3J－1

ソフトウェア性能評価のためのランダムデータ高速生成法

二村良彦　大谷啓記　青木健一
早稲田大学 理工学部

1. はじめに

アルゴリズムの性能の精密な評価を行うためには，対象となるアルゴリズムの性能に影響を及ぼす特性を制御して生成されたランダムデータを用いる必要がある．また，性能評価に要する時間を短くするためには，ランダムデータを高速に生成する必要がある．特にアルゴリズムの教育において，学生に作成させたアルゴリズムを正当に評価するためには，合理的な評価用データが不可欠である．しかし，所望の特性を有するランダムデータの高速生成は，アルゴリズムの専門家にとっても容易ではない．例えば，長さ n の順列を等確率，即ち 1/n!で生成する O(n)アルゴリズムは知られている[7]．しかし，整列法等の精密な評価のためには，このような一様乱順列による評価では不十分である．何故ならば，アルゴリズムの性能に影響を与える要因はデータのサイズだけではなく，データが有する構造もアルゴリズムの性能を左右するからである．例えば，一様乱順列および一様乱数列は，極めて偏った構造を持つ．我々は先に各種の性質を有する乱順列を計算機オーバーフロー(またはアンダーフロー)無しで高速に生成する方法を提案した[3]．そこでは順列の含む葉数(数列において自分より小さい隣人を持たない要素の個数)を制御して順列を O(n)で生成する実用的近似方式についも報告した．そこで述べた精度の高いランダムデータ生成法は，高性能確率的アルゴリズム[5]の設計にも利用可能であった．そこでの成果に基づき，我々は，アルゴリズムの研究および教育に関る者が，各種のランダムデータをインターネット(http://www.futamura.info.waseda.ac.jp/index-j.html) を通じて容易に入手可能にするために，ランダムデータサーバーを開発した．本稿は，ランダムデータ生成法，ランダムデータサーバー(RDS)およびランダムデータサーバージェネレータ(RDGG)の概要について報告する．

Fast Random Data Generation for Software Evaluation,
Yoshihiko Futamura, Hirofusa Ohtani and Kenichi Aoki,
School of Science and Engineering, Waseda University

2. 一様乱数列の問題点

アルゴリズムの性能評価の際には一様乱数列(又は乱順列)が利用される場合が多い．しかし，それ等は非常に偏った性質を有する．例えばその葉数は殆どの場合約(n+1)/3 である[3]．またランズ(数列における上昇列の個数)の平均はその性質より明らかに約(n+1)/2 である．従って例えば一様乱数列を用いて整列法の評価を行うと QUICKSORT[6]のように葉数が多い場合に速く，少ない場合に遅い方法に有利になる．一方，MERGESORT[7]や LOAS[4]のように葉数が少ない場合に速く，多い場合に遅い方法にとって一様乱数列による評価は不利である[4]．実際に整列が行われる問題領域に適合した葉数を持った数列を用いて評価をしないと，実用的評価とは言い難い．実際に整列の対象となる大規模データの葉数を我々は計測していないが，それは[7]にある通り，整列済みに近いと予想される．整列済みに近いデータの葉数は 1 に近く，QUICKSORT が 1 番不得意とする領域である．従って，アルゴリズムの性能を精密に評価するためには，アルゴリズムが実際に扱うデータ領域が持つ特性を制御しながら乱数列を生成する必要がある．

特殊な特性を持つ乱順列を生成する試みはいくつか報告されているが[1,2]，それ等は一様性の保証をしていない．例えば[2]では，最長上昇部分列(LUSSと略記する)の長さを指定して乱順列を生成している．しかしそこでの方法は，複数のLUSSを持つ順列の方が，LUSSを1つしか持たない順列よりも出現確率が高い．例えば長さ6かつLUSS長を3とした場合，順列

4,1,5,2,6,3は，順列6,5,4,1,2,3の4倍以上の確率で生成される．また[1]では葉数を指定し乱順列を生成しているが，χ自乗検定での結果が思わしくない．また2分木を等確率で生成する方法も報告されているが，これは特性を制御した一般の順列の生成には直接応用できない[3]．

3．各種乱順列の生成法

我々は[3]において，長さn，単純指標mを持つ乱順列を0(nm)で生成する計算機オーバーフロー(またはアンダーフロー)無しの方法および，単純指標が葉数である場合には順列を0(n)で生成する実用的近似方式について報告した．単純指標とは，数列の葉数，ランズ，上昇部分数等に対応する特性のクラスである．そして形式的には順列がその生成規則に従って1つ短い順列から生成される際に，次の2性質を有する順列の特性のことである：(1)特性指標(順列から非負の整数上への関数及び関数値)が高々1しか増加しない．(2)新たな要素の挿入個所を順列の長さと特性指標に基づいて決められる．

単純指標を持つ順列の総数は下記の特性方程式により表すことが出来る[3]：

$S(n,m)=x(n,m)S(n-1,m)+y(n,m)S(n-1,m-1)$.

また，次の確率関数P(n,m)を分母と分子のオーバーフローを起こさずに計算することにより長さn単純指標mの乱順列を0(nm)時間およびスペースで生成することが出来る：$P(n,m)=x(n,m)S(n-1,m)/S(n,m)$．しかも特性指標が葉数の場合には確率関数Pの簡単な近似式を見付けることができた[3]．近似方式と従来の方法との比較を通じて葉数を指定して長さ100000以上もの大きな順列を生成する場合には，近似方式を利用せざるを得ないという結論を我々は得ている[3]．

4．ランダムデータサーバー

たとえ文献[3]を読んでも，その内容を理解し所望の特性を有するランダムデータを生成するプログラムを開発することは容易ではない．特にデータを高速に生成するためには，プログラミング上のテクニカルノウハウが必要である．また，ランダムデータの生成方式は特許出願されている(特願平8−250030, 1996年9月)ので，そのプログラムを自由に配布することには不都合がある．そこで我々は，ジェネレータ自身は配布しないことにした．その代わりランダムデータの必要な研究者や学生に対し，所望のデータを供給するランダムデータサーバーを開発した．それはこの後の講演3J-02と3J-03で報告する通り，指定された特性指標と長さを持つ乱順列を，指定された個数インターネットを通して供給することが可能である．

5．おわりに

アルゴリズムの研究と教育のために利用されることを期待して，ランダムデータの生成法と，それに基づくランダムデータサーバーを開発した．そしてそれをインターネット上で誰にでも利用できるようにした．RDSのようにテクニカルノウハウと知的所有権(特許)の凝縮されたソフトウェアを，それ自身を配布せず，その機能だけを社会にサービスする機会は今後ますます増大すると思われる．

6．参考文献

[1] 浅野：各種ソーティグアルゴリズムの実際的評価，情報処理学会アルゴリズム研究会30-7, 92年11月．

[2] Cook, C.R. and Kim, D. J.: Best sorting algorithm for nearly sorted lists, CACM, Vol.23, No.11, 1980, pp.620-624.

[3] 二村, 大谷, 青木, 二村：単純指標を持つ乱順列の高速生成法，情報処理学会アルゴリズム研究会，97年1月．

[4] 二村，二村，遠藤，平井:葉数最適整列法LOASとその実現法，情報処理学会アルゴリズム研究会44-2, 95年3月．

[5] R. Guputa et al: On Randomization in Sequential and Distributed Algorithms, ACM Computing Surveys, Vol.26, No.1, March 1994

[6] Hoare : Quicksort Compt. J., 1, 1, 1962, 10-15.

[7] Knuth, D. : The Art of Computer Programming, Vol.1-3, Addison-Wesley, 1973.