拡張可能DBMS:COMMONの格納構造について

1L−1

宝珍 輝尚
NTT情報通信処理研究所

1．はじめに

データベース管理システム(DBMS)の拡張性を高める研究が盛んに行われ[1]、この中のDBMS generatorアプローチには様々なデータモデルを実現する可能性がある[2]。試作中のDBMS構築システムCOMMON[3,4]でも多種の意味的なデータモデルの実現をめざしている。

ここで、DBMSの扱う格納構造はデータモデルの実現に大きな影響を及ぼし重要であるが、多種の意味的なデータモデルの実現をめざした格納構造の提案は皆無である。意味的なデータモデルは様々なタイプ構成子およびタイプ間の関連を用いてデータの持つ意味を構造的に表現するモデルであり、格納構造にグラフを使用するのが自然である。ここで、データアクセス経路から独立な高水準DB言語を、従来のグラフのためのデータ構造に基づいて実現しようとすると、条件式に現れる属性のグラフ上での位置に条件式評価性能が依存し、一種のデータ従属になる。

本稿では、この問題を条件式評価が属性のグラフ上での位置に依存しない枝中心のデータ構造で解決する。COMMONでは本データ構造を格納構造に採用する予定である。以下、2.で前提とするデータのグラフ表現と従来のデータ構造を示し、3.で格納構造を提案し、4.で評価を行う。

2．前提

2.1 データのグラフ表現

データを表現するグラフ及び集合属性を定義する。

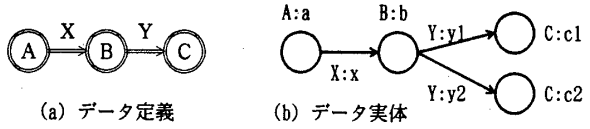
〔定義1〕属性名と値を点および枝のラベルに持つラベル付きグラフをデータ表現グラフという。

〔定義2〕データ表現グラフ中に同一属性名の要素が2以上存在するとき、この属性を集合属性と呼ぶ。集合属性でない属性を集合属性と区別するために単純属性と呼ぶ。

グラフ表現の定義と実体の例を図1に示す。図1(b)の属性Cの点は複数であるので属性Cは集合属性である。

2.2 従来のデータ構造

グラフのための従来のデータ構造はソースノード中心である。すなわち、ソースノードから該ノードに接続している要素(点および枝)を順々に検索してゆく。この概要を図2に示す。属性Bの点の値(b)は属性Aの点(ソースノード)から枝Xを経由し、該点に到達して初めて得られる。



(a) データ定義　(b) データ実体

図1　データ表現グラフの例

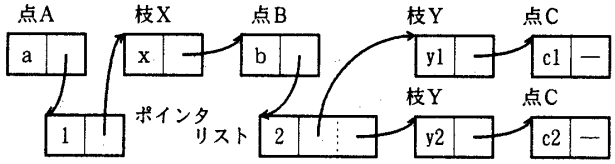


図2　従来のデータ構造

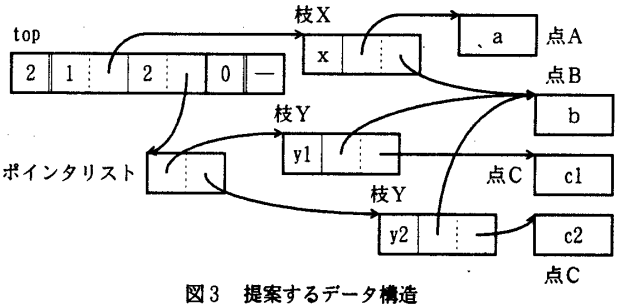


図3　提案するデータ構造

3．COMMONの格納構造

3.1 要求条件

従来のデータ構造は、構文解析木のように構造が前もって決定不可能の場合に良く使用される。これに対し、データベースでは前もってスキーマによりデータのグラフ構造か決定されてしまい、上記の構造の柔軟性は不要である。

一方、従来のデータ構造では、ある属性のデータ値に関する条件式評価は該属性に到達するまで不可能であり、条件式評価性能がデータ表現グラフ上の要素の位置に依存する。ここで、高水準なDB言語をサポートし高度なデータ独立性を実現するDBMSの格納構造としては、条件式中の属性のデータ表現グラフ上の位置という物理的な変数に条件式評価性能が依存するのは好ましくない。

そこで、構造は多少固定的でも条件式評価が要素のデータ表現グラフ上の位置に依存しない格納構造が望まれる。

3.2 枝中心のデータ構造

3.1の要求条件を満たすために、COMMONでは枝中心のデータ構造を格納構造に採用し、条件式評価の性能を点の位置から独立とする。

Storage Structure for Extensible DBMS: COMMON
Teruhisa HOUCHIN
NTT Communications and Information Processing Laboratories

提案するデータ構造の構成要素は、top，枝実体，点実体，ポインタリストである。topは 枝の定義数，枝ポインタエントリ，孤立点の数，孤立点へのポインタリストへのポインタを格納する。ｉ番目の枝ポインタエントリは定義番号ｉの枝の枝実体の数と枝実体へのポインタを格納する。ただし、枝実体数が２以上の場合はポインタリストを指し、ポインタリストが各枝実体を指す。枝実体は、枝の値，始点と終点へのポインタを格納する。点実体は点の値を格納する。図１のグラフの本データ構造での表現を図３に示す。

