Vol. 29 No. 5 情報処理学会論文誌 May 1988

分離連鎖法における挿入・探索アルゴリズムの解析†

中村良三†† 税所幹幸†††

分離連鎖法を用いた見出し探索法で，分離あふれ領域のバケットの大きさを考慮し，かつ見出しの探索頻度という重みをつけて一般化したモデルにおける挿入・探索アルゴリズムの系統的な解析は明らかでない．すなわち，バケットに格納可能なレコード数を表すバケットサイズを任意の大きさに取ったとき，任意の順序で登録される個々の見出しが，どのバケットにどのような確率で配置されるかという挿入アルゴリズムの解析から，個々の見出しの探索頻度を考慮した探索コストすなわちアクセス回数を評価する探索アルゴリズムの系統的な解析は明らかでない．本稿では，この問題に対する挿入・探索アルゴリズムの系統的な解析を示す．はじめに，挿入アルゴリズムの解析では，任意の順序で登録される見出しが，どのバケットにどのような確率で格納されるかを明らかにする．次に，探索アルゴリズムの解析では，バケットサイズと個々の見出しの探索頻度を考慮した探索コストの評価式を導出する．

1. まえがき

データの探索では，主として各々のデータを識別する見出しを指定して，適合するデータを探す方法，すなわち見出し探索法が用いられる．これらの見出し探索法の性能は，主に，ある見出しを探すために要する他の見出しとの比較回数を尺度とした探索コストと，情報を格納する領域および管理に要する領域を合わせた所要記憶容量とによって評価される．特に探索コストは，見出しが探索される頻度とその見出しを探索するに要する路の長さすなわち比較回数との積の和で定義される．このとき，見出しを探すために要する路の長さは，見出しが配置される場所に左右され，またその置かれる場所は見出しが登録される順序に依存する．

見出し探索法のひとつである分散記憶法は，見出しに依存して生成される番地によって，その格納場所を決定し，探索時には，挿入時と同様な手続きによって見出しからその格納場所を探し目的のレコードを探す方法である．

分散記憶法の中の分離連鎖法を用いた見出し探索法では，各見出しの探索頻度を一様と仮定すれば，見出しの登録順序に関係なく探索コストが決まる．したがって，各見出しの探索頻度を一様と仮定した従来の解析では，個々の見出しが配置される可能な場所を考慮にいれる必要がなく，モデルの設定が単純化されている[2),3)]．その結果，従来のアルゴリズムの解析は，一般性に欠けるとともに，現実の現象を適切に評価できないことを文献 4), 5) は指摘している．

本論文では，分離連鎖法において，バケットサイズすなわちバケットの大きさを考慮すると共に見出しの探索頻度という重みをつけて一般化したモデルのもとでの挿入・探索アルゴリズムの解析を提示する．

はじめに，挿入アルゴリズムの解析では，任意の順序で登録される見出しが，どのバケットにどのような割合で配置されるかを厳密に評価できる確率を導出する．次に，探索アルゴリズムの解析では，挿入アルゴリズムの解析で導出した確率をもとに，バケットサイズと個々の見出しの探索頻度とを考慮にいれた探索コストすなわちアクセス回数の評価式を導き出す．特に，導出した評価式で探索頻度を一様と仮定すれば，探索コストの評価式は，バケットサイズと表占有率のみの簡潔な式で表現できることを示す．

以下，2章では，分散記憶法を用いた見出し探索法について概観する．3章では，レコードを格納するバケットの大きさを考慮すると共に探索頻度という重みをつけて一般化したモデルにおける挿入・探索アルゴリズムの解析を示す．4章においては，従来の解析と比較検討する．

なお，本稿では，分散表の大きさを M，バケットサイズを b で表し，N 個の見出しからなる成功時の平均探索コストを S_N，その分散を V_N とする．また各見出しの探索頻度をその登録順に従い ρ_i $(i=1, 2, \cdots, N)$ とする．

† An Analysis of Insertion and Search Algorithms of Separate Chaining Method by RYOZO NAKAMURA (Department of Electrical Engineering and Computer Science, Faculty of Engineering, Kumamoto University) and MOTOYUKI SAISHO (Department of Information and Electronics Engineering, Yatushiro National College of Technology).

†† 熊本大学工学部電気情報工学科
††† 八代工業高等専門学校情報電子工学科

2. 分離連鎖法

2.1 分散記憶法および分離連鎖法の性質

分散記憶法は見出し探索の基本的技法のひとつであり，その適用分野はコンパイラやアセンブラにおける記号表の操作，データベースの検索，あるいは構造データの高速同定や部分的構造の共有の判定など多様にわたっている．

分散記憶法は，レコードを登録するときには，レコードの固有な見出しに対して演算を行い，その演算結果をそのレコードの格納場所の位置付けに使用し，探索するときには，登録時と同様な手続きによって，目的のレコードを探す方法である．この見出しに対する演算結果は，計算機の記憶領域に実現された表すなわち分散表の指標を示す番地となる．このように，見出しから番地を生成する演算を行う関数を分散関数といい，この番地を分散番地という．また，生成されたひとつの番地からアクセスされるレコード格納用領域の一分割単位をバケットといい，バケットサイズとはひとつのバケットに格納可能なレコードの総数を言う．

