2K-2

GRACE二次記憶系に於ける拡張多次元クラスタリング技法

伏見 信也† 喜連川 優‡ 田中 英彦† 元岡 達†

(†東京大学 工学部 ‡東京大学 生産技術研究所)

0. はじめに

関係データベースマシンGRACEはjoin等の処理負荷の重い関係代数演算を高速に実行することが可能である。従ってGRACEに於ける問合せ木全体の処理は、低速大容量の二次記憶系(磁気ディスク)へのアクセス時間、即ちselection 演算を高速化することによってより高い性能が期待できる。この為、我々は先に問合せ分布、タプル分布に応じた多次元クラスタリング技法を提案した[1]。本稿ではこの技法を更に拡張し、リレーションのトランスポーズ(射影分解)と多次元クラスタリングを組合せた新しいクラスのクラスタリング技法について考察する。

1. ページ空間に対するクラスタリング特性

一般に磁気ディスクに於けるリレーションの格納、アクセス単位は一定のタプル容量を持ったページである。ディスクに対し実際にアクセスされる単位がタプルではなくページである点にディスクアクセス回数減少に利用可能な自由度が存在し、逆に不要なタプルの読出しによるアクセスページ数増大の可能性が存在する。ここである問合せによってアクセスされたページ数をπ、これらページから selection predicate によって実際に取り出されたタプル数をτと書けば、一般に

$\tau \leqq V \cdot \pi$ [V:ページ容量(タプル)]

が成立する。$\tau \sim V \cdot \pi$であればリレーションのページ分割に関し改善の余地はなく、このことをページ空間は問合せに対しクラスタリング特性を持つと言う。先に述べた隘路解消の為にはディスクアクセス回数の平均値、即ち平均アクセスページ数$\bar{\pi}$の最小化が必要であるが、これはτの平均値$\bar{\tau}$と$\bar{\pi}$との間に$\bar{\tau} \sim V \cdot \bar{\pi}$の関係が成立すること、即ちページ空間が平均的なクラスタリング特性を有する様なページ分割を求めることに帰着する。我々が[1]で提案した多次元クラスタリングアルゴリズムは、問合せ分布及びタプル分布を考慮し、selection predicateの与えられる属性、及びその値の分布の偏りと、複数属性に対して各々与えられたpredicate の相乗的な絞り込み効果を利用することによってπの最小化を図るものであった。一方で、リレーションをタプル方向にページ分割する限りクラスタリング特性を保証出来ない場合が存在する。その最も顕著な例が、問合せ分布が一様な一次元 partial match query の場合である。各属性のselectivity が充分に小さければ、この時

$\bar{\tau} \sim 1$、$\bar{\pi} \sim N^{1-1/k} >> \bar{\tau}/V$

[k:全属性数、N:全ページ数]

となりクラスタリング特性は極めて悪い。

2. 属性/タプルクラスタリング

2.1 多次元クラスタリングに於けるトランスポーズの導入

ここで提案する新しいクラスタリング技法はリレーションを(部分)トランスポーズにより複数個のサブリレーションに分割し、各サブリレーションに対し更に多次元クラスタリングを施すことによって一層のアクセスページ数の減少、及び1.に於ける多次元クラスタリングの限界改善を図るものである。多次元クラスタリングはリレーションをタプル方向(横方向)に分割するのに対し、トランスポーズはこれを属性方向(縦方向)に分割しており、以降これらを各々タプルクラスタリング、属性クラスタリングと呼び、これらを融合した本手法を属性/タプルクラスタリングと呼ぶ。

2.2 サブリレーションに対するアクセス

selection 演算の実行に際して属性クラスタリングされた複数のサブリレーションをアクセスする場合、1)これらを独立にアクセスしその後tid joinによってselection を実行する方法、及び2)これらサブリレーションに対し selectivityの小さいサブリレーションから始めて前段のサブリレーションのアクセスにより生成されたtid を用いてタプル数を絞り込みながら順にアクセスして行く方法、が基本的である。しかし1)の方法は predicateの相乗的絞り込み効果を活かすことが出来ず、タプルクラスタリング、及び2)の方法による属性/タプルクラスタリングに比較して平均(総)アクセスページ数は常に大きい。従ってここでは2)のアクセス方法を仮定して評価を試みる。

2.3 属性／タプルクラスタリングの評価

属性クラスタリングが有効となるのは、少なくとも最初のサブリレーションアクセスにより生成されたtid の数の平均値（$\bar{\pi}_{1st}$）が充分に小さい場合に限られる。ここで対象となるリレーションRの属性集合を$\mathcal{A}$とし、問合せq 内で selection predicate が施されているRの属性集合、及び selection以外の関係代数演算によって参照されている属性を各々q の selection属性、projection属性と呼びそれぞれ$\mathcal{S}(q)$、$\mathcal{P}(q)$と書く。更にq の分布Q（q）に対し平均的にselection 、projection属性と見做される属性集合を各々$\mathcal{S}$、$\mathcal{P}$と書く。以下、簡単の為、リレーションR（$\mathcal{A}$）を属性クラスタリングにより2つのサブリレーションR（$\mathcal{A}_1$）、R（$\mathcal{A}_2$）に分離する場合を仮定し、タプルクラスタリング、属性／タプルクラスタリングが各々与える平均アクセスページ数$\bar{\pi}$、$\bar{\pi}_t$ の大小関係について、場合に分けて考える。［但し、N_i ：R（$\mathcal{A}_i$）の全ページ数、k_i ：R（$\mathcal{A}_i$）の属性数］

