低下をさらに小さく抑えることができる。

5．おわりに

本稿では、部分一致検索における検索ノイズの低減手法である単語区切り位置チェックが検索精度に対してどのような影響を及ぼすかについて評価実験を通して検証した。この結果、再現率約９８％、適合率約８９％という高い検索精度が得られ、単語区切り位置チェックが検索ノイズの低減に対してきわめて有効であることがわかった。

［参考文献］

（Frakes 1992）Edited by William B. Frakes and Ricardo Baeza-Yates: Information Retrieval, Prentice Hall PTR (1992).

（東田　1997）Masanobu Higashida: A Fully Automated Directory Assistance Service that Accommodates Degenerated Keyword Input Via Telephone, Proc. of PTC'97, pp.167-174 (1997).

（東田他　1998）東田、村上、奥：　オペレータレス自動電話番号検索システムの開発、情処学会研究報告、NL-123-4 (1998).

（菊田　1996）菊田昌弘：　用語解説：パトリシアツリー(Patricia Tree)、人工知能学会誌、Vol.11, No.2, pp.337-339 (1996).

（永井他　1997）永井、林、野田：　文字区切り・単語区切りを用いた検索解の絞り込み効果の検討－PB電話機を利用したデータベース検索への応用－、97信学会総合大会、D-6-8 (1997).

（奥他　1997a）奥、林、永井、東田：　PB電話機を利用した電話番号案内方式に適した対話誘導戦略、97信学会総合大会、D-6-7 (1997).

（奥他　1997b）奥、野田、林：　形態素解析を用いた中間部分一致検索の高速化手法、情処学会研究報告、NL-121-9 (1997).

（佐藤他　1997）佐藤、東田、林、奥、村上：PB電話機を利用した日本語入力方式、97信学会総合大会、D-6-6 (1997).