

カラム属性を利用したデータベース高速圧縮方式

6X-3

上田 尚純 郡 光則 青野 正宏 渡辺 尚 水野 忠則
 (三 菱 電 機) (東京工業高等専門学校) (静 岡 大 学)

1. はじめに

ディスク領域の削減や文書送信の高速化の目的でテキスト文書の圧縮の利用が普及してきているが、データベース (DB) への圧縮の利用はまだ限定的である。データ量の大きな DB にこそ圧縮の利用価値は大きいはずだが、圧縮の利用が限定的なのは次の理由によるものであろう。

- ・圧縮処理は CPU 負荷が高く、大量データを持つ DB 圧縮には相当の時間がかかる。
- ・DB は日々更新され、圧縮の回数も頻繁となる

文書テキスト用の圧縮プログラムは各種あるが、いずれも汎用用途を意図しており、圧縮処理に時間がかかる。ところで、リレーショナル DB ではデータは表形式に整理されており、これを利用すればより高速な圧縮を行える。DB のデータ量を圧縮できれば、大量の明細データのディスク読み出しが必要なデータウェアハウス問合わせ処理の高速化や、通信回線での DB 転送の高速化などが可能となる。

既に参考文献 3) で DB を高速に圧縮する手法が示されている。DB 中のデータ項目を文字列というよりはオブジェクトとして捉え、日々の売上伝票などのトランザクション系の詳細レコードには同じオブジェクトが頻繁に出現することを利用して、カラム単位で、オブジェクト名を対象とする簡略化した辞書生成を行う事で、高速且つ高い圧縮率で DB を圧縮できる。辞書にはオブジェクトの名称を示すテキストをまるごと、あるいは固定長で区切ったものを入れる。この方法は DB のデータの特性をうまく利用しているが、個別のカラム属性を積極的に利用する所まではしていない。本論文では文献 3) の手法を土台として、更にその上にカラム属性を利用した圧縮を行うことで、一層の処理の高速化や圧縮率向上を実現できることの可能性を述べる。

2. DB のデータの特性と圧縮の方向

DB のテキストデータは一般の文書と比較して次のような特性を持っている。

- ・表形式に整理して格納されている。表は複数のカラム (フィールド) で構成され、各カラムには同様な属性を持つデータ値が入る。データ値にはテキストと数値があり、テキストは大半が名詞である。
- ・一旦 DB の運用が始まると、データ値の日々更新や追加はあるものの、短期間にデータが大幅に入れ替わることは稀であり、データの

集合が持つ属性は時間的に持続性が高い。

DB のデータは文章構造ではないため構文解析はほとんど不要である。カラム長やデータタイプ (文字か数値かなど) も前もってわかる。また、DB 利用者の環境が急変しない限り、代表的な項目である商品、顧客や取引先、組織や社員といった多くのデータ値も短期間での急変は考えられず、継続性が高いと考えられる。圧縮の観点からは、反復して出現するテキスト (多くはオブジェクトの名称) が多く、また一度作成した圧縮用辞書の再利用性が高いと期待できる。

文献 3) で示されているように、DB の圧縮は行方向よりもカラム方向に圧縮した方が高速に行える。一般に DB のデータはレコード単位で行方向にディスクに記録されるので、データをブロック単位でメモリ上のバッファに読み込み、次いでブロック内のデータに対してカラム単位でカラム方向にデータを圧縮する。

(図 1) 伸長の場合はこの逆に、カラム単位で圧縮されたデータを 1 ブロック分単位で受け取り、カラム毎に伸長して圧縮前のブロックの形式に復元する。

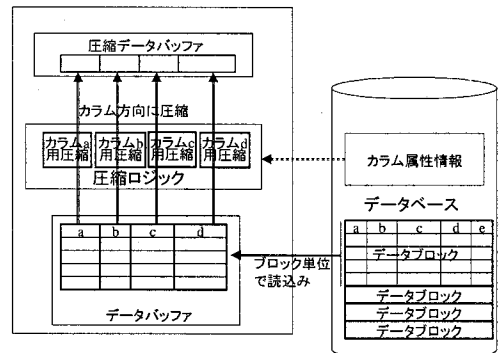


図1 カラム属性を利用した圧縮方法

3. カラム属性の利用

利用者の用途に応じて様々な DB があるが、表のカラムには同様な項目が頻繁に出現する事に気づく。手に入る以下の DB でこの調査を行った。

- ①実業務で使用している DB
- ②マイクロソフト社製 DB ソフト Access に付属しているサンプル DB
- ③情報処理技術者試験 DB スペシャリストに出題された DB

企業などで運用されている DB は秘匿性が高く調査は困難であり、上記の②と③を調査対象に含めた。

これらは実用の DB ではないが、公開されており誰でも参照可能なこと、実用 DB の代表的特徴を模倣していると考えられることから、調査対象とした。調査結果を表 1 に示す。表 1 のデータ項目に対して高速の圧縮手法を用意しておけば、DB 全体の高速圧縮が期待できる。表 1 に出てこない類型化困難なデータ項目に対しては、汎用の圧縮手法を用いるか圧縮の対象外にするか等の対処が考えられるが、出現頻度が少なければ実用上は問題とならない。