1) $\mathcal{A}_1 \subset \mathcal{S} \cup \mathcal{P}$, $\mathcal{A}_2 \subset \mathcal{A} - \mathcal{S} \cup \mathcal{P}$

この場合はR（$\mathcal{A}_1$）のみがアクセスされる。従って、属性クラスタリングによって

$$\bar{\pi}_t \sim N_1 \cdot \bar{\pi} / N < \bar{\pi}$$

と平均アクセスページ数は線型的に改善される。

2) $\mathcal{A}_1 \subset \mathcal{S}$, $\mathcal{A}_2 \subset \mathcal{P}$

この場合は常にR（$\mathcal{A}_1$）を先にアクセスし、選択された tidを用いてR（$\mathcal{A}_2$）にアクセスする。そこでR（$\mathcal{A}_1$）をタプルクラスタリングした時の平均アクセスページ数、平均アクセスタプル数を各々改めて$\bar{\pi}_{1st}$、$\bar{\tau}_{1st}$とし、この2つの値によって更に場合分けを行なう。

ア） $\bar{\pi}_{1st} \sim 1$、$\bar{\tau}_{1st} \sim 1$の時

これはR（$\mathcal{A}_1$）に対する問合せ分布（$Q^1(q^1)$）がexact match に近い分布をしている場合に相当し、R（$\mathcal{A}$）のタプルクラスタリングは$\mathcal{A}_1$の属性のみによって行なわれる。従って

$$\bar{\pi} \sim 1、\ \bar{\pi}_t \sim \bar{\pi}_{1st} + \bar{\tau}_{1st} \sim 2 > \bar{\pi}$$

となり、分離は不利となる。

イ） $\bar{\pi}_{1st} \sim 1$又は$>>1$、$\bar{\tau}_{1st} >> 1$の時

この場合、一般に次式が成立する。

$$\bar{\pi}_t \sim N_1 / N (1 + v) \bar{\pi}$$

$$[1 \leq v \leq N \cdot V / N_1]$$

$v \sim N \cdot V / N_1$ の時は一般的なrange query が発行される場合で分離は不利となる。一方 vが十分に小さい場合は分離が有利で、これは$Q^1(q^1)$がselectivity の小さい、一様なpartial match query の分布である場合に相当する。

ウ） $\bar{\pi}_{1st} >> 1$、$\bar{\tau}_{1st} \sim 1$の時

これは$Q^1(q^1)$が一様な低次元partial match query に近い分布をなし、各属性のselectivity が十分に小さい場合に相当し、

$$\bar{\pi} \sim N^{1-\frac{1}{k_1}}、\ \bar{\pi}_t \sim N_1^{1-\frac{1}{k_1}} + 1 < \bar{\pi}$$

となって属性クラスタリングにより平均アクセスページ数は減少する。

3) $\mathcal{A}_1 \subset \mathcal{S}$, $\mathcal{A}_2 \subset \mathcal{S}$

ア） R（$\mathcal{A}_1$）、R（$\mathcal{A}_2$）に対するアクセスパタン（問合せによって参照される属性集合）が互いに素である時

これは 1）の様な分布を対称的に重ね合せた場合に相当し、 1）と同様な議論により分離によって線型的改善が得られる。

イ） R（$\mathcal{A}_1$）、R（$\mathcal{A}_2$）に対するアクセスパタンが互いに素ではない時

この場合、$\bar{\tau}_{1st}$が小であり$\bar{\pi}_t < \bar{\pi}$である可能性があるのは 2）の様な分布を重ね合せた場合に限られ（問合せが平均的に両者にまたがる場合は絞り込み効果を活かせない）、 2）と同様な議論によって$Q^2(q^2)$が一様なselectivity の小さい partial match query で任意のq に対し$\mathcal{S}(q) \subset \mathcal{A}_1$ 又は$\mathcal{S}(q) \subset \mathcal{A}_2$ である時に限って分離が有効となる。

以上の議論から、タプルクラスタリングのみの場合に比較して属性／タプルクラスタリングが有効となるのは次の2つの場合であるとの結論が得られる。

t1） アクセスパタンが互いに素な属性集合が存在する場合。

t2） selection predicate の分布が一様なpartial match queryであり、各属性の selectivityが比較的小さく、且つpredicate 中、値が指定される属性集合が互いに素な場合。

これ以外の場合については、タプルクラスタリングが充分なクラスタリング特性を保証する。又、特に2）の場合、属性クラスタリングにより各サブリレーションのクラスタリング特性は向上し、タプルクラスタリングの限界が解消される。

3. おわりに

リレーションを属性方向、タプル方向縦横にページ分割する属性／タプルクラスタリングについて考察した。現在、より詳細な評価、実装方法等について検討中である。

［参考文献］

［1］伏見他、情処全大第26回、4f-3、1983年