この分散記憶法の基本的な考え方は，分散関数を適当に選択することによって，レコードを分散表になるべく一様に分散して格納し，探索時には，少ない探索回数で目的とするレコードを探索しようとするものである．

ところで，見出しを分散番地に変換する際，生成可能な見出しの集合は利用可能な分散番地の集合に比べて非常に大きい．その結果，理想的な分散関数が作られたとしても，異なった見出しが分散表の上で同じ番地になる現象が生ずる．このような現象を衝突という．この衝突を処理する方法には，衝突を起こした見出し同士をポインタで繋ぐ連鎖法と新たな格納番地を計算によって求める計算法に大別される．分離連鎖法は前者に属する方法である．

分離連鎖法では，同じ分散番地になる見出しすなわち同族の見出しを，分散表とは別に設けたあふれ領域に，その分散番地を探索根とする一つの線形リスト(以下リストと書く）として構成する．すなわち分散表には，同族の見出しからなるリストの先頭を指す番地すなわちポインタを格納し，リストの各要素となるバケットは見出しを含むレコードを格納する領域と次の要素を指すポインタ領域とからなる．

2.2 挿入・探索アルゴリズム

分離連鎖法においては，あらかじめある大きさの分散表を配列などで構成し，分散表のポインタの値をすべて何も指していない状態に設定する．また，バケットは必要に応じて動的に生成されるものとする．このような設定のもとで，任意の順序で登録される見出しの挿入アルゴリズムは次のようになる．

[A.1] 見出しを分散関数により，分散表の指標に置換し，この指標からポインタの値を求める．

[A.2] ポインタの値が何も指していないならば，新たなバケットを生成し，ポインタの値をそのバケットを指す番地とし連結する．そして，生成したバケットのポインタ領域を何も指していない状態に設定する．

[A.3] バケットに空きがあるか調べる．空きがあれば，空いている左端から順次レコードすなわち見出しを格納する．空きがなければ，そのバケットのポインタの値を求め[A.2]に帰る．

次に，ある見出しを探索するときには，前述した挿入アルゴリズムによって構成されたデータ構造に対して，バケット単位でアクセスする方法を考える．したがって，見出しを探す場合のアルゴリズムは次のようになる．

[B.1] 見出しを分散関数により，分散表の指標に置換し，この指標からポインタの値を求める．

[B.2] ポインタの値が何も指すものがなければ，探索は失敗し，不成功探索となる．

[B.3] ポインタが指すバケットに探す見出しを含むレコードが存在すれば，探索は成功し，成功探索となる．そのバケットに探す見出しがなければ，そのバケットのポインタの値を求め[B.2]に帰る．

3. 挿入・探索アルゴリズムの解析

バケットサイズを考慮に入れ，かつ見出しの探索頻度という重みをつけて一般化したモデルにおける挿入・探索アルゴリズムを解析する．

3.1 挿入アルゴリズムの解析

見出しが挿入される場所は見出しが登録される順序に依存する．したがって，任意の順番で登録された見出しが，リスト上のどのバケットにどのような確率で

配置されるか，その挿入アルゴリズムの厳密な解析を示す．

解析に先立ち，理想的な分散関数によって，N 個の見出しを大きさ M の分散表に一様に分配するとき，分散表に連結されるあるリストの見出しの個数が k になる．すなわち同族の見出しが k 個になる確率 P_{Nk} は次のようになる．

$$P_{Nk}=\binom{N}{k}\left(\frac{1}{M}\right)^{k}\left(1-\frac{1}{M}\right)^{N-k} \qquad (1)$$

このとき，前述の挿入アルゴリズムから，同族の見出しのリスト上での位置の順序は，見出しの登録される順序関係に等しい．

ここで，N 個の見出しからなる任意の登録順序を

$$a_1a_2\cdots a_i\cdots a_N$$

と表し，これらの見出しを一様に分散したとき，同族の見出しの数が k 個からなるリストを対象に考察する．

はじめに，上述の条件のもとに，i 番目に登録された見出し a_i が，その同族の見出し列の先頭から数えて j 番目に位置する確率 S_{ikj} を導き出す．

a_i が k 個の同族の見出し列の先頭から j 番目に位置していることから，a_i より前方には，$j-1$ 個の見出しの配置場所があり，そこには登録順序が i 番目以前の $i-1$ 個の見出しが配置可能である．同様に，a_i より後方には，$k-j$ 個の配置場所があり，i 番目以後の $N-i$ 個の見出しが配置可能である．したがって，i 番目に登録された見出しが，k 個からなる同族の見出し列の先頭から数えて j 番目に配置される確率 S_{ikj} は次のようになる．

$$S_{ikj}=\binom{j-1}{j-1}\binom{N-i}{k-j}\Big/\sum_{j=1}^{k}\binom{i-1}{j-1}\binom{N-i}{k-j}$$
$$=\binom{i-1}{j-1}\binom{N-i}{k-j}\Big/\binom{N-1}{k-1} \qquad (2)$$

このとき，$\sum_{j=1}^{k} S_{ikj}=1$ が成立する．

たとえば，5個の見出し ($N=5$) を登録するとき，見出し列のひとつ ($a_1a_2a_3a_4a_5$) を理想的な分散関数によって，ある大きさの分散表に一様に分配したとき，同族の見出しの数が3個になったと仮定する．この前提のもとに，3番目に登録された見出し a_3 が同族の見出しの列の先頭から数えて1番目，2番目および3番目の位置に配置される場合の数とその確率 S_{ikj} について考察する．