表 1 の分類区分について補足する。これはトランザクションを処理するアプリケーションにおける処理の区分けや、主キー項目の属性を示すのに利用される。特徴として、その中に出現する値がいくつかの値に限定され、同一値の反復出現率が高く、高い圧縮率を達成できる。

DB 圧縮にカラム属性を積極的に利用するのは、文書の圧縮プログラムが汎用化を指向するのは逆方向であるが、以下の理由でこれを正当化できる。

- ・カラム属性は類型化されるため、類型の個数分の圧縮方法を用意すればよい。
- ・DB の表やカラムは設計を終えて一旦運用を開始すると長期間利用し続ける。
- ・DB のデータは時間軸で見ると短時間で急激には変化せず、同じ特性を示し続ける。

DB を設計して運用開始する時に、各カラム毎に適用する圧縮方法を指定しておけばよい事になる。

4. カラム属性利用の具体例

カラム属性を利用することで圧縮に役立つ例を以下にいくつか示す。

①前回圧縮時の知識、辞書の利用。データは短期間で急に変わる事は稀であり、以前の圧縮で作成した辞書を保存しておき、それを再利用する。DB のデータ更新に伴い保存辞書の効果は徐々に劣化するが、これは一定間隔で辞書を作り直す事で対処する。また、辞書の有効度を示す統計データを探り、この値がある閾値より悪くなった時点で辞書を作り直すような対策も取りうる。

②表の主キーとして順序番号がよく利用されている。例えば売上伝票など。この場合、開始番号と行毎の刻み値がわかれば、レコード件数に関係なく数バイト以内に圧縮できる。カラム属性を利用することが有効に働く典型的な例である。

③カラム間の関係の利用。例えば、納期日が受注日の 1 週間後と決まっている場合は、納期日のカラムは受注日のデータから算出できるので圧縮は必要なくなる。郵便番号と都道府県名、氏名や製品名の漢字表記と読みなども、カラム間の関係を利用できる。

④氏名、地名、住所、固有名詞などの辞書を利用。氏名や住所などはカラムとして頻繁に利用される。汎用

表 1 使用されるカラム属性の分類

	分類	例	出現頻度(%)
テキスト	人名	社員、学生、担当者、患者 など	5
	会社名	本社名、顧客、仕入先 など	6
	商品名、サービス名	商品、仕入れ品、中間品、材料 など	7
	命名した名称	表題、プロジェクト名など	1
	道具・設備	器具、建物、製造機械 など	1
	住所	県名、フル住所、建物名 など	8
	所属名、職位	支店名、部署名、職位 など	3
数字	識別番号	IDコード、順序コード、分類コード など	25
	電話・FAX番号	人名、会社名に対応した電話番号など	3
	郵便番号	本社名、顧客、仕入先 など	2
	金額	単価、総額、預金額 など	8
	計量	個数、計量 など	5
	日付	誕生日、受注日、締切日 など	11
	分類区分(テキストか数字)	色、形状、重さ、性別 など	13
	注釈	人、会社、伝票などに付随する注釈	1

の辞書を利用できるようにしておけば、圧縮毎に圧縮用辞書を作成する必要がなくなる。

⑤高出現単語用辞書の用意。製品やサービスに関連して、仕入れ品や中間製品なども含めると、ある範疇の単語が高い頻度で出現する。例えば、食品メーカーなら食べ物に関する単語がよく利用される。これら単語の辞書を予め作成しておき、最長一致法で探索すれば、効率よい圧縮が期待できる。

⑥カラム内データの分割。1つのカラム内のデータの表記形式が決まっており、データが複数の要素から構成されている場合、要素単位に分割して処理すれば高速化できる。例えば 2001/07/30 という形式の日付の場合、年と月と日に分割して各々を圧縮すれば、同じ年や月や日は連続して出現するので、圧縮効果を上げる事ができる。

5. あとがき

文書テキスト用圧縮プログラムでの圧縮速度では 2MB/秒前後の処理速度しか得られないが、テストデータによる評価では、カラム属性を利用すればより高い圧縮率を 10 倍以上の処理速度で実現できている。詳細な評価を行っているところである。また、より多くの DB の事例を調べ、カラムの類型化を深化させる努力も今後必要である。

参 考 文 献

- 1) Ziv,J. and Lempel,A., "A Universal Algorithm for Sequential Data Compression", IEEE Tras. on Information theory, Vol.23, No.3,pp337-343, May 1977
- 2) Ziv,J. and Lempel,A., "Compression of Individual Sequences via variable-rate coding", IEEE Tras. on Information theory, Vol.24, No.5,pp530-536, Sep 1978
- 3) 郡 光則、佐藤 重雄 他、"大規模データ処理向け超並列可逆データ圧縮伸長処理技術の開発 "、IPA 成果報告書 (2000)