順位符号に基づく英文二次文献情報のデータ圧縮法†

二村祥一†† 松尾文碩††

本稿では，英文二次文献情報に対する実用的データ圧縮プログラムの実現とその性能について述べる．このプログラムの圧縮技法は，QOC と呼ぶテキスト圧縮符号を核にしたものである．QOC は，基本的には順位符号と呼ぶ単語を圧縮単位とする可変長符号であり，最適符号ではないが最適に近く，また符号器と復号器が高速であるため，きわめて実用性が高い．このデータ圧縮プログラムを1年分の INSPEC テープを使って圧縮したところ，圧縮率は分野によって多少異なるが，平均 2.20 ビット/字～2.42 ビット/字と良好な結果を得た．また，FACOM M-382 計算機上での平均符号化時間および復号化時間は，分野によらず，それぞれ 3.90 マイクロ秒/字，1.09 マイクロ秒/字と非常に高速であった．このデータ圧縮プログラムは，九州大学大型計算機センターにおいて会話型情報検索システム AIR による INSPEC テープの検索サービスに使用されている．

1. まえがき

オンライン文献検索は，大量のディスク領域を必要とするため，データ圧縮技法を最も必要とする分野の一つであろう．文献データは，抄録や標題など文章情報の占める割合が大きいので，このデータ圧縮にはテキスト圧縮技法の影響が大きい．これまで種々のテキスト圧縮技法が考案され，実用化されている[1]．テキストは，通常 EBCDIC のような1文字当り8ビットの符号によって表現されているので，かなり冗長であり，文字や単語をその生起頻度に応じて可変長符号で表すことにより，冗長度を減らすことができる．この方法では，ハフマン符号[2]が最適であることが証明されている[3]が，文字を圧縮単位としたハフマン符号による英文の圧縮では，1/2 倍程度の圧縮効果しかないことが知られており，さらに圧縮率を上げるには，単語や N-gram などの文字列を単位とした圧縮方法を取らざるをえない．しかし，ハフマン符号は，単語を圧縮単位とした場合のように要素数が非常に大きい場合は処理がやや面倒で，符号化および復号化に要する計算量が大きくなる．

筆者らは，QOC (Quasi-Optimum Code) と呼ぶテキスト圧縮技法を開発した．QOC は，高頻度単語に対しては順位符号 (Rank Code) と呼ぶ単語を圧縮要素とする圧縮技法を用い，低頻度単語に対しては文字を要素とする圧縮を行う．順位符号は，ハフマン符号のように最適ではないが，ジップの法則[4]が成立する言語に対しては，その圧縮率はハフマン符号より 2.7% 悪いだけである[5]．一方，符号化と復号化に要する計算量において，順位符号はハフマン符号に決定的に勝っている[5]．QOC により INSPEC テープ[6],[7]の圧縮を試みたところ，抄録や標題のような文章情報のみならず，誌名，著者の所属機関名などの書誌情報に対しても，この技法が有効であることを確認することができた．すなわち，これらの項目については約 1/4 にデータを圧縮することができた．また，FACOM M-382 計算機上で実現した符号器および復号器の速度は，1文字当りそれぞれ 4.01 マイクロ秒，1.03 マイクロ秒と非常に高速であった．

ここでは，QOC を中心とした英文二次文献データの符号化法について述べ，この技法によって INSPEC テープを圧縮した結果について報告する．

2. 英文二次文献情報

本稿では，二次文献情報とは抄録誌に記載されている情報をいう．二次文献情報には，このほか索引誌や文献目録などがあるが，本稿はこれらの情報を対象とはしない．しかし，本稿の技法は，索引誌や文献目録，雑誌目録などの情報に対しても有効であろう．

さて，ここでは機械可読型の二次文献情報を二次文献データと呼ぶことにする．現在，ほとんどの英文抄録誌は冊子体だけではなく，二次文献データの形で流通しており，それらのデータ量はきわめて大きい．また，二次文献情報の利用者も検索能率の点から冊子体よりもオンライン文献検索システムの利用に移りつつある．しかし，オンライン文献検索に要する費用が比較的高いことがこの移行を阻害する最大の要因となっている．検索コストが高い原因の一つは，オンライン文献検索が大量の計算機資源，とりわけ二次記憶媒体を必要とするためである．したがって，英文二次文献

† A Data Compression Method Based on Rank Code for Secondary Document Information Written in English Language by SHOUICHI FUTAMURA and FUMIHIRO MATSUO (Computer Center, Kyushu University).

