5F－10

全文検索のためのパトリシア構造化シグネチャファイルのテキストデータベース上での実現

権藤 夏男，　金子 邦彦，　牧之内 顕文
九州大学大学院システム情報科学研究科

1. はじめに

全文検索を高速に行うためのインデックスであるシグネチャファイルは，他の全文検索用インデックスに比べサイズが小さいとい利点がある．シグネチャファイルの問題点は，全文検索におけるシグネチャファイル走査コストがテキスト数に比例することである．この走査コストを削減するために，いくつかのシグネチャファイル分割法が提案されてきた[2]．これらの研究は，シグネチャファイルを分割格納し，検索時におけるディスクI/Oコストの削減を目的とするものである．

しかし，メモリの大容量化により，シグネチャファイル全体をメモリ上に置くことが可能になりつつある．そこで，我々はメモリ上におけるシグネチャファイルによる全文検索コストの削減を目的に，ビットデータ用データ構造であるパトリシア[1]構造のシグネチャファイルを提案した．本稿では，パトリシア構造化シグネチャファイルと，分割シグネチャファイルであるビットスライスドシグネチャファイル[3]を，実メモリ上に作成し，検索コストの比較を行った．実験の結果，シグネチャファイルによる全文検索の後処理に要するコストを考慮すると，パトリシア構造化シグネチャファイルは，検索文字列が多い場合に有効であり，検索文字列が少ない場合には，2つのシグネチャファイルの走査コストの差は全検索コストから見ると小さいことが分かった．

2. シグネチャファイル

テキストシグネチャとは，各テキストごとに作成した固定長のビット列のことで，次の手順で作成される．(1) テキストの各単語ごとにハッシュ関数[4]により固定長のビットパターン(ワードシグネチャ)を作成する．(2) テキストに登場する単語のワードシグネチャの論理和をテキストシグネチャをとする．これらをまとめたものがシグネチャファイルである．

検索では，検索文字列から問合わせシグネチャを作成し，問合わせシグネチャの'1'であるビットに対応する全てのビットが'1'であるテキストシグネチャを検索する．テキスト全体の長さよりテキストシグネチャの長さの方が短いため，シグネチャファイルの利用により全文検索は速くなる．しかし，シグネチャのマッチングをテキストの数だけ繰り返すため，走査コストがテキスト数に比例し検索コストの大きな要因となる．このコストを削減るためのシグネチャファイル構成法がこれまでに提案されてきた．

代表的な分割シグネチャファイルであるビットスライトシグネチャファイルは，各シグネチャを1ビットごとに分割格納する方法である．これによって，検索では，問い合わせシグネチャの'1'であるビットに対応するビットのみを調べればよく，高速な検索が可能となる．

3. パトリシア構造化シグネチャファイル

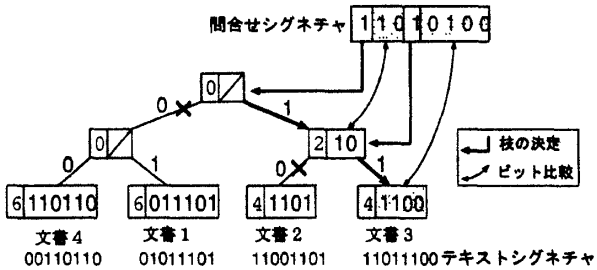

図1: パトリシア構造化シグネチャファイルの検索

パトリシアでは，'0'のビットは左の枝に，'1'のビットは右の枝に対応する．パトリシアトライ構造のシグネチャファイルによる検索では，S_T をテキスト T のシグネチャ，S_Q を問い合わせシグネチャとすると，$S_T \wedge S_Q \equiv S_Q$ を満たす T を検索する．S_Q で'1'のビットでは右の枝のみをたどり，'0'であるビットでは左右両方の枝をたどる．パトリシアの各節点はスキップする部分シグネチャ S'_T と

Implementation of Patricia Structured Signature File for Full-Text Retrieval on Text Database
Natsuo Gondo, Kunihiko Kaneko, Akifumi Makinouchi
Graduate School of Information Science and Electrical Engineering, Kyushu University

