2E−1

複合オブジェクトに対する索引機構の研究

福島慶明† 石川佳治†† 于旭††† 北川博之††† 大保信夫†††
(†筑波大学理工学研究科,††筑波大学工学研究科,†††筑波大学電子・情報工学系)

1.はじめに

近年CAD等の高度応用における複雑な対象をデータベース化する要求が高まっている。この要求に対応するために、nested relationalモデルやオブジェクト指向モデルの研究がなされてきている。これらの新しいデータモデルでは階層構造や集合値を取り扱うための索引機構が必要となる。これまでに階層構造を扱うための索引機構としてnested attribute index[1]等が考案されているが、集合演算をサポートしていないので充分とはいえない。

本論文ではnested relationalデータベースの効率良い検索をサポートするためにシグネチャ(signature)を用いた索引機構を考案し考察する。

2.問い合わせの種類

nested relationに対するデータ操作を記述するのに幾つかの非正規関係代数が提案されている。検索操作を記述するのにここでは下記の選択演算(σ)を考える[2]。

σ[G,C](R)　G:選択をするグループ
C:選択条件
R:選択をする関係

図1に対する問い合わせの種類としては、

1.σ[DEPT;EMP.NAME∋nishimoto](D)
2.σ[DEPT;EMP.NAME⊇{nishimoto,hashimoto}](D)
3.σ[DEPT;EMP.NAME⊆{nishimoto,hashimoto}](D)
4.σ[DEPT;EMP.NAME≡{nishimoto,hashimoto}](D)

等が考えられる。1の問い合わせはnested attribute indexで効率的に検索することができるが、2,3,4に対しては集合演算を含むために大変コストが掛かる。一方superimposed coding法[3]を用いたシグネチャを利用することで集合をのコード化を行なうことが可能であり、上記のような検索に対して有効であると思われる。

3.シグネチャを用いた索引機構

シグネチャは元来図書館や医療記録など大きなデータを持つデータベースのデータを篩い分けるために考案された手法である[4]。

3.1 シグネチャファイル(signature file)の構成

本論文では、superimposed coding法を用いた図2のような索引機構を考案した。各インデックス属性のインスタンスはハッシュされエレメントシグネチャ(element signature)を生成する。各エレメントシグネチャはスーパーインポーズされセットシグネチャ(set signature)を作る。各セットシグネチャはタプルIDと結び付けられる。

3.2 シグネチャを用いた検索

上記の過程では幾つかの異なる集合が同じセットシグネチャを生成する可能性がある。したがって検索の際にシグネチャがマッチして実際には検索条件を満たさないタプル(false dropと呼ばれる)を除去せねばならない。

2で述べた問い合わせに対し、検索は以下のステップで行なわれる。

1) 問い合わせ中の各エレメントをシグネチャに変換し、スーパーインポーズすることにより質問シグネチャ(query signature)を生成する。
2) シグネチャファイルをスキャンし、各問い合わせで以下の条件を満たすとき、候補タプルとなる。
1. (query signature)∧(set signature)=(query signature)
2. (query signature)∧(set signature)=(query signature)
3. (query signature)∧(set signature)=(set signature)
4. (query signature)＝(set signature)
3)2)でマッチしたタプルから実際にマッチするタプルを調べる。

3.3 検索の例

検索の例として2の場合について説明する(図3)。問い合わせのエレメントはそれぞれシグネチャに変換し、これをスーパーインポーズする事で質問シグネチャを生成する。質問シグネチャで1がセットされているビット位置がセットシグネチャにおいても全てセットされていれば候補タプルとなり、T1が候補となる。

