1Aa-04

シグネチャを持つ入れ子型索引を用いた集合値検索のコスト解析

永戸 克裕 † 北川 博之 ††

† 筑波大学 理工学研究科 †† 筑波大学 電子・情報工学系

1 はじめに

集合は基本的なデータ構造であり、複合オブジェクトの部分構造としても頻繁に現れる。このため、集合値に関する検索条件を効率的に支援する索引機構は、先進的な応用分野を対象としたデータベースシステムにおいて重要なものとなる。

集合値検索を支援する索引機構として、入れ子型索引 (NIX) や、シグネチャファイルなどがある。この2つの索引機構を比較すると、問合せ条件やデータベースの規模によってその優劣は異なる。そこで、シグネチャと B^+ 木を組み合わせることで、それぞれの長所を生かしたより優れた集合値検索のための索引機構を提案し、そのコストを評価する。

2 集合値検索とその処理

2.1 集合値検索

例として学生に関するデータベースを考える。Student クラスには属性として name と hobbies があり、hobbies 属性は文字列の集合を値とする。ここで、次の問合せ Q_1 を考える。

Q_1: **select** name
from Student
where hobbies $\supseteq$ {"Baseball", "Fishing"}

Q_1 は、趣味に "Baseball" と "Fishing" の両方を持つ学生の名前を求める問合せである。問合せ条件中に現れる集合を**問合せ集合** (query set, Q) と呼び、データベース中に格納され問合せ集合との比較の対象となるそれぞれの集合を**ターゲット集合** (target set, T) と呼ぶ。

問合せ集合を部分集合として含むターゲット集合を検索する問合せを $T \supseteq Q$ (has-subset) の問合せと呼び、問合せ集合の部分集合となるターゲット集合を検索する問合せを $T \subseteq Q$ (is-subset) の問合せと呼ぶ。

2.2 集合値検索の処理

集合値検索を支援する索引機構として、入れ子型索引 (NIX) とシグネチャファイルについて考える。

NIX は、B^+ 木に基づく索引手法で、リーフページにキー値と対応する識別子 (OID) リストからなる索引エントリが格納される。$T \supseteq Q$ の問合せでは、与えられた D_q 個の要素を持つ問合せ集合に対して、対応する D_q 個の OID 集合が検索される。次に、この OID 集合の積集合中の各 OID を基にデータオブジェクトが検索される。$T \subseteq Q$ の場合は、同様に D_q 個の OID 集合が検索され、次にそれらの和集合を求める。この和集合に対応するデータオブジェクトは必ずしも検索条件を満たすとは限らないため、実際にデータオブジェクトを検索して、条件を満たすか調べなければならない。

シグネチャファイル (signature file) は、主にテキストデータベースにおけるキーワード検索において利用されてきた索引手法である。**シグネチャ**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であり、シグネチャを対応するデータオブジェクトの OID とともに格納したものがシグネチャファイルである。

集合シグネチャは以下のように生成される。集合値が与えられると、まず、ハッシュ法などにより、集合の各要素から**要素シグネチャ** (element signature) が作られる。次に、すべての要素シグネチャのビットごとの論理和をとることにより、**集合シグネチャ** (set signature) が生成される。作成された集合シグネチャ (**ターゲットシグネチャ** (target signature)) は、対応する OID とともにシグネチャファイルに格納される (**図 1**)。

集合の要素		要素シグネチャ ($F = 8$, $m = 2$)
"Baseball"	$\longrightarrow$	01000100
"Golf"	$\longrightarrow$	00100001
"Fishing"	$\longrightarrow$	00010100
		⇓ 論理和
集合シグネチャ		01110101

図 1: 集合シグネチャの生成

シグネチャファイルの物理的な構成手法として、**ビットスライストシグネチャファイル** (Bit-slised Signature File,BSSF)[1, 2] がある。BSSF ではシグネチャはビットごとに別々のファイル (**ビットスライス**) に格納されるため、検索では一部にアクセスすればよいことになる。

$T \supseteq Q$ に対する問合せ処理は以下のようになる。問合せが与えられると、問合せ集合より**問合せシグネチャ** (query signature) が生成される。次に、問合せシグネチャにおいて "1" の値を持つ各ビット位置に対応するビットスライスが検索される。次に、読み出したすべてのビットスライスの論理積をとり、結果 "1" の値を持つエントリについて、対応する OID を OID ファイルから検索する。対応するデータオブジェクトは問合せを満たす候補 (**ドロップ** (drop)) となる。最後に、それぞれのドロップが実際に問合せ条件を満たすか否か調べる。

$T \subseteq Q$ に対する問合せ処理は、$T \supseteq Q$ の処理で "1" と "0" の役割を入れ換えたものである。

検索対象のデータベースが小・中規模の場合、$T \supseteq Q$ では NIX と BSSF は同程度の性能であり、$T \subseteq Q$ では BSSF の方が優れる。しかし、データベース中のオブジェクト数が増加すると、BSSF に対する NIX の優位性は徐々に低減する。

3 シグネチャと B^+ 木を組み合わせた索引

上の2つの索引機構の長所を生かすため、次のようにしてこの2つを統合することを考える。

シグネチャ付き NIX(S-NIX) は、NIX のリーフページに、キー値だけでなく、対応する OID リスト、ターゲットシグネチャが格納される (**図 2**)。

3.1 検索処理

S-NIX における、$T \supseteq Q$ に対する検索処理は NIX と同様で、シグネチャ部は使用しない。$T \subseteq Q$ に対する処理は、まず、NIX 部を用いて OID 集合が検索される。次に、それらに対応するシグネチャファイルと問合せシグネチャの間に問合せ条件が成立するかを調べる。条件が成立すると判定されたターゲットオブジェクトについて、オブジェクトを検索して実際に問合せ条件を満たすか調べる。