まず，3個の見出しからなる同族の見出しの列は次のようになる．

$a_1a_2g_3$, $a_1a_2a_4$, $a_1a_2a_5$, $a_1g_3a_4$, $a_1g_3a_5$
$a_1a_4a_5$, $a_2g_3a_4$, $a_2g_3a_5$, $a_2a_4a_5$, $g_3a_4a_5$

上記の 10 通りの列において，3番目に登録された見出し a_3 を含んでいる列の数は6個である．これは (2)式の分母 $\binom{N-1}{k-1}$ における $N=5$, $k=3$ を与えた値6を示している．また，この値は他の見出し，a_1, a_2, a_4, a_5 についても同様であることを示している．

ここで，a_3 は6通りの列のうち，列の1番目の位置に1回，2番目の位置に4回，3番目の位置に1回，それぞれ位置している．

上記の結果は，提示した(2)から次のように求められる．すなわち，3番目に登録された見出しが3個の見出しからなる同族の見出し列の先頭から1番目，2番目および3番目の位置に配置される確率 S_{331}, S_{332} および S_{333} はそれぞれ次のようになる．

$$S_{331}=\binom{2}{0}\binom{2}{2}\Big/\binom{4}{2}=\frac{1}{6}$$

$$S_{332}=\binom{2}{1}\binom{2}{1}\Big/\binom{4}{2}=\frac{4}{6}$$

$$S_{333}=\binom{2}{2}\binom{2}{0}\Big/\binom{4}{2}=\frac{1}{6}$$

このとき，すべての見出しが任意の順序で登録されることを考慮すると，任意の見出しが k 個の見出しからなるリストの先頭から j 番目の位置に配置される確率は次のようになる．

$$\frac{1}{N}\sum_{i=1}^{N} S_{ikj}=\frac{1}{N}\sum_{i=1}^{N}\binom{i-1}{j-1}\binom{N-i}{k-j}\Big/\binom{N-1}{k-1}$$
$$=\frac{1}{N}\binom{N}{k}\Big/\binom{N-1}{k-1}=\frac{1}{k}$$

すなわち，任意の登録順番には，すべての見出しが等確率で出現するので，任意の見出しは k 個の見出しからなるリストのどの位置にも等しい確率 $1/k$ で配置される．

ところで，このモデルでは，同一の分散番地をもつ同族の見出しは，その分散番地を探索根とするひとつのリストにバケット単位 (バケットサイズ b) で連結される．したがって，任意の順序で登録される N 個の見出しがリストの先頭から k 番目のバケットに挿入される確率 Q_{Nk} は次のように算出される．

任意の順序で登録される見出しが，任意のリストの先頭から h 番目のバケットに格納されるためには，同族の見出しの数が少なくとも $(h-1)b+1$ 個より等しいか大きいこと，すなわち(1)式の確率 P_{Nk} の k は $(h-1)b+1$ より等しいか大きいことが必要である．

また，ある順番で登録される見出しが同族の見出し列で，先頭から数えて $(h-1)b+1$ から hb までの位置にあることが必要である．その結果，Q_{Nh} は次のように表される．

$$Q_{Nh}=\sum_{j=(h-1)b+1}^{hb}\sum_{k=j}^{N}\frac{1}{N}\sum_{i=1}^{N}S_{ikj}P_{Nk}$$
$$=\sum_{j=(h-1)b+1}^{hb}\sum_{k=j}^{N}\frac{1}{k}P_{Nk} \qquad (3)$$
$$(h=1,2,\cdots,\lambda)$$

ただし，$\lambda=\lceil N/b\rceil$ を示す．

($\lceil X\rceil$ は X 以上の最小の整数を表す)

たとえば，分散表の大きさ M が5，バケットサイズ b が2で，見出しの数 N が5という場合を仮定したとき，任意の順序で登録される見出しが，リストの先頭から2番目のバケットに配置される確率 Q_{52} について考察する．

まず，先頭から2番目のバケットに配置されるには，同族の見出しが3個以上からなるリストの先頭から数えて，見出しが3番目と4番目の位置に配置される場合を考えればよい．すなわち，同族の見出しの数が3個，4個および5個からなる各リストにおいて，見出しが各リストの先頭から3番目の位置に配置される場合と，同族の見出しの数が4個および5個からなる各リストの先頭からの4番目の位置に見出しが配置される場合とを考えることになる．したがって，Q_{52} は次のようになる．

$$Q_{52}=\sum_{j=3}^{4}\sum_{k=j}^{5}\frac{1}{k}P_{5k}$$
$$=\sum_{k=3}^{5}\frac{1}{k}P_{5k}+\sum_{k=4}^{5}\frac{1}{k}P_{5k}$$
$$=\frac{1}{3}P_{53}+\frac{1}{4}P_{54}+\frac{1}{5}P_{55}+\frac{1}{4}P_{54}+\frac{1}{5}P_{55}$$

