

TM-0013

Prologによる

ビットテーブルの扱い

石井 暁

August, 1983

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03) 456-3191~5
Telex ICOT J32964

Institute for New Generation Computer Technology

Prolog による

ビットテーブルの扱い

石井 暁

概要

Prologでプログラム作成, 実行し, 参照速度等を評価することによりビットテーブルの扱いの面から論理型言語を考察した. ここではビットテーブルの使用目的は, ある文字列が辞書にあるか否かの判定を想定した.

手法を変えた4通りのプログラムおよびFortranプログラムによる評価の結果, ユニファイと"arg"による要素のとり出しを行うPrologプログラムの性能はかなり良く, Fortranと比較して速度, 容量共 $\frac{1}{3}$ の性能を持っていることが解った.

目 次

| | |
|--------------------------|----|
| 1. はじめに | 1 |
| 2. プログラム | 3 |
| 2.1 Prolog (ユ=ファイ・ターム方式) | 3 |
| 2.2 Prolog (タームのターム方式) | 5 |
| 2.3 Prolog (ユ=ファイ・リスト方式) | 5 |
| 2.4 Prolog (リストのリスト方式) | 8 |
| 2.5 Fortran | 10 |
| 3. 評価 | 11 |
| 3.1 大きさの評価 | 11 |
| (1) Prolog (ユ=ファイ・ターム方式) | 11 |
| (2) Prolog (タームのターム方式) | 11 |
| (3) Prolog (ユ=ファイ・リスト方式) | 12 |
| (4) Prolog (リストのリスト方式) | 13 |
| (5) Fortran | 13 |
| 3.2 速度の評価 | 14 |
| (1) Prolog (ユ=ファイ・ターム方式) | 15 |

1. はじめに

ビットテーブルの扱いは、おけるデータ構造の表現能力、参照速度等の面から論理型言語を考察した。論理型言語として Prolog を用い、実証的な考察のプログラムの作成、実行により実証的な考察とした。

このビットテーブルとは、ビット列を定状態とし、各ビットが1か0かによって何らかの状態を表わすためのものである。例えば OS におけるディスクのアクセスに用いられるものである。

この評価実験においてビットテーブルの使用目的として自然言語においてある文字列が辞書にあるか否かの判定を想定する⁽³⁾。例えばエディタのスペルチェック⁽²⁾はウェブスター辞書の25000語をビットテーブルの50000ビットの1/10に対応させておき、ある文字列がウェブスター辞書にあるかを、50000ビットのうち最大10ビットの1/10を調べることに断言している。(エディタの方法は辞書にない文字列を辞書にあると判断することもある)

エディタの方法では、ビットテーブルの大きさは辞書の語数の10から数10倍必要であり、

数10万ビットの大きさが必要となる⁽¹⁾。また、ビットの参照速度が充分速いことが必要となる。例えば1行に10単語、10行に1単語のミススペルと例定し、1単語のチェックに10ビットを参照し、1ビットの参照に0.1m秒かかることを、対話型システムにおいて1つのミススペルの処理が終わってから次のミススペルがみつかるまでの時間は

$$0.1 \frac{\text{m秒}}{\text{参照}} * 10 \frac{\text{参照}}{\text{単語}} * 10 \frac{\text{単語}}{\text{行}} * 10 \frac{\text{行}}{\text{エラー}} = 0.1 \text{秒} / \text{エラー}$$

となり、この程度の速度は必要であろう*。
なお、ビットの書きかえの速度はそれ程要求
されないので、本評価実験において考えていな
い。

使用した計算機はICOTのdec 2060であ
る。PrologはProlog-20 version 1.0であり、
全ての述語にmode指定を付けてある。

* 各単語について参照される全ビットがばらばらにはなっていると仮定している。

2. プログラム

次の5通りのプログラムにより評価した。

- Prolog (ユニファイ・ターム方式)
- Prolog (タームのターム方式)
- Prolog (ユニファイ・リスト方式)
- Prolog (リストのリスト方式)
- Fortran