4. False Drop

false dropの確率は本索引機構の性能を知る上で最も重要な点である。またシグネチャのビット長とも関連しこれはデータ格納持のオーバーヘッドに直接影響する。

ここではシグネチャで表現できる属性値の数とデータベース中の別個のエレメントの数(V)はインデックスを付ける属性の要素数の平均(N_T)よりも充分に大きいものとする。またランダムに選択したエレメントはお互い独立にタプルに出現するものとする。シグネチャはFビット長を持ち、エレメントシグネチャではそのmビットが1にセットされる(F>>m)。この時、各セットシグネチャのあるビットが1である確率は、

$$1-e^{-(mN_T/F)}$$

で与えられる[5]。問い合わせ中のエレメントの数が1の時は質問シグネチャでmビットが1にセットされるので、false dropの確率をF_dとすると、

$$F_d=\left[1-e^{-(mN_T/F)}\right]^m$$

が得られる。またmに対し最適化を行なうと、

$$F_d=\left(\frac{1}{2}\right)^{m_{opt}},\ m_{opt}=\frac{F\ln 2}{N_T}$$

となる。

問い合わせ中のエレメントの数が複数の時は、質問シグネチャでセットされるビット数がスーパーインポーズされる事で増加してくる。問い合わせのエレメントの数が2の時、質問シグネチャでセットされるビット数は平均して、

$$\overline{m_2}=\frac{\sum_{i=0}^{m}(m+i)\cdot {}_FC_{m+i}\cdot {}_{m+i}C_m\cdot {}_mC_{m-i}}{({}_FC_m)^2}$$

となり、この時false dropの確率は、

$$F_{d2}=\left[1-e^{-(mN_T/F)}\right]^{\overline{m_2}}$$

となる。図4に問い合わせ中のエレメントの数が1と2の時、異なるN_Tの値に対するfalse dropの確率を示す。これからfalse dropはFを大きくすることで減少するが、一方ストアのオーバーヘッドは増加する。また問い合わせのエレメント数の増加に伴い、false dropが減少する事も分かる。インデックス属性の要素数が大きいほどfalse dropの確率は大きくなる。

A Signature Indexing Scheme for Handling Complex Objects

Yoshiaki Fukushima,Yoshiharu Ishikawa,Xu Yu,Hiroyuki Kitagawa,Nobuo Ohbo
Univesity of Tshukuba

5 Nested Attribute Indexとの比較

5.1 記憶容量

TIDのサイズをSTビットとし、属性値のサイズをSaビットとする。ストアのスペースは単純には、

signatre　　Ss=(F+ST)T

nested attribute index　　Sn=(Sa+S T TNI/V)V

で表せる(実際にはnested attribute indexでは更に索引部のためのスペース等もある)。シグネチャのビット長にもよるがnested attribute indexではTIDの重複情報があるのでサイズが大きくなる(図5)。

5.2 検索効率

問い合わせ中のエレメントの数が1つの時は本索引機構ではfalse dropの処理があるのでnested attribute indexの方が効率が良い事が予想される。問い合わせ中のエレメントの数が増えていくとnested attribute indexではインデックスファイルのアクセス回数がそれだけ増加し、効率が落ちる(本索引機構ではファイルアクセスの回数は常に1回で済む)。本索引機構ではfalse dropの処理がまた必要になるが、nested attribute indexにおいては検索されたTIDの集合に対して集合演算が必要となりコストが高くなる。また検索の種類が2の3や4である時は、nested attribute indexの検索だけでは対応することが困難である。

以上のことより二つのインデックスの有効な場面を検証し使い分けることにより、検索効率は向上するものと思われる。

6.おわりに

本研究ではシグネチャを用いたnested relationに対する索引機構を考案した。今後更に解析を進め、シミュレーションを行なう予定である。また検索効率の改善のために、bit sliceを用いた技法[4]やシグネチャの階層化[6]等の導入も考えられる。

参考文献

[1] Elisa Bertino,Won Kim," Index Techniques for Queries on Nested Objects ",IEEE Transactions on Knowledge and Data Engineering,Vol.1 ,No.2 ,1989.

[2] Kitagawa H.,Kunii.T.L.,"The Unnormalized Relational Data Model",Springer,1989.

