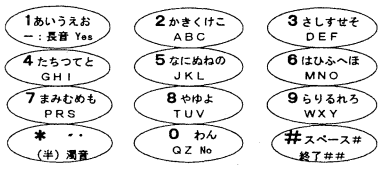
部分一致検索における単語区切り位置情報インデックス化による高速化とその効果

永井　良史、奥　雅博
NTT 情報通信研究所
〒239-0847 横須賀市光の丘 1-1
{nagai, oku}@isl.ntt.co.jp

１．はじめに

電話帳掲載名義をプッシュボタン電話機(以下PB 電話機）を使って検索する電話番号検索システムを構築している(東田　1997)(東田他　1998)。このシステムでは、利用者は音声による対話誘導ガイダンス(奥他　1997)に従って、PB 電話機を用いた入力技術である日本語入力方式(佐藤他　1997)をもとに入力を行う。しかし、利用者が入力できる情報は曖昧であることが多く、そのような曖昧な情報を利用者が入力した場合にも、高速でしかも、検索ノイズの少ない検索を行うことが求められる(奥他 1998)。

PB 電話機を用いた入力技術である日本語入力方式では、プッシュボタンに下図に示す文字の割り当てを行い、「とうきょう」の入力では、「４１２８１」と入力する（図１）。



投入例：とうきょうと → ４１２８１##

図１．ＰＢボタンへの文字の割り当て

この日本語入力方式では、「とうきょう」の「と」のPB 数字列は４となり、「たちつてと」のどれを入力したのかを特定することはできない。そこで「とうきょう（PB 数字列では４１２８１）」というPB 数字列の実在性をDB 検索することによって、曖昧性の解消を図っていく。

また、実際の電話番号検索システムでは、利用者の不完全な入力に対しても何らかの検索解を見つけることが求められるため、部分一致検索が必要となる。しかし、部分一致検索では、検索ノイズを多く含むことから、インデックスに工夫を凝らした階段状インデックス(野田他　1997)を適用している。階段状インデックスとは、元の名義「電信機産業」に対して形態素解析を行い、意味のある単語単位に区切るための情報（単語区切り位置情報）のついた文字列「電信/機/産業/」を得て、この文字列をもとに、各単語を先頭とする末尾までの「電信/機/産業/」「機/産業/」「産業/」を派生させるものである（図２）。この派生により、文字列の中間からの部分一致検索（中間部分一致検索）で検索しなければ見つからない候補も、文字列の前方からの部分一致検索（前方部分一致検索）で検索することが可能になる。

この前方部分一致検索によって得られた検索候補に対して、利用者が入力した文字列の末尾に相当する個所に、単語区切り位置情報が付与されている候補を最終候補とすることで、解として有効である割合（適合率）を高めている(永井他　1997)。

本稿では、この単語区切り位置情報をＤＢ検索時のインデックスとすることによって、ＤＢ検索のスループットを向上させる手法について述べる。

単語区切り位置情報の照合手順	検索文字列	備考
1.利用者入力情報	***電信機***	
2.ＤＢに登録されている検索文字列（階段状インデックス）	***電信機***産業	検索候補
	機産業	
	産業	
	電信機械	検索候補
	機械	
3.単語区切り位置の照合	***電信/機/***産業/	○:区切一致
	電信/機/械/	×:区切不一致
4.最終的な検索候補	***電信機***産業	

図２．単語区切り位置情報を用いた検索

２．提案モデル

２－１．従来の単語区切り位置情報の問題点

利用者が入力した文字列を、階段状インデックスに対して前方部分一致検索することで、局部的な文字列からはじめるだけの検索ノイズを排除することができる。しかし、依然として、文字列の後方部

分では、局部的に一致しただけの意味のない検索ノイズが含まれている。

そこで従来の手法では、ＤＢに登録する文字列に対して形態素解析を行い、あらかじめ単語区切り位置情報を与えておき、利用者が入力した文字列の末尾部分に、単語区切り位置情報が存在する検索候補だけを最終的な候補とすることによって適合率を高める方法をとってきた（図２）。

しかし、この方法では、階段状インデックスを検索するステップと、その検索候補に対して単語区切り位置を照合するステップが必要となる。

実際に構築しているシステムでは、階段状インデックスに対して検索候補を抽出するステップまでをサーバ側で行い、検索候補に対して単語区切り位置を照合するステップについては、クライアント側で行っていた。しかしこの方法では、検索候補がＤＢ検索ステップにおいて得られた不要な検索ノイズを多く含んでいても、取捨選択を行わずにクライアントに転送しなければならず、適合率の低下および、検索候補過多の要因となっていた。

このため、サーバ側からクライアント側への検索結果の転送時には、 ＮＷ上のトラフィックを圧迫しないように転送数に上限を設けたため、一部の検索要求については全ての検索候補を転送することができなかった。