以下に各々の説明をする。

2.1 Prolog (ユニファイ・ターム方式)

ビットテーブルを表現する際、第1図に示す様に複数のアサーションに分解し、各アサーションにおいてはコンパウンドタームとして表現するものである。ここで n は整数であり、最大18ビット使えるが、本実験では右側10ビットのみ用いている。また、 $m = 50$, $n = 300$ とした。

n ビット目のビットの値を知るアルゴリズムは自明であるが、次のステップで行われる。

- (i) どのアサーションの中のタームにそのビットがあるかを計算し、ユニファイによりそのタームを取り出す。
- (ii) そのタームの何番目の要素に求めるビットがあるかを計算し、組み込み手続き"arg"によりその要素を取り出す。
- (iii) その要素の何番目のビットが求めるものかを計算し、シフトによりそれを取り出す。

プログラムの大きさは約20行である。

$\text{bit}(1, b(a_{11}, a_{12}, \dots, a_{1m-1}, a_{1m}))$.

$\text{bit}(2, b(a_{21}, a_{22}, \dots, a_{2m-1}, a_{2m}))$.

⋮

$\text{bit}(n, b(a_{n1}, a_{n2}, \dots, a_{nm-1}, a_{nm}))$.

※ 1 ㉔ $\Gamma = \text{ライ・ター・ン}$ 式のビットテーブル

(ビットテーブルそのものおよびコメント行を除く)

2.2 Prolog (タームのターム方式)

ビットテーブルを表現する際、第2図に示す様に1つのアサーション中のタームのタームとして表すものである。その他は前のものと同じである。ただし、 $n=100$ とした。

又ビット目のビットの値を知るアルゴリズムは前のユニファイ・ターム方式と同様であるが、(i)が異り、

- (i) ターム中の何番目のタームにそのビットがあるかを計算し、"arg"によりそのタームを取り出す。

となる。

プログラムの大きさはユニファイ・ターム方式と同様である。

この方式では、ビットテーブルが大きくなると、エディタで参照している時、その何行目を扱っているかが解りにくくなる。第1図の方式においては、述語bitの第1アーギュメントにより、かなりその点で解り易くなる。大きなプログラム作成時には考えておくべき特徴であろう。

2.3 Prolog (ユニファイ・リスト方式)

ビットテーブルを表現する際、第3図に示す様に複数のアサーションに分解し、各アサーションにおいてはリストとして表現するものであ

$$\text{bit} \left(\text{bit} \left(b(a_{11}, a_{12}, \dots, a_{1m-1}, a_{1m}), \right. \right. \\
\left. \left. b(a_{21}, a_{22}, \dots, a_{2m-1}, a_{2m}), \right. \right. \\
\vdots \\
\left. \left. b(a_{n1}, a_{n2}, \dots, a_{nm-1}, a_{nm}) \right) \right).$$

第2図 ターミナルのビット列

bit (1, [a₁₁, a₁₂, ..., a_{1 m-1}, a_{1 m}]).

bit (2, [a₂₁, a₂₂, ..., a_{2 m-1}, a_{2 m}]).

⋮

bit (n, [a_{n1}, a_{n2}, ..., a_{n m-1}, a_{n m}]).

第3回 ヲニファイ・リスト方式のビットテーブル

る。ユニファイ・ターム方式と比較するとコンパウンドタームがリストに変わったものであり、その他は同様にしてある。

スビット目のビットの値を知るアルゴリズムはユニファイ・ターム方式と同様であるが、次の様である。

- (i) どのアサーションの中のリストにそのビットがあるかを計算し、ユニファイによりそのリストを取り出す。
- (ii) そのリストの何番目の要素に求めるビットがあるかを計算し、その要素を取り出す。
- (iii) その要素の何番目のビットが求めるものを計算し、シフトによりそれを取り出す。

プログラムの大きさは、リストの要素を取り出す手続き(2~3行)分だけユニファイ・ターム方式より多い。

2.4 Prolog (リストのリスト方式)

ビットテーブルを表現する際、第4図に示す様に1つのアサーションのリストのリストを用いる。その他はユニファイによるものと同様である。ただし、 $n=100$ とした。

スビット目の値を知るアルゴリズムは、ユニファイ・リスト方式のプログラムと比較して(i)のみが異り、

- (i) リスト中の何番目のリストにそのビットがあるかを計算し、そのリストを取り出す。

list ([[a₁₁, a₁₂, ----- a_{1m-1}, a_{1m}],
[a₂₁, a₂₂, ----- a_{2m-1}, a_{2m}],
⋮
[a_{n1}, a_{n2}, ----- a_{nm-1}, a_{nm}]]).