産經新聞	35538
毎日新聞	27519
讀賣新聞	61124
合計	124181

表 1: 実験データの HTML ファイル数

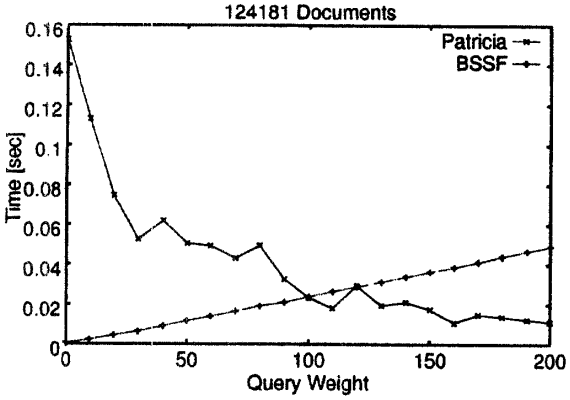

図 2: 問合わせシグネチャのビット数と検索時間

スキップ数を格納しておき，訪れた節点の S'_T と，対応する S_Q の部分シグネチャ S'_Q について $S'_T \wedge S'_Q \equiv S'_Q$ の条件判定を行う．条件を満たせば，スキップ数と S_Q のビット値によりたどる枝を決定し次に進む．条件を満たさなければその節点より以下は調べない．(図 1)

4. 実験

パトリシア構造化シグネチャファイルとビットスライストシグネチャファイルを実メモリ上に作成し，検索コストの比較を行った．

i) 実験データ — 実験データとして，産經新聞社 [5]，毎日新聞社 [6]，讀賣新聞社 [7] の各社が WWW で公開している HTML ファイルの日本語記事を使用した．(表 1)

ii) 実験パラメータ — シグネチャのビット長を 1024 ビットとし，テキスト 1 バイトあたり 2 ビットを '1' とするハッシュ関数を用いた．

iii) シグネチャ生成方法 — 日本語のコードは EUC とし，2 文字組 (bigram) を単位として，ワードシグネチャを作成した．

iv) 実験方法 — 問合わせシグネチャの '1' のビットの位置をランダムに選択し，'1' のビットの数を 0 から 200 まで，10 きざみで変化させ，その検索処理時間を測定した．

検索時間の測定の結果を図 2に示す．

5. 考察

実験の結果，問合わせシグネチャの '1' のビットの数が増えるにしたがって，ビットスライストシグネチャファイルの検索処理時間は増加するが，パトリシア構造化シグネチャファイルでは逆に減少することが分かった．

この実験では，問合わせシグネチャの '1' のビットの数が 100 以上のとき，パトリシア構造化シグネチャファイルのほうが速くなった．ビットの数が 100 以上とは，日本語の検索文字列 1 つを 4 文字とすると文字列 4 つ (16 文字) に相当する．したがって，パトリシア構造化シグネチャファイルが有効になる場合は有り得ると思われる．

シグネチャファイルによる全文検索では後処理が必要である．後処理に要する時間を考慮したところ，問合わせの '1' のビットの数が 100 以下の場合には，後処理に要する時間が大きく，シグネチャファイルの走査コストの差が相対的に小さくなることも分かった．

6. まとめ

本稿では，パトリシア構造化シグネチャファイルによる全文検索は，検索文字列が少ないときには，ビットスライスドシグネチャファイルと比べ検索時間はあまり変わらないことを示し，検索文字列が多いほど検索が高速に行えることを示した．これは，多数のテキストから少数のテキストを得る場合に有効であるということである．

参考文献

[1] D.R.Morrison. "Practical algorithm to retrieve information coded in alphanumeric", Journal of the ACM, 15(4):514-34, 1968

[2] 渡辺悟康，北川博之，"分割ビットスライストシグネチャファイルの提案と集合値検索への適用"，情報処理学会論文誌，vol.37，No.12，pp.2314-2325, 1996

[3] C.Faloutsos, and R,Chan. "Fast test access methods for optical disks: Designs and Performance comparison," in Proc. 14th Int. Conf. on VLDB. 1988. pp.280-293.

[4] C.Faloutsos, "Signature-Based Text Retrieval Method : A Survey", Data Eng. Bulletin, Vol.13, No.1, pp.25-32, 1990

[5] 産經新聞社 URL: *http://www.sankei.co.jp/*

[6] 毎日新聞社 URL: *http://www.mainichi.co.jp/*

[7] 讀賣新聞社 URL: *http://www.yomiuri.co.jp/*