[3] Kam-Fai Wong,M.Howard Williams,"A Superimposed Codeword Indexing Schema for Handling Sets in Prolog Databases",International Symposium on Database Systems for Advanced Applications,April,1991.

[4] A.Kent,R.Sachs-Davis,K.Ramamohanarao,"A Superimposed Coding Schema Based on Multiple Block Descriptor Files for Indexing Very Large Data Bases",14th Proc.of VLDB,1988.

[5] Chris Faloutsos,Stavros Christodoulakis,"Signature Files:An Access Method for Documents and Its Analytical Performance Evaluation",ACM Transaction on Office Information Systems,Vol.2,No.4,October 1984.

[6] Walter W. Chang,Hans J. Schek,"A Signature Access Method for the Starburst Database System",15th Proc. of VLDB,1989.

D:

(DEPT)					
TID	DNO	MGRNO	MNAME	(EMP)	
				EMPNO	ENAME
T1	314	56194	kawakami	39582	kondo
				56019	nishimoto
				69011	hashimoto
T2	125	43586	ogawa	58912	wakamatsu
				90011	iida
				78218	nishimura
				98902	aoyagi
T3	218	71349	ikeda	72227	hasegawa
				89211	nakamura
				92100	nagano
				89921	aoyama
T4	417	91093	takagi	87710	simoda
				81193	tayama
				75913	ohara
				96001	nakazima

図 1 nested relationの例

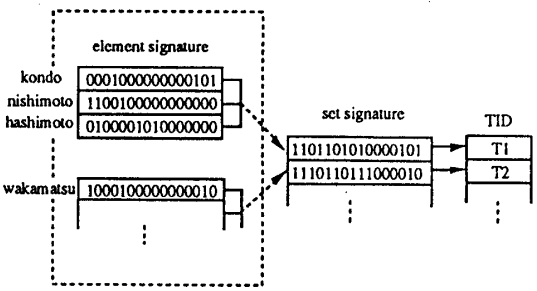

図 2 signature fileの構成

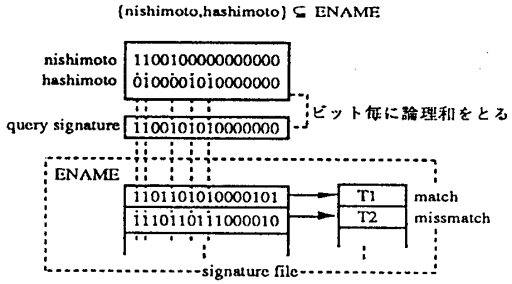

図3 本索引機構による検索

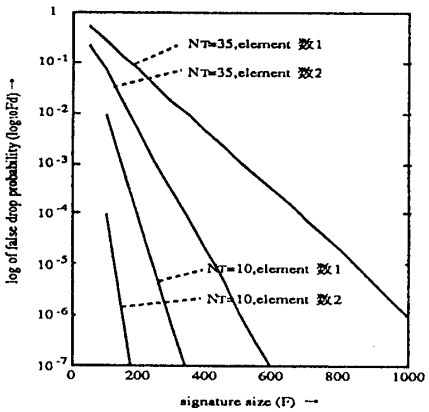

図4 $\log_{10}Fd \times F$

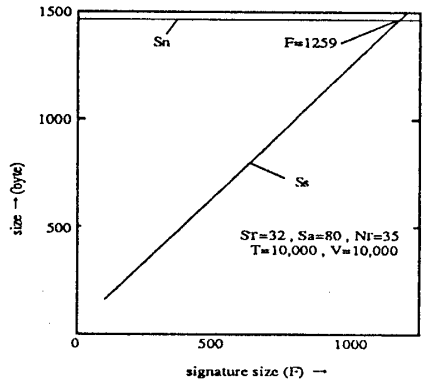

図 5 index fileのサイズ