第4回 リストのリスト形式のビットテーブル

となる。

プログラムの大きさについてはユニファイ・リスト方式のプログラムと同様である。

2.5 Fortran

ビットテーブルとして整数の1次元配列を用いる。Fortranでは整数は36ビットであるが、符号ビットを除いた35ビットを用いた。本実験では配列の大きさを10,000 (350,000ビット)とした。

又ビット目の値を知るアルゴリズムは自明である。ただし、整数に対してビットのシフトを行う関数、サブルーチンが用意されていない様なので、それを2の何乗かの数との演算により陽に書いた。

プログラムサイズは20行余りで、Prologの場合と大差はない。(付録のプログラには使われていない部分や、ビットに値をセットするサブルーチンも含まれており、もっと長い)

3. 評価

評価は定義し得るビットテーブルの大きさと速度について行われた。

3.1 大きさの評価

(1) Prolog (ユニファイ・タ-ム方式)

一応本実験では 15,000 整数で 150,000 ビット (フルに使用は 270,000 ビット) で行った。しかし、本実験で直接の限度となったのは、Prolog における整数の値の上限である。即ちビットテーブルの "何番目" のビットかを表わすのに整数を用いたが、それが 18 ビットなので約 130,000 の数までしか表現できないことがネックになった。(それ以上の数を扱う拡張もされているが、それが全ての場合に思い通りには動作をしない様である)

ビットテーブル自体はコンパイラを通すとき

90,000 整数で 900,000 ビットは可

95,000 整数で 950,000 ビットは不可。

また、コンパイラを通さなければ

95,000 整数で 950,000 ビットは可

である。

(2) Prolog (タ-ムのタ-ム方式)

本方式においては上の数字よりかなり減り、

コンパイラを通すとき

7,500 整数で 75,000 ビットは可
10,000 整数で 100,000 ビットは不可

であり、コンパイラを通さないとき

5,000 整数で 50,000 ビットは可
7,500 整数で 75,000 ビットは不可

であった。コンパイラを通す方が多くのビットを定義できるのは、これのみであった。なお、この方式の速度の評価は 50,000 ビットを定義して行った。

この方式や後に述べるリストのリスト方式の様な1つのアサーションでビットテーブルを表わす方式において、上記の制約は1つのアサーションについての制約であり、アサーションを複数にすることにより、更に大きなビットテーブルを定義することができる。

(3) Prolog (ユニファイ・リスト方式)

ユニファイ・ターム方式と同様な条件で行った。直接の上限が Prolog の整数が 18 ビットであることで決まっている点も同様である。

ビットテーブル自体はコンパイルする時、しない時共、

45,000 整数で 450,000 ビットは可
50,000 整数で 500,000 ビットは不可

であった。

(4) Prolog (リストのリスト方式)

コンパイラを通す場合

5,000 整数で 50,000 ビットは可
6,000 整数で 60,000 ビットは不可

であった。整数の 18 ビットをフルに使えば、約 100,000 ビットまで扱える。なお、コンパイラを通さなければ少くとも

7,000 整数で 70,000 ビットは可

であった。

速度の評価の実験は 50,000 ビットを定義して行った。

(5) Fortran

本実験は 10,000 整数で 350,000 ビットを扱ったが、更に大きなテーブルを扱えて

130,000 整数で 4,550,000 ビットは可
140,000 整数で 4,900,000 ビットは不可

であった。

3.2 速度の評価

基本的には速度は1ビットを参照するのに必要な時間で評価される。しかし、リストを用いた場合には、参照したいビットがリストのどの位置にあるかによって必要な時間が変化すると思われる。

そこで、本実験においては、ビットテーブルの大きさ(ビット数 S で表す)を変化させ、各々の場合についてビットテーブルの全ビットを順に1ビットずつ参照するのに必要なCPU時間を測定することとした。*

これは、Prologにおいては"statistics"によって得られる。ただし、時間がXWD表示になる(129.071秒以上になる)場合には、表示に不明な点があるため、測定を行わなかった。