Cost Estimation of Set-valued Object Retrieval Using Nested Index Including Signatures

Katsuhiro Nagato†, Hiroyuki Kitagawa††

†Master's Degree Program in Sci. and Eng., Univ. of Tsukuba

††Institute of Info. Sci. and Elec., Univ. of Tsukuba

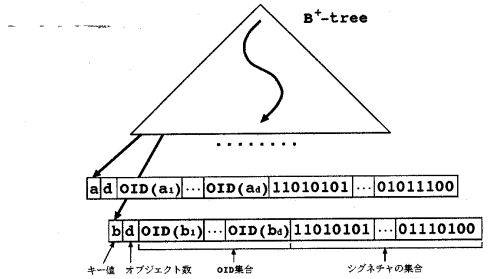

図 2: S-NIX の構成

4 コスト解析

S-NIX、BSSF、NIX の 3 種類の索引機構について、コスト解析を行った。但し、$T \supseteq Q$ における索引、$T \subseteq Q$ における BSSF による索引については、スマート検索方式 [1] を評価した。

変数名	意味
V	集合要素の定義域の要素数
N	オブジェクトの総数
D_t	ターゲット集合の要素数
F	シグネチャのビット長
m	要素シグネチャにおける “1” の数 (ウェイト)

表 1: 記号とその意味

4.1 検索コスト

4.1.1 $T \supseteq Q$ の問合せの検索コスト

V=16000、N=160000、D_t=10 における検索コストを**図 3** に示す。この場合、S-NIX のコストは NIX とほぼ同等で、BSSF に比べ優れたものになる。

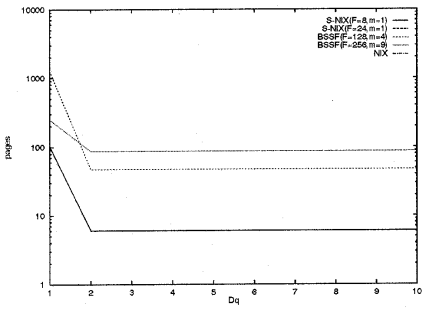

図 3: $T \supseteq Q$ の検索コスト (V=16000,N=160000,D_t=10)

4.1.2 $T \subseteq Q$ の問合せの検索コスト

D_t=10 における検索コストを**図 4**、**5** に示す。S-NIX では、V、N が大きくなるほど、B^+ 木部の絞り込みにより、BSSF に比べコストの増大が抑えられる。

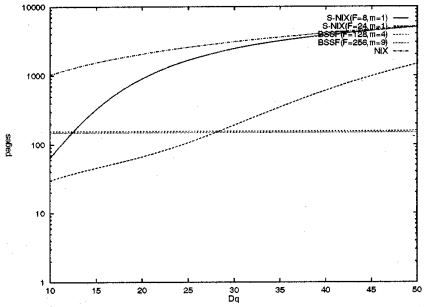

図 4: $T \subseteq Q$ の検索コスト (V=16000,N=160000)

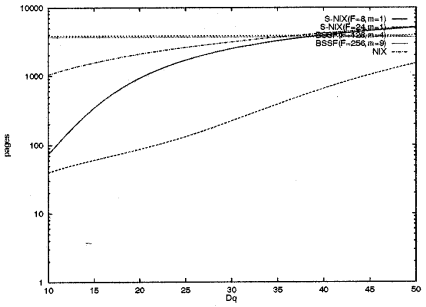

図 5: $T \subseteq Q$ の検索コスト (V=400000,N=4000000)

4.2 格納コスト

V=400000、N=4000000、D_t=10 における格納コストを**表 2** に示す。S-NIX の格納コストは、BSSF, NIX に対して劣っている。しかし、F の長さを必要最小限にすることによって、コストの増大を抑えることができる。

D_t	ファイル	ページ数
10	S-NIX(F=8)	100463
	S-NIX(F=24)	133950
	BSSF(F=128)	23557
	BSSF(F=256)	39301
	NIX	80370

表 2: 格納コスト (V=400000,N=4000000)

4.3 更新コスト

V=400000、N=4000000、D_t=10 における更新コストは、**表 3** のようになる。S-NIX は、F のサイズによりリーフノードサイズが大きくなり過ぎない限り、更新コストは NIX とほぼ同等に抑えられる。

D_t	ファイル	挿入コスト	削除コスト
10	S-NIX(F=8)	40	40
	S-NIX(F=24)	40	40
	BSSF(F=128,m=4)	36	3941
	BSSF(F=256,m=9)	77	3983
	NIX	40	40

表 3: 更新コスト (V=400000,N=4000000)

5 おわりに

本研究では、シグネチャ付き NIX を提案し、そのコスト解析を行った。S-NIX は、V、N の両方が大きくなるほど、シグネチャと B^+ 木を組み合わせた効果が発揮され、BSSF、NIX に対して検索コストの低減が実現された。

今後の課題としては、シグネチャの格納に関する冗長性の解決、圧縮を組み合わせたより効率的な索引構成法の検討等がある。

参考文献

[1] Y. Ishikawa, H. Kitagawa, N. Ohbo. Evaluation of Signature Files as Set Access Facilities in OODBs, *Proc. ACM SIGMOD Conf.*, pp.247-256. May.1993.

[2] 石川佳治, 北川博之, 大保信夫. シグネチャファイルによる集合値検索のコスト評価, 情報処理学会論文誌, 第 36 巻第 2 号. Feb.1995.