6W－1

階層化されたシグネチャファイルを用いた集合値検索方式の検討

渡辺 悟康 *　北川 博之 †　石川 佳治 ‡

* 筑波大学工学研究科　† 筑波大学電子情報工学系　‡ 奈良先端科学技術大学院大学

1 はじめに

集合データは、複雑なデータ構造を支援するデータベース中において、頻繁に現れる基本的なデータ構造である。そのため、集合データを用いるようなデータベースシステムでは、集合値を効率良く支援する索引機構が必要となる。我々は、従来テキスト検索に用いられてきたシグネチャファイルを集合値検索機構として取り上げ、比較的小規模のデータベース上での、様々なコスト評価を行なってきた[1, 2]。しかし、シグネチャファイルを用いた集合値検索には、検索コストがデータオブジェクトの数に比例して増加するという問題点がある。このためデータオブジェクト数が増加した時には、入れ子型索引より性能が劣ることがある[3]。

本稿では、シグネチャファイルによる検索の高速化の手段として知られている階層化の概念を用い、主に性能改善の必要性が大きい集合値問い合わせ条件を対象に階層化されたシグネチャファイルの有効性を検討する。

2 シグネチャファイルを用いた集合値検索

シグネチャ (*signature*) とは、個々のデータオブジェクトから生成される固定長のビット列のことである。

シグネチャの作成法

1. 集合の各要素から、長さが F ビットで、その内 m ビットが"1" にセットされている**要素シグネチャ** (*element signature*) を作成する。
2. すべての要素シグネチャのビットごとの論理和をとる**スーパーインポーズドコーディング** (*superimposed coding*) を行ない、**集合シグネチャ** (*set signature*) を作成する。

このようにして生成されたシグネチャと、各データオブジェクトの識別子 (OID) の組を格納したのが**シグネチャファイル** (*signature file*) である。シグネチャファイルの構成法としては、シグネチャをビットごとに別々のファイルに格納する**ビットスライストシグネチャファイル** (*bit-sliced signature file*, BSSF) を用いる。

問い合わせが与えられた際に、問い合わせ条件中に現れる集合を**問い合わせ集合** (*query set*, Q)、データベース中の集合を**ターゲット集合** (*target set*, T) と呼ぶ。また、それぞれから作成される集合シグネチャを、**問い合わせシグネチャ** (*query signature*, S_Q)、**ターゲットシグネチャ** (*target signature*, S_T) と呼ぶ。

集合値検索の問い合わせ条件には様々なものがあるが、本稿では、Q が T に含まれる場合 ($T \supseteq Q$) を対象とする。この問い合わせ条件の場合、$S_Q \wedge S_T \equiv S_Q$ を満たすターゲットシグネチャが、問い合わせ条件を満たす候補となる。しかし、これらの候補の中には、実際に問い合わせ条件を満たす**アクチュアルドロップ** (*actual drop*) と、実際には条件を満たさない**フォルスドロップ** (*false drop*) があるので、実際のデータからそれらを区別する必要がある。

3 階層化されたBSSF

本稿では、2段階に階層化されたBSSFを考える。

階層化されたBSSFの作成法

1. 前述のようにターゲットシグネチャを作成し、BSSFに格納する。このBSSFをレベル1 BSSFと呼ぶ。
2. レベル1の集合シグネチャ M 個について、さらにスーパーインポーズドコーディングを行ない、1つのシグネチャを作成する。それらのシグネチャをまとめたものを、レベル2 BSSFと呼ぶ。

階層化されたBSSFを作成した時のシグネチャファイルの概略は、図1のようになる。($F = 8, m = 2, M = 4$)

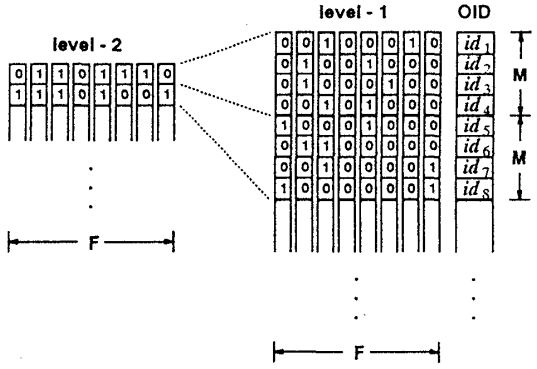

図 1: 階層化されたBSSFの構成

4 コスト解析

表1に使用する記号の定義を示す。また、特に式を示さなかったものについては[3]の式を用いる。

4.1 フォルスドロップ確率

レベル2 BSSFのシグネチャ1つに対応した集合の平均要素数 D_{t2} は、$D_t \ll V$ の時、(1) 式で与えられる。

$$D_{t2} = V\left\{1 - \left(1 - \frac{\binom{V-1}{D_t-1}}{\binom{V}{D_t}}\right)^M\right\} \approx V(1 - e^{-\frac{D_t}{V}M}) \quad (1)$$

よって、[1, 2] より、レベル2 BSSFを検索する際のフォルスドロップ確率 Fd_2 は、(2) 式で与えられる。

$$Fd_2 \approx (1 - e^{-\frac{mD_{t2}}{F}})^{mD_q} \quad (2)$$

また、レベル1 BSSFのみを用いた場合のフォルスドロップ確率 Fd_1 は、(3) 式で与えられる。

$$Fd_1 \approx (1 - e^{-\frac{mD_t}{F}})^{mD_q} \quad (3)$$

Evaluation of Two-level Signature Files as Set-valued Object Retrieval Facilities