Fortranにおいて、このための適切なサブルーチンが見当らなかつたので、RUNコマンドを実行するのにかかった時間を用いた。

これは測定したい時間の他に多くの時間を含むが、それは測定値の解析の時に考慮した。

全てのケースにおいて測定値は3回同条件で測定した値の平均である。また、Prologによる場合、コンパイルしての実行を中心に扱う。マシンの混み方などによる測定値のバラツキは、100ms以上の値においては3~4%程度におさまっている様である。

* 実験は、ビットテーブルの大きさは固定し、その一部の全ビット参照を行った。これによる誤差は特に問題にならない。また、この状況は実際の応用システムでの使用状況とは異なるかも知れない。注意が必要である。

(1) Prolog (ユニファイ・ターム方式)

大きさ S ビットのビットテーブルの全ビットの参照にかかる CPU 時間 T を推定する。
ビット列の任意のビットを参照するには

- ・ アサーションのユニファイ
- ・ arg による要素のとり出し

の二つの処理が主に問題になる。前者は処理系の工夫次第であるが、例えばハッシングなどによってどのアサーションがユニファイされる場合でも一定の時間ですむと考えられる。
後者についても、同様にどの要素がとり出される場合でも一定の時間ですむと考えられる。
結局、

T は S に比例

と考えられる。

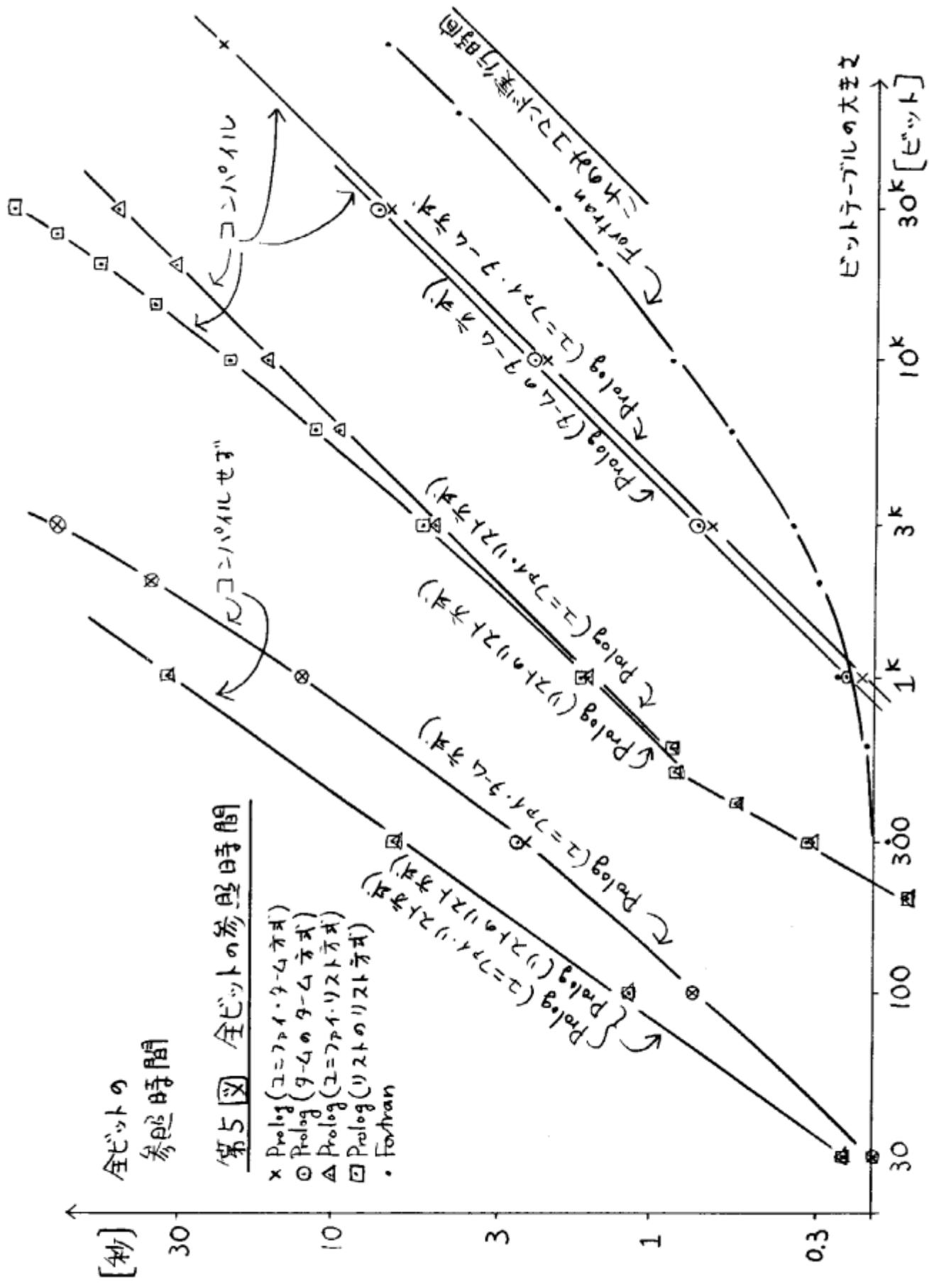
実測値によるグラフを第5図に示す。コンパイルした時は上記の議論が成り立ち、

T は $S^{1.0}$ に比例

となっている。

一方、コンパイルをしない場合、 S が 100 以下付近では上記議論の通りであるが、 S がそれ以上になると、それ以上の時間がかかる様になる。次に述べるユニファイ・リスト方式ではこの様な傾向が見られないので、これは " arg " の処理方式がコンパイルする場合の様に工夫されていないためと考えられる。

上記をまとめると、コンパイルによるこの方式では 1 ビットの参照には



約

0.22 m 秒

かかる。また、コンパイルの効果は S により一定ではないが、第5図のグラフの範囲内 ($30 < S < 3,000$ 程度で) は