このとき，同族の見出しの数が3個，4個，5個となる各リストの生成確率 P_{53}, P_{54}, P_{55} は，(1)式からそれぞれ $160/5^5$, $20/5^5$, $1/5^5$ となる．したがって，Q_{52} は次のような確率となる．

$$Q_{52}=\frac{956}{46875}$$

3.2 探索アルゴリズムの解析

前述の探索アルゴリズムに従い，見出しの探索はバケット単位でアクセスされるので，h 回のアクセスを要して探索されるバケットは，リストの先頭から h 番目のバケットになる．したがって，ある見出しが h 回のアクセスで探索される確率を R_{Nh} とすると，その確率は次のように導出できる．

前述の挿入アルゴリズムの解析から，任意の順序で登録される見出しが，同族の見出し列の先頭から h 番目のバケットに挿入される確率 Q_{Nh} は(3)式のように導出される．したがって，(3)式を導出した過程に各見出しの探索頻度 ρ_i を考慮することによって，確率 R_{Nh} は次のようになる．

$$R_{Nh}=\sum_{j=(h-1)b+1}^{hb}\sum_{k=j}^{N}\sum_{i=1}^{N}\rho_i S_{ikj}P_{Nk} \qquad (4)$$
$$(h=1,2,\cdots,\lambda)$$

ただし $\lambda=\lceil N/b\rceil$

ここで，成功探索時のアクセス回数の平均と分散の評価式を導出する．このとき考慮しなければならないことは，成功探索では，少なくともひとつ以上の見出しがリスト上に存在していることが必要である．

したがって，その条件を加味し，すなわち $\sum_{k=1}^{N}P_{Nk}$ を分母に導入して定式化すると次のようになる．

$$S_N=\sum_{h=1}^{\lambda}hR_{Nh}\Big/\sum_{k=1}^{N}P_{Nk}$$
$$=\sum_{h=1}^{\lambda}h\sum_{j=(h-1)b+1}^{hb}\sum_{k=j}^{N}\sum_{i=1}^{N}\rho_i S_{ikj}P_{Nk}\Big/\sum_{k=1}^{N}P_{Nk} \qquad (5)$$

$$V_N=\sum_{h=1}^{\lambda}h^2\sum_{j=(h-1)b+1}^{hb}\sum_{k=j}^{N}\sum_{i=1}^{N}\rho_i S_{ikj}P_{Nk}\Big/$$
$$\sum_{k=1}^{N}P_{Nk}-S_N^2 \qquad (6)$$

ここで，上記の一般式(5)(6)において，見出しの探索頻度を一様と仮定すると，

$\rho_i=1/N$ $(i=1,2,\cdots,N)$ から，$\sum_{i=1}^{N}\rho_i S_{ikj}$ は次のようになる．

$$\sum_{i=1}^{N}\rho_i S_{ikj}=\sum_{i=1}^{N}\frac{1}{N}\binom{i-1}{j-1}\binom{N-i}{k-j}\Big/\binom{N-1}{k-1}$$
$$=\frac{1}{N}\binom{N}{k}\Big/\binom{N-1}{k-1}$$
$$=\frac{1}{k}$$

したがって，アクセス回数の評価式は次のようになる．

$$S_N=\sum_{h=1}^{\lambda}h\sum_{j=(h-1)b+1}^{hb}\sum_{k=j}^{N}\frac{1}{k}P_{Nk}\Big/\sum_{k=1}^{N}P_{Nk}$$
$$=1+\sum_{h=1}^{\lambda-1}\sum_{k=hb+1}^{N}\frac{k-hb}{k}P_{Nk}\Big/\sum_{k=1}^{N}P_{Nk}$$
$$=1+\sum_{h=1}^{\lambda-1}\left\{\sum_{k=hb+1}^{N}P_{Nk}-hb\sum_{k=hb+1}^{N}P_{Nk}/k\right\}\Big/$$
$$\sum_{k=1}^{N}P_{Nk} \qquad (7)$$

ここで，確率 P_{Nk} をポアソン分布で近似して，上式を簡潔にする．このとき表占有率を $\alpha\,(=N/Mb)$ とする．

まず，$\sum_{k=hb+1}^{N} P_{Nk}$ をポアソン分布で近似する．

$h=1$ のとき

$$\begin{aligned}\sum_{k=b+1}^{N} P_{Nk} &\fallingdotseq \sum_{k=b+1}^{N} \frac{e^{-(N/M)}(N/M)^k}{k!}\\ &= \sum_{k=b+1}^{N} \frac{e^{-\alpha b}(\alpha b)^k}{k!}\\ &= e^{-\alpha b}\left\{\frac{(\alpha b)^{b+1}}{(b+1)!}+\frac{(\alpha b)^{b+2}}{(b+2)!}+\cdots\right\}\\ &= e^{-\alpha b}\frac{(\alpha b)^{b+1}}{b!}\left\{\frac{1}{b+1}\right.\\ &\quad +\frac{\alpha b}{(b+1)(b+2)}+\frac{(\alpha b)^2}{(b+1)(b+2)(b+3)}\\ &\quad \left.+\cdots\right\}\\ &= \frac{e^{-\alpha b}(\alpha b)^{b+1}}{bb!}\left\{\frac{b}{b+1}\right.\\ &\quad +\frac{\alpha b^2}{(b+1)(b+2)}+\frac{\alpha^2 b^3}{(b+1)(b+2)(b+3)}\\ &\quad \left.+\cdots\right\}\\ &= \frac{e^{-\alpha b}(\alpha b)^{b+1}}{bb!}R(\alpha, b)\end{aligned}$$