２－２．単語区切り位置情報のインデックス化

そこで、単語区切り位置情報の照合過程をクライアント側で行うのではなく、サーバ側で行う方式を提案する。

まず、元の文字列に対して形態素解析を行い、単語区切り位置を表す記号が文字列の何文字目にあるのかをカウントし、数値表現したものをインデックスとしてＤＢに登録しておく（図３）。検索時には、利用者が入力した文字列と、文字列の長さの２つを検索キーとして検索し、これらの条件を満たす検索候補を最終候補とする。

この方法によれば、サーバ側で単語区切り位置を照合した検索候補のみをクライアント側に転送できるようになる。即ち、サーバとクライアント間のＮＷ上の負荷を軽減させ、全体のパフォーマンスを向上させることができる。

文字列	従来の 文字列レベルの 単語区切り位置情報	提案の インデックス化する 単語区切り位置情報
電信機産業	電信/機/産業/	2,3,5
機産業	機/産業/	1,3
産業	産業/	2
	…	

図３．単語区切り位置情報のインデックス化

３．評価

３－１．評価方法

単語区切り位置情報をＤＢのインデックスとすることによって、全体のパフォーマンスをどれほど向上させることができるのか、また、インデックス化した単語区切り位置情報をサーバ側で検索する処理が追加されたことによって、負荷がどの程度膨らむのかを評価のポイントとした。

評価項目としてサーバ側の負荷の増大を把握するために、ＣＰＵ使用率を１０分間測定し、そのうちの５分間の平均をとった。また、全体のパフォーマンスを把握するためにレスポンスタイムとネットワーク転送量を測定した。

なおサーバは、ＨＰ Ｋ４００シリーズ４ＣＰＵでメインメモリ１Ｇ、ディスクはレイド５構成としてキャッシュ６４Ｍの合計１２８Ｇ、ＮＷはFDDIをＤＡＳで接続した。ただし、検索対象がキャッシュに載った状態での測定になることを避けるため、レスポンスタイムの評価では測定開始から３００個までの入力情報に限定して測定した。

３－２．ＣＰＵの負荷

クライアント側のＣＰＵ負荷は、従来型検索よりも減少し、サーバ側のＣＰＵ負荷は、従来型検索より高くなった。これは単語区切り位置情報の照合プロセスをクライアントからサーバへ移行させたことに起因する。しかし、 サーバ側の負荷の増加は１０％以内に収まったのに対して、クライアント側の負荷の軽減では、３５％～５５％もの効果が得られた。これは提案手法では、単語区切り位置情報をインデックス化することによってサーバ側での前

方部分一致検索が行えるようになったことから、ＤＢＭＳが得意なパターンマッチングに置き換わったことの効果であると考えられる（図４、図５）。

また、図４の数字列数が４文字から６文字の短い部分において、サーバ側の負荷が減少している理由のひとつとして、disk I/O による処理待ちが発生し、ＣＰＵがアイドリング状態になったことも考えられ、さらなる改善の余地が見いだされた。

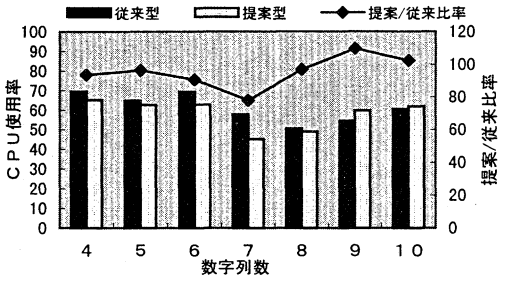


図４．サーバのＣＰＵ使用率

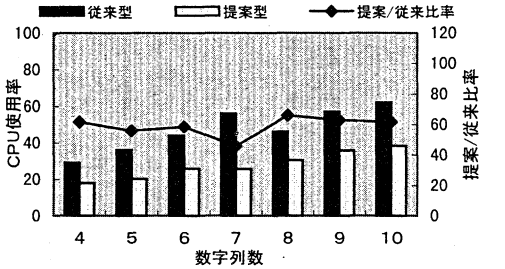


図5. ｸﾗｲｱﾝﾄのCPU使用率

３－３．レスポンスタイム

レスポンスタイムは、キャッシュによる効果を最小限に抑えるために、入力情報ファイルの先頭から３００個の入力情報を処理するまでの時間に限定して測定を実施した。提案手法では期待に反して、平均応答時間が従来手法よりも遅くなった（図６）。これは単語区切り位置情報をインデックスとして読み込む disk I/O のプロセスが増えたためであると考えらる。そこで、 disk I/O の影響がない場合を想定して、キャッシュに全てのインデックスが載った部分のログに限定した評価を行ったところ、提案手法は従来手法に比べて最大２０％の高速化が図られていた（図７）。