大体 28 ~ 110 倍の効果

である。また、コンパイルの時間は約30秒で、コンサルトする場合の7割増程度である。

(2) Prolog (タ-ムのタ-ム方式)

前のユニファイ・タ-ム方式と同様の結果であり、それとの相違を中心に述べる。

ビット列の任意のビットの参照には

- arg による要素の取り出し

が2回行われる。これはどの要素についても一定と考えられるので、大きさ S ビットのビットテーブルの全ビットの参照にかかる CPU 時間 T について

T は S に比例

となる。

第5図の実測値によると、コンパイルをした時は

T は $S^{1.0}$ に比例

となっている。しかし、前のユニファイ・タ-

ム方式の場合と比較すると約1割多く時間がかかっており、1ビット参照に約

0.24 m 秒

必要である。

コンパイルしない場合は、前のユニファイ・タム方式と全く同じ結果になっており、特に付け加えて説明する点はない。

コンパイルの効果は第5図のグラフの範囲内 ($30 < S < 3,000$ 程度で) は

大体 25 ~ 100 倍の効果

である。また、コンパイル時間は約10秒で、コンサルトする場合の3割増程度である。

(3) Prolog (ユニファイ・リスト方式)

大きさ S ビットのビットテーブルの全ビットの参照にかかる CPU 時間 T は、次の様になると考えられる。

(i) $0 < S \leq 500$ のとき

ビット列を表現している複数のアサーションのうち、1つのみで処理が終る。そのためアサーションをユニファイする時間は特に問題にならない。一方、求めるビットが、そのアサーションの中のリストのどの値にあるかの影響は、リストの先頭からの位置に比例した時間がかかると考えられる。そこで、

T は S^2 に比例

となる。

(ii) $S \gg 500$ のとき

(i)とは逆にリストの中からビットをさがす時間は平均すれば一定と考えてよい。一方ユニファイするための時間は、処理系の工夫の結果、どのアサーションがユニファイされる場合でも一定の時間であると考えられる。そこで