ただし，関数 $R(\alpha, b)$ は次の式で表される[2]．

$$R(\alpha,b)=\frac{b}{b+1}+\frac{\alpha b^2}{(b+1)(b+2)}+\frac{\alpha^2 b^3}{(b+1)(b+2)(b+3)}+\cdots$$

また，$R(\alpha, b)$ を漸近表示すると次のようになる[1]．

$$R(\alpha,b)\fallingdotseq\frac{1}{1-\alpha}-\frac{1}{(1-\alpha)^3 b}+O(b^{-2})$$

$h=2$ のとき

$$\sum_{k=2b+1}^{N} P_{Nk}\fallingdotseq\frac{e^{-\alpha b}(\alpha b)^{2b+1}}{2b(2b)!}R(\alpha/2, 2b)$$

$h=3$ のとき

$$\sum_{k=3b+1}^{N} P_{Nk}\fallingdotseq\frac{e^{-\alpha b}(\alpha b)^{3b+1}}{3b(3b)!}R(\alpha/3, 3b)$$

これから帰納的に次の関係が成立する．

$$\sum_{k=hb+1}^{N} P_{Nk}\fallingdotseq\frac{e^{-\alpha b}(\alpha b)^{hb+1}}{hb(hb)!}R(\alpha/h, hb) \qquad (8)$$

次に，$\sum_{k=hb+1}^{N} P_{Nk}/k$ をポアソン分布で近似する．

まず，$\sum_{k=hb+1}^{N}\frac{P_{Nk}}{k}\fallingdotseq\sum_{k=hb+1}^{N}\frac{P_{Nk}}{k+1}$ と近似して，

$h=1$ のとき

$$\begin{aligned}\sum_{k=b+1}^{N}\frac{P_{Nk}}{k+1} &\fallingdotseq \sum_{k=b+1}^{N}\frac{e^{-\alpha b}(\alpha b)^k}{(k+1)!}\\ &= e^{-\alpha b}\left\{\frac{(\alpha b)^{b+1}}{(b+2)!}+\frac{(\alpha b)^{b+2}}{(b+3)!}+\cdots\right\}\\ &= \frac{e^{-\alpha b}(\alpha b)^{b+1}}{b!}\left\{\frac{1}{(b+1)(b+2)}\right.\\ &\quad \left.+\frac{\alpha b}{(b+1)(b+2)(b+3)}+\cdots\right\}\\ &= \frac{e^{-\alpha b}(\alpha b)^{b+1}}{\alpha b^2 b!}\left\{\frac{\alpha b^2}{(b+1)(b+2)}\right.\\ &\quad \left.+\frac{\alpha^2 b^3}{(b+1)(b+2)(b+3)}+\cdots\right\}\\ &= \frac{e^{-\alpha b}(\alpha b)^{b+1}}{\alpha b^2 b!}\left\{\left(\frac{b}{b+1}\right.\right.\\ &\quad \left.+\frac{\alpha b^2}{(b+1)(b+2)}+\frac{\alpha^2 b^3}{(b+1)(b+2)(b+3)}\right.\\ &\quad \left.\left.+\cdots\right)-\frac{b}{b+1}\right\}\\ &= \frac{e^{-\alpha b}(\alpha b)^{b+1}}{\alpha b^2 b!\,(b+1)}\left\{(b+1)\left(\frac{b}{b+1}\right.\right.\\ &\quad \left.+\frac{\alpha b^2}{(b+1)(b+2)}+\frac{\alpha^2 b^3}{(b+1)(b+2)(b+3)}\right.\\ &\quad \left.\left.+\cdots\right)-b\right\}\\ &= \frac{e^{-\alpha b}(\alpha b)^{b+1}}{\alpha b^2 b!\,(b+1)}\{(b+1)R(\alpha, b)-b\}\end{aligned}$$

$h=2$ のとき

$$\sum_{k=2b+1}^{N}\frac{P_{Nk}}{k+1}\fallingdotseq\frac{e^{-\alpha b}(\alpha b)^{2b+1}}{2\alpha b^2(2b)!\,(2b+1)}*$$
$$*\ \{(2b+1)R(\alpha/2, 2b)-2b\}$$

$h=3$ のとき

$$\sum_{k=3b+1}^{N}\frac{P_{Nk}}{k+1}\fallingdotseq\frac{e^{-\alpha b}(\alpha b)^{3b+1}}{3\alpha b^2(3b)!\,(3b+1)}*$$
$$*\ \{(3b+1)R(\alpha/3, 3b)-3b\}$$

これから，帰納的に次の関係が成立する．

$$\sum_{k=hb+1}^{N}\frac{P_{Nk}}{k}\fallingdotseq\frac{e^{-\alpha b}(\alpha b)^{hb+1}}{h\alpha b^2(hb)!\,(hb+1)}*$$
$$*\ \{(hb+1)R(\alpha/h, hb)-hb\} \qquad (9)$$

また，

$$\sum_{k=1}^{N} P_{Nk}\fallingdotseq 1-e^{-\alpha b} \qquad (10)$$

で表される．

よって，(7)式を(8)，(9)，(10)で近似すると平均アクセス回数は次のようになる．

$$S_N \fallingdotseq 1+\sum_{h=1}^{\lambda-1}\frac{e^{-\alpha b}(\alpha b)^{hb+1}}{(hb+1)!\,(1-e^{-\alpha b})}\{R(\alpha/h, hb) * * (\alpha-h)(hb+1)/(h\alpha b)+h/\alpha\} \quad (11)$$