４．評価

4.1 性能評価

条件式評価時間とデータ返却時間で評価する。グラフは線状とし、属性は単純属性とした。測定は、SUNﾜｰｸｽﾃｰｼｮﾝ(SPARC SLC)においてメモリ上，データ数100，データ長20バイトで行った。

(1) 条件式評価時間　グラフを構成する点の数を11，枝の数を10(固定)とし、条件式中の述語数と述語に出現する最遠点を変化させて評価した。条件式は最後に評価される述語で条件を満たさなくなるものとした。この結果を図４に示す。従来のデータ構造では最遠点に比例して評価時間が増加する。提案したデータ構造では最遠点に無関係である。

(2) データ返却時間　無条件にデータを取り出し返却時間を測定した。ここでは、定義上の要素(点および枝)の数を変化させて評価した。この結果を図５に示す。提案したデータ構造は従来のデータ構造に比べ性能は良くない。両データ構造ともに格納データ数には依存しない。

4.2 メモリ量評価

定義上の要素数を変化させ、利用者のデータ以外に必要な領域の一要素当りの平均メモリ量で評価した。この結果を図６に示す。提案したデータ構造は要素数が多くなると一要素当りのメモリ量が減少する。これは、要素数が多くなると 一要素当りのオーバヘッドが少なくなるためである。

4.3 考察

データ表現グラフの点の数は、関係モデルの属性数に対応する。従って、点の数が３以下(すなわち、最遠ノードが３以下)ということはほとんどない。さらに、応用プログラムでデータを検索する場合は、高速化のためにインデックスを付与するが、インデックス検索で評価できない述語評価の性能は、従来のデータ構造では、4.1.1で測定したように属性のデータ表現グラフ上の位置に依存してしまう。さらに、これらの属性はインデックスキーを構成しない属性であり、ソースノードから遠い点が多く、検索ヒット率が低い場合は性能劣化が問題となる。提案したデータ構造は、条件式評価性能が属性のデータ表現グラフ上の位置に依存せず、データ定義要素数に従って条件式評価性能を見積ることが可能である。また、属性のデータ表現グラフ上の位置によっては従来のデータ構造よりも性能が良い。

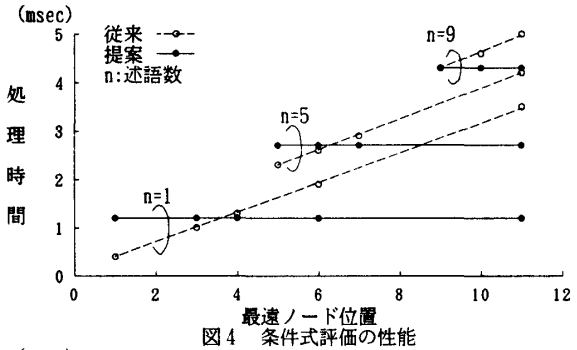


図4　条件式評価の性能

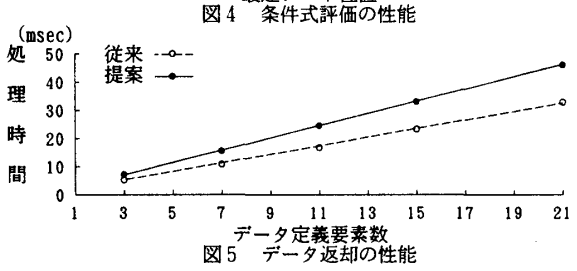


図5　データ返却の性能

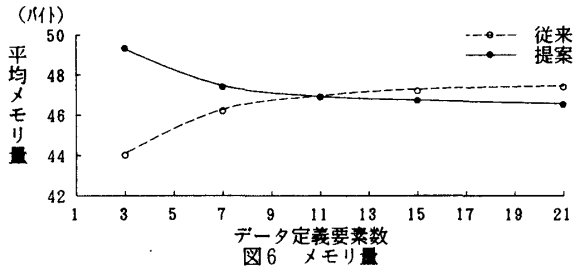


図6　メモリ量

さらに、想定しているデータ定義要素数(11以上)では、提案したデータ構造の方が少ないメモリ量で実現できる。

５．おわりに

ＤＢＭＳ構築システムＣＯＭＭＯＮで採用予定のグラフに基づく格納構造について述べた。本格納構造は、枝を中心にしたものであり、条件式に現れる属性のグラフ上の位置に条件式評価性能が依存せず、高いデータ独立性の要求される高水準なＤＢ言語のサポートには不可欠である。

データ挿入時の性能，巡航検索（これは、集合属性が存在する場合に必要であり、高水準なＤＢ言語を使用せずユーザにデータの定義グラフを計算機で直接表示し検索要求を行わせる場合に必要となる）を行った場合の性能はあまり良くない。これらに対する対処と提案した格納構造の各種データモデルへの適応能力の評価等が今後の課題である。

参考文献

(1) D.S.Batory and M.Mannino," Panel on Extensible Database Systems", Proc. ACM SIGMOD 1986.
(2) S. Hong and F. Maryanski:" Using a meta model to represent object-oriented data models", Proc.6th Intl.. Conf. on DATA Engineering, pp.11-19(1990).
(3) 宝珍,"拡張可能ﾃﾞｰﾀﾍﾞｰｽ管理ｼｽﾃﾑ構築についての一考察",第40回情処全大 5H-1.
(4) 宝珍,"拡張可能DBMS:COMMONの基礎的性能評価",第41回情処全大 1D-8.