成長型ハッシュ検索法の検討

1H-9

森　達男

ＮＴＴ電気通信研究所

１．まえがき

本検討では、オンラインでのファイル拡張法[1]および再編成法等を付加したより実用的な成長型ハッシュ検索法についての提案をする。検討内容としてはファイル再編成技術（再編成時での新たなデータ格納・検索位置をアドレッシングする技術）、検索時間の高速性保証技術について報告する。

２．ファイル再編成技術での前提条件

（１）ハッシュ検索での検索キー‥‥ボックスをデータ格納箱と定義し利用者毎に一つ割り当てる。汎用性の高さと通信システムでの検索キーという２点を重点に考え、ボックス利用者（特定小集団：ｍ）は電話網加入者を母集団（ｎ）とする。　ここで、ｍ／ｎの割合が小さいと検索キーのランダム性は低くなるため、バラツキを持つボックス利用者番号（電話番号）にハッシュ法を作用させてランダム性の高いシーケンシャル番号に変換する。

（２）ハッシュ関数のロジック‥‥‥集中域とローカル域についての電話加入者番号傾向は①～⑧の桁のランダム性の順位は①②③⑤となる。これより、①②③⑤の桁を利用する数字分析法を採用する。

（集中域電番）　$03-\underset{⑦}{X_7}\underset{⑥}{X_6}\underset{⑤}{X_5}-\underset{④}{X_4}\underset{③}{X_3}\underset{②}{X_2}\underset{①}{X_1}$

（ローカル域電番）$04\underset{⑧}{X_8}\underset{⑦}{X_7}-\underset{⑥}{X_6}\underset{⑤}{X_5}-\underset{④}{X_4}\underset{③}{X_3}\underset{②}{X_2}\underset{①}{X_1}$

また、①②③⑤の桁を利用する数字分析法を採用した場合のハッシュ関数の基本型は

$f(X)=X_1\times10^3+X_2\times10^2+X_3\times10+X_5$　となる。

（３）ファイル配置と格納条件‥‥‥本方式でのファイル配置を図１に示す。ボックスファイルの手前にブロックファイルを位置させ、ハッシュ法によりブロックをハントし、さらにブロック内の検索テーブル（３節内記述の数値分布表とＶアドレス表）より該当ボックスをサーチする。ボックスファイルはハッシュ法のロジックの影響を受けない構造としているため、ファイル単位での増設（拡張）が可能となる。また、ファイル再編成対象のファイルはブロックファイルである。

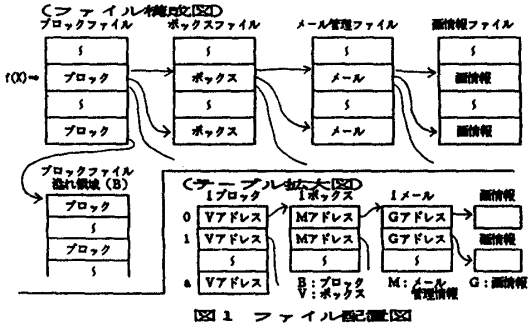

図１　ファイル配置図

３．高速データ検索法（数値分布比較法の改良）

大規模データ蓄積システムでのデータ高速検索技術である数値分布比較法の導入については文献(1) で既に報告しており、本検討では、その改良として数値分布表に必要なデータ量の圧縮を図った。数値分布比較法で比較照合する桁は数字分析法（ブロックのサーチロジックに使用）では未使用の桁（例：４桁目、６桁目）を用いる。ブロック内の数値分布表の構造を図２に示す。各桁の取り得る値（０～９）を行とし、ブロック内の管理ボックス番号（０～ａ）を列としたマトリックス（S_{ij}：$0\leq i\leq 9$、$0\leq j\leq a$ ）を考える。$S_{ij}=1$ はｊ番目のボックスのＳ桁目の値がｉであることを示し、$S_{ij}=0$ はＳ桁の値はｉではないことを示す。S_{ij}のビットのセットはボックス生成時に、リセットはボックス消去時に行う他、数値分布表はブロック内に設置する。

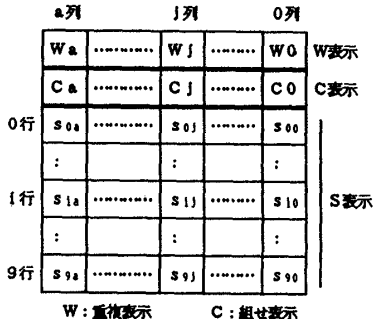

図２　数値分布マトリックス

（１）数値分布表の個数条件‥‥‥数値分布比較法での照合桁数は２桁とし、検索範囲は1/100 程度に絞ることを可能とさせる。また、ブロックのボックス管理数は128(ａ=127に対応）個程度を考える。照合桁数が３桁以上だと数値分布のランダム性が低い桁が対象となったり、空塞表の設定に処理ステップがかかる。

（２）数値分布比較法の改良ロジック‥‥‥ボックスが既に作成済みかの判定は、比較対象桁（S_1 桁、S_2 桁）の値が同じであるボックスが存在するか否かを数値分布表より調べる。利用者番号のS_1 桁目の値がｊであるボックスのサーチには空塞表で“１”が立ったbit 位置を見つける。