†† 九州大学大型計算機センター

データに対する効率的なデータ圧縮技法は，オンライン文献検索に寄与するところが大きい．

本稿では，データ圧縮の対象となる英文二次文献データとして INSPEC テープを取り上げたが，本稿の技法は他の英文二次文献データについてもINSPECテープと同程度に有効であると考えられる．INSPECテープは，英国 IEE (the Institution of Electrical Engineers) が1969年から提供している代表的な英文二次文献データで，物理学や電気工学，電子工学，制御工学，計算機科学，情報工学の分野の文献を含んでいる．収録文献数は，1985年11月の時点で約255万である．

図1に INSPEC テープのレコードの例を示す．これは，雑誌からの文献の例で10項目からなる．この項目は，標題 (title), 著者名 (author), 著者の所属機関名 (affiliation), 誌名 (journal title), 巻数 (vol/issue no's), ページ番号 (page numbers), 発行年月日 (publishing date), CODEN, 抄録 (abstract), 自由索引句 (free-indexing terms) である．

会議録からの文献には，上記の項目に会議名 (conference title), 会議開催場所 (location of conference) など会議に関する情報が追加される．本の場合は，出版社名 (publisher), 出版社所在地 (place of publication), ページ数 (no. of page) などの情報が追加されるが，誌名，巻数，CODEN などの項目がない．

INSPEC テープにおける項目は，二つに大別することができる．一つは，標題や抄録，自由索引句の文章情報であり，もう一つは著者名，誌名などのような定型データである．データ量は前者が圧倒的に多い．例えば，1983年の INSPEC テープにおける各項目のデータ量の割合は，抄録，自由索引句，標題，著者の所属機関名，誌名，著者名が，それぞれ 59.7%, 17.3%, 7.2%, 4.4%, 3.5%, 3.0% で，それら以外の項目については，各項目は 1% 以下であり，合計では 4.9% であった．このように，文章情報のデータ量は全体の 84% を占める．したがって，二次文献情報のデータ圧縮は文章情報に対するテキスト圧縮技法を中心にしたものでなければならない．

3. 符号化法

本稿の英文二次文献情報に対するデータ圧縮符号は，テキスト圧縮符号である QOC に，固定長符号と文字単位符号を組み合わせたものである．QOC は，文章情報だけではなく，幾つかの定形データ型書誌的事項に対しても効果的である．

3.1 QOC

QOC は，基本的には単語を圧縮要素とし，その生起頻度の順位 (rank) を可変長2進数によって表現した符号である．順位 r の単語とは，r 番目に出現頻度が高い単語のことである．いま，順位 r の単語 w の符号を ω とすると，ω の本体はビット長 $\lfloor \log_2 r \rfloor$ の2進数 $r-2^{\lfloor \log_2 r \rfloor}$ である．これをマトリックスと呼

Title	Advanced feedback methods in information retrieval
Authors	Salton, G., Fox, E.A., Voorhees, E.
Affiliation	Dept. of Comput. Sci., Cornell Univ., Ithaca, NY, USA
Journal title	J. Am. Soc. Inf. Sci. (USA)
Vol/issue no's	vol.36, no.3
Page numbers	200-10
Publishing date	May 1985
CODEN	AISJB6
Abstract	Automatic feedback methods may be used in online information retrieval to generate improved query statements based on information contained in previously retrieved documents. In this study automatic relevance feedback techniques are applied to Boolean query statements. The feedback operations are carried out using both the conventional Boolean logic, as well as an extended logic producing improved retrieval effectiveness. Experimental output is included to evaluate the automatic feedback operations.
Free-indexing terms	online information retrieval, query statements, automatic relevance feedback techniques, Boolean logic, retrieval effectiveness, automatic feedback

図 1 INSPEC テープのレコードの例
Fig. 1 A record of INSPEC tapes.