同様に，分散も次のようになる．

$$\begin{aligned} V_N &= \sum_{h=1}^{\lambda} h^2 \sum_{j=(h-1)b+1}^{hb}\sum_{k=j}^{N}\frac{P_{Nk}}{k}\Big/\sum_{k=1}^{N}P_{Nk}-S_N{}^2 \\ &= 1+\sum_{h=1}^{\lambda-1}\sum_{k=hb+1}^{N}((2h+1)(k-hb)/k)P_{Nk}\Big/ \sum_{k=1}^{N}P_{Nk}-S_N{}^2 \\ &\fallingdotseq \sum_{h=1}^{\lambda-1}\frac{(2h-1)e^{-\alpha b}(\alpha b)^{hb+1}}{(hb+1)!(1-e^{-\alpha b})}\{R(\alpha/h, hb) * * (\alpha-h)(hb+1)/(h\alpha b)+h/\alpha\} \\ &\quad + \left\{\sum_{h=1}^{\lambda-1}\frac{e^{-\alpha b}(\alpha b)^{hb+1}}{(hb+1)!\,(1-e^{-\alpha b})}\{R(\alpha/h, hb) * * (\alpha-h)(hb+1)/(h\alpha b)+h/\alpha\}\right\}^2 \end{aligned} \quad (12)$$

上述したように，探索アルゴリズムの解析では，一般化したモデルでの探索コストを厳密に評価することができる．特に見出しの探索頻度を一様と仮定すれば，探索コストの評価式は，(11)，(12)式のようにバケットサイズ b と表占有率 α によって簡潔に表現することができ，その導出過程は見通しよく系統的である．

ここで，探索頻度を一様と仮定したとき，平均アクセス回数の評価式(11)，(12)にもとづき，バケットサイズ b と表占有率 α をパラメータとし平均成功アクセス回数およびその分散の数値結果をそれぞれ**表 1**，**表 2** に示す．上述したようにバケットサイズ b，および表占有率 α によってアクセス回数を簡潔に評価することができる．

4. 従来の解析との比較検討

分離連鎖法についての従来の解析は，参考文献 2) に詳しく述べられている．しかし，その解析では，見出しの探索頻度を一様と仮定したもとで論議されている．

ここでは，従来の Knuth の解析[2]を詳細に考察し，提示する解析と比較検討する．

はじめに，バケットサイズ b が 1 の場合について考察する．

N 個の見出しを大きさ M の分散表に分散する，す

表 1 提示した解析に基づく平均成功アクセス回数
Table 1 Average accesses in a successful search based on the proposed analysis.

Bucket size (b)	Load factor (α)									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%
1	1.0352	1.0712	1.1080	1.1456	1.1840	1.2232	1.2632	1.3039	1.3453	1.3663
2	1.0033	1.0127	1.0275	1.0471	1.0709	1.0986	1.1296	1.1636	1.2001	1.2193
3	1.0004	1.0031	1.0097	1.0211	1.0378	1.0596	1.0864	1.1176	1.1527	1.1715
4	1.0001	1.0009	1.0040	1.0109	1.0230	1.0409	1.0648	1.0943	1.1287	1.1475
5	1.0000	1.0003	1.0018	1.0061	1.0151	1.0300	1.0516	1.0796	1.1135	1.1323
10	1.0000	1.0000	1.0001	1.0006	1.0028	1.0092	1.0225	1.0446	1.0758	1.0945
20	1.0000	1.0000	1.0000	1.0000	1.0002	1.0015	1.0068	1.0205	1.0466	1.0645
50	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0005	1.0045	1.0211	1.0365

表 2 提示した解析に基づく成功アクセス回数の分散
Table 2 Variance accesses in a successful search based on the proposed analysis.

Bucket size (b)	Load factor (α)									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%
1	0.0357	0.0732	0.1127	0.1540	0.1972	0.2423	0.2894	0.3383	0.3891	0.4152
2	0.0033	0.0126	0.0273	0.0467	0.0702	0.0973	0.1275	0.1606	0.1962	0.2148
3	0.0004	0.0031	0.0097	0.0208	0.0370	0.0578	0.0827	0.1112	0.1425	0.1591
4	0.0001	0.0009	0.0040	0.0109	0.0226	0.0397	0.0618	0.0882	0.1178	0.1336
5	0.0000	0.0002	0.0018	0.0061	0.0149	0.0292	0.0493	0.0744	0.1033	0.1188
10	0.0000	0.0000	0.0000	0.0005	0.0028	0.0091	0.0330	0.0426	0.0702	0.0858
20	0.0000	0.0000	0.0000	0.0000	0.0002	0.0015	0.0067	0.0201	0.0445	0.0603
50	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0045	0.0207	0.0352