つまり、従来手法では、利用者が入力した文字列を元にＤＢ検索を行えば、検索候補を抽出できたのに対して、提案手法では、利用者が入力した文字列を元にＤＢ検索した後、続けて単語区切り位置情報のインデックスを読み込む処理が必要になったことでレスポンスタイムが悪化したと考えられる。

改善案としては、前方部分一致検索時用に、文字列と単語区切り位置情報を一つのインデックスとしてまとめたものを作成することで、disk I/O 回数を減らす方法がある。

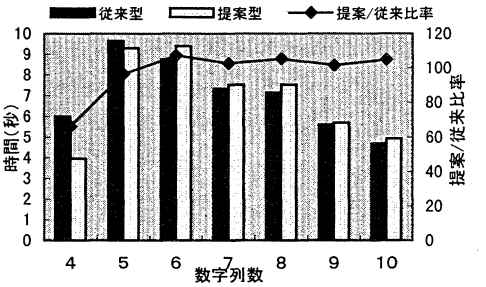


図6. ｸﾗｲｱﾝﾄ側からみた平均ﾚｽﾎﾟﾝｽﾀｲﾑ

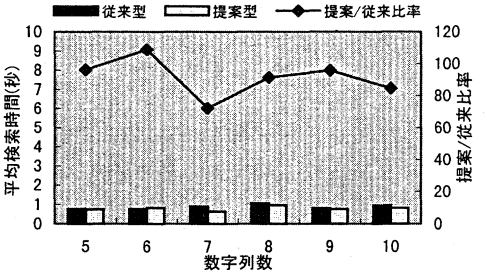


図7. ｷｬｯｼｭ後の平均ﾚｽﾎﾟﾝｽﾀｲﾑ

３－４．Ｃ／Ｓ間ネットワーク転送量

クライアントとサーバ間のネットワーク上のパケット量としては、図８及び図９に示すように送受信ともに減少している。このうち、前述の disk I/O による影響を考慮しても、提案手法によってパケット量を減少できることがわかる。

また、クライアントが受信するパケット量が減少することによって、パケットを送受信する回数その

ものも減らせることができ、結果的にクライアントが送信するパケット量まで削減できることから、全体的なスループットを向上させることができる。

このようにネットワークにかかる負担が軽減できることから、ネットワークに接続された機器のトータルスループットの向上が図られ、多重度を高めることが可能になり、接続可能なクライアント数を増加させる見込みが得られた。

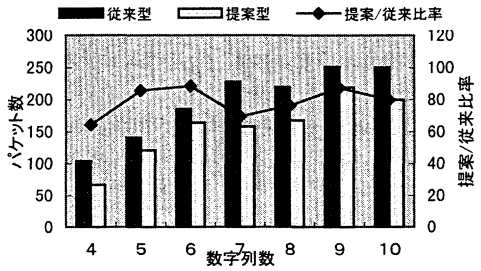


図8. クライアントの送信パケット数

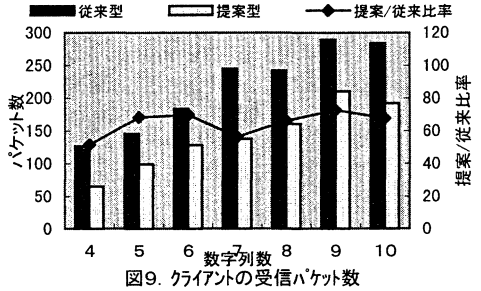


図9. クライアントの受信パケット数

3－5．考察

単語区切り位置情報を検索インデックスとして追加することによって、ネットワーク上へ転送する検索候補の絞り込み効果を確認できた。クライアントのＣＰＵ使用率やネットワークへのパケット送出量は、検索候補の削減効果により大きく減少させることができる。即ち、全体的なトータルスループットの向上という観点から、十分な改善効果が期待できる。

しかし、サーバ側において、単語区切り位置情報をインデックス化した領域を検索するための disk I/O が新たに発生し、レスポンスが若干悪化することもわかった。この点については、インデックスの張り方を変更し、disk I/O そのものを減少させるインデックス構成をとることによってレスポンスを改善できると考えられる。

4．まとめ

提案手法では、単語区切り位置情報をサーバ側の検索条件に追加することで、検索ノイズを低減させることに成功した。検索ノイズが低減したことによって利用者が入力した検索条件に一致する全候補をクライアント側へ転送させることができる。

今後、全候補をクライアント側へ転送できる利点を利用することによって、各検索候補がもつ詳細な情報を分析し、次に行うべき有効な絞り込み手段を、システム側から利用者に提示するなど、検索そのものの利便性を高めるための改善を盛り込んでいく予定である。

【参考文献】

(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).

(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　1997)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　1998)奥、永井、野田、”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”、言語処理学会第４回年次大会 P-2-5(1998).

(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).

(野田他　1997)野田、藤岡、村上、奥、“形態素解析を利用したデータベース検索高速化方法”、97信学会総合大会、D-6-9(1997).