Noriyasu Watanabe*, Hiroyuki Kitagawa† and Yoshiharu Ishikawa‡

*Doctoral Degree Program in Engineering, Univ. of Tsukuba

†Institute of Information Sciences and Electronics, Univ. of Tsukuba

‡Nara Institute of Science and Technology

記号	定義
N	オブジェクトの総数 (= 320,000)
P	1 ページのバイト長 (= 4096)
b	1 バイトのビット長 (= 8)
M	レベル 2 のシグネチャを 1 つ作るのに必要なレベル 1 のシグネチャの個数
V	ドメインの要素数 (= 10,000)
D_t	ターゲット集合の要素数
D_q	問い合わせ集合の要素数
F	シグネチャのビット長
m	要素シグネチャで "1" のセットされるビット数 (= 2)
A	アクチュアルドロップ数
P_s	成功検索時の 1 オブジェクト当たりのページアクセス数 (= 1)
P_u	不成功探索時の 1 オブジェクト当たりのページアクセス数 (= 1)

表 1: 変数の定義

4.2 検索コスト

総検索コストは、各レベルでの BSSF を検索するコスト、OID ファイルをアクセスするコスト (LC_{OID})、候補となったターゲットシグネチャがアクチュアルドロップかフォルスドロップかを実際のデータから区別するコストの和になる。

レベル 1,2 BSSF の検索コスト LC_1, LC_2 は、以下のように与えられる。

$$LC_1 = \left\{1-(1-Fd_2)^{\frac{Pb}{M}}\right\}\left\lceil\frac{N}{Pb}\right\rceil m_q \quad (4)$$

$$LC_2 = \left\lceil\frac{N}{PbM}\right\rceil m_q \quad (5)$$

ここで、m_q は (6) 式で与えられる問い合わせシグネチャのウエイト ("1" の立っているビット数) の期待値である。

$$m_q \approx F(1-e^{-\frac{m}{F}D_q}) \quad (6)$$

以上より、総検索コスト RC は (7) 式で与えられる。

$$RC = LC_2+LC_1+LC_{OID}+P_sA+P_uFd_1(N-A) \quad (7)$$

階層化された BSSF (BSSF2)、通常の BSSF (BSSF)、入れ子型索引 (NIX) の 3 種類の検索機構の検索コストのグラフを図 2に示す。

傾向として、D_q が一定の値より大きくなった時、階層化された BSSF は有効であることが分かる。これは、D_q が大きくなると、Fd_2 が減少するために、レベル 2 BSSF での絞り込みが効果的に行なわれるためである。しかし、D_q が小さい時は Fd_2 が大きいため、階層化された BSSF の検索コストは通常の BSSF とほぼ同じになる。また、M が大きいと Fd_2 が悪化し、レベル 2 BSSF での絞り込みが効果的にできない。逆に M が小さいと LC_2 が悪化してしまい、やはり検索コストは大きくなる。

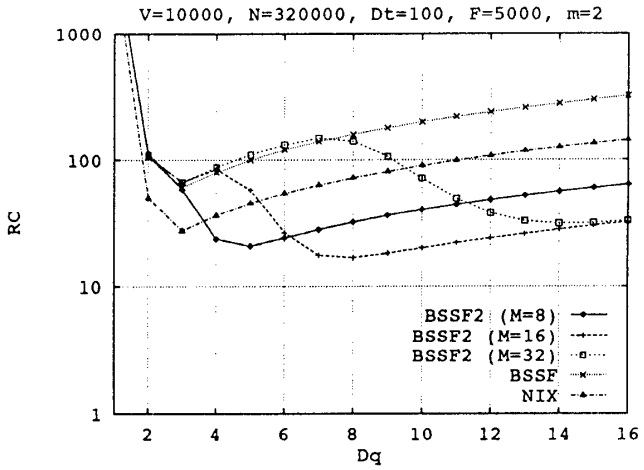

図 2: $F = 5000, D_t = 100$ の検索コスト

4.3 格納コスト

階層化された BSSF の格納コスト SC は、各レベルの BSSF の格納コスト (SC_1, SC_2) と OID ファイルの格納コスト (SC_{OID}) との和になる。

$$\begin{aligned} SC &= SC_2 + SC_1 + SC_{OID} \\ &= \left\lceil\frac{N}{PbM}\right\rceil F + \left\lceil\frac{N}{Pb}\right\rceil F + SC_{OID} \end{aligned} \quad (8)$$

格納コストは、通常の BSSF に比べて、レベル 2 BSSF を作るために増加する。しかし、検索コストがほぼ同等の入れ子型索引と比べると、低く抑えることができる。

5 まとめ

本稿では階層化された BSSF の検索コストと、格納コストについて見積り式を示した。本稿で検討したパラメタ設定の下では、$T \supseteq Q$ の問い合わせに対する検索コストは、通常の BSSF の 1 ～ 2 割程度となり、一定以上の D_q 値に対しては、入れ子型索引以上の性能を示した。格納コストは通常の BSSF より若干大きくなるが、入れ子型索引と比べると、小さい格納コストで同等の検索コストを得られることが分かった。

今後の課題としては、[1] において提案されたスマート検索方式を用いた時の評価、N が増加した時の多段化などがあげられる。

参考文献

[1] Y. Ishikawa, H. Kitagawa, and N. Ohbo. Evalution of signature files as set access facilities in OODBs. In *Proc. ACM SIGMOD*, 1993.

[2] H. Kitagawa, Y. Fukushima, Y. Ishikawa, and N. Ohbo. Estimation of false drops in set-valued object retrieval with signature files. In *Proc. 4th International Conference on Foundations of Data Organization and Algorithms*, 1993.

[3] 石川, 北川, 大保. シグネチャファイルによる集合値検索のコスト評価. 情報処理学会研究会報告 93-DBS-94-27, 7 1993.