なわち M 個のリストに分配するとき，各リストに分配された見出しの個数を，$k_1, k_2, \cdots, k_M$ とすると，その確率分布は $\dfrac{\binom{N}{k_1, k_2, \cdots, k_M}}{M^N}$ となる．ここで，N 個のすべての見出しを探索するときの探索路長を見出しの数 N で割った，すなわち見出しひとつ当たりの探索路長 $\dfrac{\binom{k_1+1}{2}+\binom{k_2+1}{2}+\cdots+\binom{k_M+1}{2}}{N}$ を確率変数として，成功時の平均探索路長 S_N を次のように定式化している．

$$
\begin{aligned}
S_N=&\sum_{k_1+k_2+\cdots+k_M=N}\left\{\frac{\binom{k_1+1}{2}+\binom{k_2+1}{2}+\cdots+\binom{k_M+1}{2}}{N}\right\}*\\
&*\frac{\binom{N}{k_1, k_2, \cdots, k_M}}{M^N}\\
=&\sum_{k=0}^{N} M\frac{\binom{k+1}{2}}{N}\frac{\binom{N}{k}(M-1)^{N-k}}{M^N}\\
=&\frac{M}{N}\sum_{k=1}^{N}\binom{k+1}{2}P_{Nk}\\
=&\frac{M}{N}\left\{P_{N1}+3P_{N2}+\cdots+\frac{N(N+1)}{2}P_{NN}\right\}\\
=&\frac{M}{N}\left\{\frac{1}{2}G''(1)+G'(1)\right\}\\
=&\frac{1}{2}\frac{N-1}{M}+1\\
\fallingdotseq&\frac{\alpha}{2}+1
\end{aligned}
\tag{13}
$$

ただし，$G(z)$ は確率 P_{Nk}，$(k=0, 1, 2, \cdots, N)$ の母関数で，$G(z)=\left(1+\dfrac{z-1}{M}\right)^N$ を表す．

また，同様にして探索路長の分散 V_N も次のように導出されている．

$$
\begin{aligned}
V_N=&\sum\left\{\frac{\binom{k_1}{2}+\binom{k_2}{2}+\cdots+\binom{k_M}{2}}{N}\right\}^2*\\
&*\frac{\binom{N}{k_1, k_2, \cdots, k_M}}{M^N}-\left\{\frac{N-1}{2M}\right\}^2\\
=&\frac{1}{M^N N^2}\left\{M(M-1)\sum\binom{N}{k_1, k_2, \cdots, k_M}\binom{k_1}{2}\binom{k_2}{2}\right.\\
&\left.+M\sum\binom{N}{k_1, k_2, \cdots, k_M}\binom{k_1}{2}^2\right\}-\left\{\frac{N-1}{2M}\right\}^2\\
=&\frac{1}{M^N N^2}\left\{M(M-1)\left(\frac{1}{4}M^{N-4}N^{\underline{4}}\right)+M^{N-3}\right.\\
&\left.\times\left(\frac{1}{4}N^{\underline{4}}+N^{\underline{3}}M+\frac{1}{2}N^{\underline{2}}M^2\right)\right\}-\left\{\frac{N-1}{2M}\right\}^2\\
=&\frac{(M-1)(N-1)}{2NM^2}
\end{aligned}
\tag{14}
$$

ここで，$N^{\underline{k}}$ は $\prod_{0\le j<k}(N-j)$ を表す．

前述したように，従来の解析の導出前提では，すべての見出しを分配し，その任意の分配に対して，すべての見出しを探索したときの探索路長の総和を見出しの総数で割った，すなわち見出しひとつ当たりの探索路長を確率変数として評価式を導出している．

一方，提示する解析では，探索回数それ自身を確率変数にとって評価式を定式化している．すなわち，(5)式において，見出しの探索頻度を一様と仮定し，$b=1$ とすれば，平均成功探索路長 S_N は次のようになる[4]．

$$
\begin{aligned}
S_N=&\sum_{h=1}^{N} h\sum_{k=h}^{N}\frac{1}{k}P_{Nk}\Big/\sum_{k=1}^{N}P_{Nk}\\
=&\frac{1}{2}\{G'(1)+G(1)-G(0)\}/\{1-G(0)\}\\
=&\frac{1}{2}\left\{\frac{N}{M}\Big/\left\{1-\left(1-\frac{1}{M}\right)^N\right\}+1\right\}\\
\fallingdotseq&\frac{1}{2}\left\{\frac{\alpha}{(1-e^{-\alpha})}+1\right\}
\end{aligned}
\tag{15}
$$

同様な導出前提から分散を導出すると次のような評価式になる．

$$
V_N\fallingdotseq\frac{\alpha(\alpha+1)}{3(1-e^{-\alpha})}-\frac{\alpha^2}{4(1-e^{-\alpha})^2}-\frac{1}{12}
\tag{16}
$$

したがって，従来の解析と提示する解析の基本的な相違点は，導出前提における確率変数の設定の違いにある．

従来の Knuth の解析から導出された評価式(13)，(14)は提示する解析から導出された評価式(15)，(16)に比べ，平均は若干大きめであるが，分散は著しく小さくなる．特に，分散の評価式(14)は $N\gg1$，$M\gg1$ となると $1/2M$ に近似され，分散表の大きさ M のみに依存する形となり，現実の現象を適切に評価していない．

次に，バケットサイズ b を考慮した場合の従来の解析は，次のようなモデルのもとで考察されている[2]．

同族の見出しは，バケットの先頭から順次登録し，バケットサイズを超えた見出しは，あふれ領域に見出し単位にひとつずつ格納し順次連結される．一方，探索するときには，基本的にはバケット単位でアクセス

表 3 Knuth の表現式を用いたときの平均アクセス回数
Table 3 Average accesses in a successful search by Knuth's formula.