【改良点】これまでの数値分布法では、対象桁単位に数値分布表を用意していたが、改良型数値分布法ではデータ圧縮のため１個の数値分布表で圧縮表現する方法を採る。

①S_1 桁目の値がｊ、S_2 桁目の値がｋである利用者番号のサーチは　　$P=(S_1\ \mathrm{AND}\ S_2)$

STUDY ON GROWING-TYPE HASHING METHOD

Tatsuo MORI

NTT Electrical Communications Laboratories

の論理積演算を行い、Pに“1”が立っているビットがあるか否かで存在の判断をする。
②j＝Kの場合には矛盾を生じない様にその旨を示す重複表示（図2内のW表示の該当ビットに1を立てる）を数値分布表示に用意する。利用者番号のサーチは　　T＝（P　AND　W）
の論理積演算を行いTに“1”が立っているビットがあるか否かで存在の判断をする。
③（S_1、S_2)の組合せは(j,k) と(k,j) の2通りあり、組合せ表示（図2内のC表示はj＞kの場合に該当ビットに“1”を立てる）を用意する。利用者番号のサーチは　　T＝（P　AND　C）
の論理積演算を行いTに“1”が立っているビットがあるか否かで存在の判断をする。
【データ圧縮効果】旧方式では数値分布表（10a bit）は2個であったことより、合計20a bit が必要であった
新方式では（10＋2）a bitでよいため、8a bitのデータ量削減となる。
（3）ボックスの新規作成法……上記（2）の方法で該当ボックスが未作成であることが判明した場合には関連テーブル（数値分布表、ボックス空塞表等）に制御情報を記憶させボックスを生成する。

4．オンラインでのファイル再編成条件

2節の前提条件（3）より、ファイル再編成はブロックファイルを対象とする。ここでは、ファイル再編成技術として、①プログラム処理でサイズ見合いブロックファイルの自動選択技術、
②再編成時での新たなデータ格納・検索位置のアドレッシング技術、　について述べる。

（1）成長型ハッシュ関数設定条件……規模見合いのハッシュ関数の設定法として次の2案が考えられる。案1は処理が複雑で、再編成の自由度が低いため、案2が秀れていると考える。案2の場合にはファイルサイズ種別の記憶用テーブル（図3参照）を設ける。テーブル内設定情報として、旧ファイル（基本部ブロック）か新ファイル（再編成部ブロック）かのファイル識別子、ブロックアドレスを記憶する。
【案1】予めハッシュ関数を成長段階数分だけ用意しておき、再編成時には新ハッシュ関数および新ファイルへの切替えを行う。
【案2】ハッシュ関数は変更しないで、ファイルサイズ種別を記憶するテーブルを新たに用意する。

（2）ブロック内各種テーブルの再設定条件……各種テーブル（数値分布表、ボックスアドレステーブル等）はファイル再編成時にはテーブルの入れ替えが必要となる。2つのファイル再編成案および、両案の処理を図3に示す。また、両案の比較を表1に示す。この比較より案2が優れていると考える。
【案1】細胞分裂タイプ……同一ハッシュ値を持つボックスを群と定義し、ブロックを群の数に等分割する。
ブロック内で該群が満杯になったら、該群を別ブロックに移して拡張させる。
【案2】成長時移転タイプ…ブロック内の各種テーブルを成長時にさらに大きなエリアに移転させる。

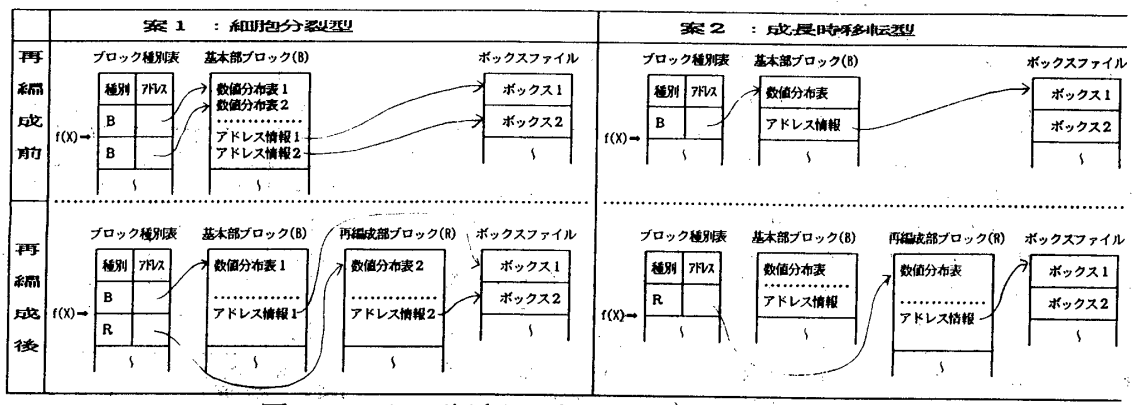

図3 ファイル再編成処理案　B：基本部ブロック　R：再編成部ブロック

表1　ファイル再編成案の比較

項　目	案1（細胞分裂型）	案2（成長時移転型）
ファイルの有効利用度	○ 再編成の場合でも基本ファイルが利用できる	△ 再編成後、該ブロックの旧エリアの利用不可。
ファイル再編成の自由度	△ ファイルがN分割のみ可。Nは整数	○ ファイルは独立となる。拡張の比率は任意。
再編成用処理の難易度	△ 数値分布表の再編成が必要	○ 数値分布表の移し替えのみ。
ロジックへの影響度	△ ファイルがN分割を意識するため若干影響する	○ 影響しない
評　価	△	○

（3）ファイル設定条件……ファイル再編成をしない間は溢れ領域（図1参照）を使用する。また、再編成時には旧ファイルの基本部ブロックと合せて溢れ領域のブロックも再編成ブロックに移し替える。

5．あとがき

……成長型ハッシュ法の検討としてオンラインでのファイル再編成基本技術、大規模データ蓄積システムでの高速データ検索技術を中心に述べた。今後は、より多段成長型のハッシュ検索法の検討を進める。

[文献]（1）森、南、佐々木“ファイル拡張を考慮したハッシュ法の検討”昭60信学情報・システム全会 N0585