T は S に比例

となる。

実測値によるグラフを第5図に示す。コンピュータした場合には、ほぼ上記の議論が成り立っている。即ち

(i) $0 << S \leq 500$ のとき

グラフから

T は $S^{1.8}$ に比例

となる。べき数が2まで達しないのは S に比例の項などが影響しているためと思われる。

(ii) $S \gg 500$ のとき

グラフから

T は $S^{1.0}$ に比例

となる。

(iii) $500 < S < 1000$ 付近

(i)と(ii)の部分がなめらかにはつながらないことが読みとれる。

一方コンパイルをしない場合。

T は $S^{1.4}$ に比例

となっている。これは $300 < S < 1,000$ 程度のち
ようと(i)と(ii)の中間で、上記議論が成り立たな
い範囲のためと考えられる。

上記をまとめると、コンパイルした場合数万
ビットのビットテーブルにおいて、1ビットの
参照にほぼ一定の時間

1.6 m 秒

かかることが解る。

コンパイルの効果は S が30から1000程度の範
囲で

大体20 ~ 24倍

である。なお、コンパイル時間は1分余りであ
り、コンサルトするのに比較して約3倍の時間
がかかる。

このプログラムで、リストAのB番目の要素
をCに取り出す述語を

`select (A, B, C)`

で呼ぶ様にしてある。これを

`select (B, A, C)`

とアーギュメントの順を変えることにより、コ
ンパイルする時、3割 ~ 4割の速度の向上があ
る。処理系のやり方を知ってプログラムを書く
か、書かないかにより、大きな差が出ることに
解る。なお、コンパイルしない場合の速度の向
上は、無視し得る(値のバラツキの)範囲内であ
る。

(4) Prolog (リストのリスト方式)

(i) $0 \ll S < 500$ のとき

(3) に述べたユニファイ方式と同様の議論が成り立ち

T は S^2 に比例

と考えられる。

(ii) $S \gg 500$ のとき

小さいリストからビットをさがす時間は平均すれば一定と考えてよい。しかし、該当する小さいリストを得るためには、リストの先頭からの位置に比例した時間がかかると考えられるので、(i)と同様に

T は S^2 に比例

と考えられる。

第5図の実測値によると

(i) $0 \ll S < 500$ のとき

グラフでは、(3)のユニファイ・リスト方式と重なり、

T は $S^{1.8}$ に比例

となる。

(ii) $S \gg 500$ のとき

かなり予想と異った結果である。これは、実は

$T = S^2$ に比例する頃 (小さいリスト

を 選ぶ操作)
+ S に 比例 する 項 (小 さ な リ ス ト
か ら ビ ッ ト を 選 ぶ 操 作 な ど)
で あり、 図 で は、 両 方 の 項 を 考 え な け れ ば
な ら ない 範 囲 を 表 わ し て い る た め と 考 え ら
れ る。 S に 比 例 する 項 は ユ ニ フ ァ イ ・ リ ス ト
方 式 で の 時 間 T と ほ ぼ 等 し い と 考 え ら れ
る の で、 そ の 差 を と る と
差 は $S^{2.0}$ に 比 例
と な っ た。

図 の S が 10 K ビ ッ ト の 点 で も、 小 さ な リ
ス ト に は 50 の 要 素 が あ る の に、 大 き な リ ス ト
に は 20 の 要 素 し か な い た め、 $S > 500$ で
は あ る が、 $S \gg 500$ で は な く、 上 記 の 様 に
な る と 考 え ら れ る。

上 記 か ら、 こ の 方 式 で は 数 万 ビ ッ ト の ビ ッ ト
テ ー ブ ル に お い て、 1 ビ ッ ト の 参 照 に は 約

2.2 m 秒

か か る こ と が 解 る。

コ ン パ イ ル の 効 果 は S が 30 か ら 1000 の 範 囲 で

大 体 20 ~ 24 倍 の 効 果

で あり、 ユ ニ フ ァ イ ・ リ ス ト 方 式 と 同 様 で あ る。
な お、 コ ン パ イ ル 時 間 は 約 1 分 で あり、 コ ン サ
ル ト する の に 比 較 し て 約 8 倍 の 時 間 が か か る。

余 談 に な る が、 こ の プ ロ グ ラ ム で、 リ ス ト A
の B 番 目 の 要 素 を C に 取 り 出 す 述 語 $select(A, B,$
 $C)$ を

$select([C|-], 1, C):-!$

$select([-|D], B, C):-BM \text{ is } B-1, !,$
 $select(D, BM, C).$

としてある。この2番目の文を

```
select (A, B, C) :- second (A, D), BM is B-1,  
!, select (D, BM, C).
```

```
second ([_ | D], D).
```