Bucket size (b)	Load factor (α)									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%
1	1.0500	1.1000	1.1500	1.2000	1.2500	1.3000	1.350	1.450	1.450	1.5
2	1.0063	1.0242	1.0520	1.0883	1.1321	1.1823	1.238	1.299	1.364	1.4
3	1.0010	1.0071	1.0216	1.0458	1.0806	1.1259	1.181	1.246	1.319	1.4
4	1.0002	1.0023	1.0097	1.0257	1.0527	1.0922	1.145	1.211	1.290	1.3
5	1.0000	1.0008	1.0046	1.0151	1.0358	1.0699	1.119	1.186	1.286	1.3
10	1.0000	1.0000	1.0002	1.0015	1.0070	1.0226	1.056	1.115	1.206	1.3
20	1.0000	1.0000	1.0000	1.0000	1.0005	1.0038	1.018	1.059	1.150	1.2
50	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.001	1.015	1.083	1.2

されるが，あふれ領域の見出しに対しては，見出し単位にアクセスするモデルになっている．

上記のモデルのもとで，前述したバケットサイズがひとつの場合の解析と同じ導出前提に従い，成功時の平均探索路長すなわち平均アクセス回数の評価式を次のように導出している．

$$S_N = 1 + \frac{M}{N}\sum_{k>b}^{N}\binom{k-b+1}{2}P_{Nk}$$

$$\fallingdotseq 1 + \frac{1}{2}e^{-\alpha b}(\alpha b)^b b!^{-1}(\alpha b - b + 2 + (\alpha^2 b - 2\alpha(b-1) + b - 1)R(\alpha, b)) \qquad (17)$$

アクセス回数の分散についての評価式は明らかでないが，Knuth の解析の導出前提では，分散の評価式を導出することは極めて難しいと考える．

ここで，従来の Knuth の解析から導出された評価式(17)の数値例を**表 3** に示す[2]．

結局，提示する解析と従来の Knuth の解析との違いは，まず，見出しの探索頻度を考慮するかしないかによる導出前提の違い，すなわち確率変数の設定の違いにある．次に，バケットサイズを考慮するときのモデルの設定の違いである．

その結果，提示する解析では，成功探索路長すなわちアクセス回数の平均と分散の評価式は，表占有率とバケットサイズとによって簡潔に表現することができる．また，提示する解析方法は，見通しが良いうえに現実の現象を適切に評価することができる．

5. むすび

分離連鎖法を用いた見出し探索において，あふれ領域のバケットサイズをパラメータとし，かつ見出しの探索頻度を考慮に入れて一般化したモデルにおける挿入・探索アルゴリズムの厳密な解析を提示した．

まず，挿入アルゴリズムの解析では，任意の順序で登録される見出しが同族な見出しからなるリスト上のどのバケットにどのような確率で挿入されるかを明らかにした．次に探索アルゴリズムの解析では，バケットサイズと個々の見出しの探索頻度とを考慮に入れた探索コストすなわちアクセス回数を評価する表現式を導出した．特に探索頻度を一様と仮定すれば，探索コストの評価式はバケットサイズと表占有率のみの簡潔な表現となることを示した．導出した評価式は内部記憶の探索のみならず外部記憶上の探索に対して，いずれの場合にも現実の現象を厳密に評価できる．

謝辞 日頃から御助言，御指導を賜っている松山公一熊本大学名誉教授および九州大学工学部牛島和夫教授に感謝いたします．また，有益な御指摘をいただいた査読者の方に感謝いたします．

参考文献

1) Knuth, D. E.: *The Art of Computer Programming, Vol. 1, Fundamental Algorithms*, pp. 104-119, Addison-Wesley, Reading (1973).
2) Knuth, D. E.: *The Art of Computer Programming, Vol. 3, Sorting and Searching*, pp. 509-549, Addison-Wesley, Reading (1973).
3) 渋谷，山本：データ管理算法，岩波講座 情報科学 11, pp. 29-52 (1983).
4) 中村，松山：分散記憶法における探索頻度を考慮した探索路長とその評価，情報処理学会論文誌，Vol. 24, No. 1, pp. 125-130 (1983).
5) 中村，松山：見出しの探索頻度を考慮した探索路長の考察，情報処理学会論文誌，Vol. 24, No. 4, pp. 505-512 (1983).

(昭和 62 年 6 月 25 日受付)
(昭和 63 年 3 月 9 日採録)

中村　良三（正会員）

昭和15年生．昭和39年防衛大学応用物理専攻卒業．昭和43年熊本大学大学院電気工学専攻修士課程修了．中部電力(株)を経て，昭和50年から熊本大学工学部勤務．現在，同工学部電気情報工学科助教授．工学博士．算法解析，計算機言語，推論処理等に興味を持っている．電気学会，電子情報通信学会各会員．

税所　幹幸（正会員）

昭和25年生．昭和49年熊本大学工学部電気工学科卒業．昭和51年同大学院工学研究科修士課程修了(電子工学専攻)．同年(株)日本電気中央研究所入社．昭和57年八代工業高等専門学校情報電子工学科に勤務．現在，助教授．CAI，知識情報処理，データベースに興味を持つ．電子情報通信学会会員．