とすると、実行時間が2~3割多くかかる。何
度も呼ばれる述語ではあるが、予想以上の影響
が出てくる。

また、selectのアーギュメントの順による影
響は前のユニファイ・リスト方式の場合と同様
である。(もっと大きなビットテーブルではよ
り大きな影響があると思われる。

(5) Fortran

ビットテーブルをリストではなく配列にして
いること、およびCPU時間は参照の時間のみ
ではなく、runコマンド全体の実行時間を測定
したため、

T は S に比例 + ゲタ

と考えられる。第5図に示す様に実測値は、そ
の様になっている。

S が小さい時の T の値からゲタを推定するこ
とにより

$$T = 0.067 S + 0.20 \text{ [m秒]}$$

となる。

4. おわりに

Prolog によりビットテーブルを参照するプログラムを作成し、動作させて評価し、論理型言語を考察した。評価にあたり、4種のデータ構造のPrologプログラムおよびFortranによる同機能のプログラムの比較によった。

その結果、Prologでも、ユニフィケーションと"arg"による要素の取り出しによる方法ではかなり効率がよく、コンパイルすれば

- 速度はFortranの3倍程度遅く。
- ビットテーブルは一応百数十万ビット (Fortranの場合の約 $1/3$)程度まで定義できる。ただし、ビットの"何番目"を整数で表すと、それが18ビットであることが直接のネックとなり、工夫なしでは13万ビット (Fortranの場合の数十分の1)程度までとなる。
なお、大きなビットテーブルを定義する際、コンサルトされた語数としてマイナスの数が表示されたりする。正しく動作しているかは、更に確認する必要があるだろう。

リスト構造はPrologにむいたデータ構造であると言われている。しかし、本文で述べた様なかなりの量のデータに対してランダムなアクセスをする応用については速度、記憶容量の点から、なるべくリストを避けなければならないということが定量的に明らかになった。この種の"論理"に関係ない処理も容易に記述できる様な改良(従来からの言語機能を取り込むこと)を希望したい。本実験では扱えなかったビットテーブルへの書き込みにおいては、その種の改良が一層望まれるであろう。

文献

- [1] Carter, L. et al, "Exact and Approximate Membership Testers", Proc. of 1978 Tenth Annual Symposium on Theory of Computing, ACM, pp. 59~65 (1978)
- [2] Wood, S., "Z - the Last Word in a Series of Single Letter Yale Editors", Yale Univ., pp. 34~35 (1980)
- [3] 川合, "英文綴り検査法", 情報処理 vol. 24, no. 4, pp 507~513 (1983)

付録5. Fortran ヲ-リスト

```
dimension ibit(10000)
c do 10 i= 1,10000
c10 ibit(i)=0
c
c
c go to 1
c
c
c do 20 i = 1,10
write(5,99)
99 format(1h ,3hset)
read(5,100) j
100 format(i6)
20 call setbit(ibit,j)
1 write(5,199)
199 format(17h check(<= 350000))
c read(5,100)i
c
c read(5,100) imax
call time(x,y)
write(5,900) x,y
do 1000 ii=1,imax
c x=ran(0) * 350000.0
c i=x
c i=ii
if(i .le. 0) i=1
if(i .ge. 350000) i=350000
call refbit(ibit,i,j)
c write(5,200) i
c write(5,200) j
200 format(1h ,i10)
1000 continue
c go to 1
call time(x,y)
write(5,900) x,y
900 format(1h ,2a)
stop
end
subroutine setbit(ibit,i)
dimension ibit(10000)
k=(i+34)/35
l=ibit(k)
kk=i-(k-1)*35
m=0
if (kk .ne. 1) m=(1/(2**(35-kk+1))) * (2**(35-kk+1))
n=mod(l,(2**(35-kk)))
ibit(k)=m+n+2**(35-kk)
c write(5,100) i,k,l,kk,m,n,ibit(k)
100 format(1h ,2i8,o,i,3o)
return
end
subroutine refbit(ibit,i,j)
dimension ibit(10000)
k=(i+34)/35
l=ibit(k)
k=i-(k-1)*35
j=1/(2**(35-k))
j=mod(j,2)
return
end
```