ぶ．しかし，このままでは復号ができないため，マトリックスの前に4ビットのプレフィックスと呼ぶ固定長部を置き，ここにマトリックスのビット長の2進数を入れる．例えば，$r=45$ の場合マトリックスは 101101 −100000＝01101 で，プレフィックスは 0101 であるので，符号は 010101101 となる．この符号を順位符号と呼ぶことにする．順位符号のマトリックスにおいて最上位けたを除去するのは，このけたが常に1であるため，これがなくても順位がわかるためである．プレフィックスの長さを4にしたのは，この場合が最も圧縮率が高いからである[8]．順位 r の単語の生起確率を $p(r)$ とすると，ジップの法則は，次の式で表される[4]．

$$p(r)=0.1/r.$$

順位符号は，ジップの法則が成立する任意の言語に対して，

圧縮効率＝エントロピー/平均符号長＝97.21%

であることを示すことができる[5]．ただし，順位符号はハフマン符号のように最適ではない．ハフマン符号は，同じ条件のもとで効率が 99.92% に達する．しかし，符号器と復号器の性能は，順位符号の方がハフマン符号よりすぐれていて，10倍以上高速である[5]．

さて，自然言語の単語の数は，事実上無限と言ってよいので，順位符号による符号化が可能であるのは高頻度単語だけであり，低頻度単語は別の符号化を行わねばならない．また，順位符号は単語を圧縮単位とする他の符号同様，原理的には単語＋1空白が符号化の単位であり，単語を構成する文字が大文字であるか小文字であるかの区別をつけない．すなわち，順位符号は可逆符号 (reversible code) ではない．つまり，順位符号によって符号化したものを復号すると，大文字と小文字の区別がなくなり，カンマやピリオド，2個以上の空白などの単語間のデリミタが1個の空白に置き替わる．

そこで，プレフィックスが0から 12 までの値をもつときだけ，そのあとに順位符号のマトリックスが続いているものとする．このとき，単語の英字は小文字とみなす．順位符号の符号化の対象となる高頻度単語が大文字を含む符号の場合は，プレフィックスの値に13をおく．単語の英字が大文字ばかりのとき，続く1ビットを0とする．先頭の文字が大文字のときは，そのビットを1とする．そして，その後に順位符号を置く．ただし，文頭の単語の場合，この処理は次のように変わる．単語の英字が大文字ばかりのとき続く1ビットを1とし，小文字ばかりのときは0とする．先頭の文字のみ大文字である単語には，値 13 のプレフィックスを置かない．以上の処理は，テキストが大文字と小文字の両方で表現されている場合であって，テキストに小文字が使われていない場合は，当然この処理を行わない．この場合，プレフィックスが0〜12のとき，英字は大文字として復号される．

次に，低頻度単語の符号について述べる．順位符号のプレフィックスの最大が 12 であるので，順位符号で符号化可能な単語数は 8,191 である．したがって，この数より大きい順位の単語は，別の符号により符号化しなければならない．この符号は，種々の方式が考えられるが，順位符号で符号化する単語数をある程度大きくすれば，低頻度単語符号化法が全体の効率に及ぼす影響は小さい[5]．そこで，QOC では単純に1文字を5ビットで表す符号を採用した．このときは，プレフィックスを 15 にする．また，この符号列の最後には5ビットの終端符号を置く．ただし，5ビットでは，すべての文字を表すことができないので，シフトコードにより英大文字，英小文字，数字，特殊記号の字種を切り替える．このシフトコードも，すべて5ビットである．

可逆符号化の最後として，単語間のデリミタの符号化について述べる．デリミタが1個の空白の場合，QOC ではそれを陽に指定しないが，それ以外の1文字以上のデリミタは，その出現頻度に応じて 0, 100, 101, 11000, 11001, 11010, 11011, 111000, 111001, … のような可変長符号で表す．もちろん，出現頻度が高いほど短い符号を割り当てる．この符号では，最初に現れる0が符号の終りを指示する．順位が 1, 2, 3 のデリミタにはそれぞれ 0, 100, 101 を，順位が4以上のものについては最初の0の後の2ビットまでを符号として割り当てる．デリミタは異なり数が小さく，その頻度分布は高順位部分への偏りが非常に大きいためこのような符号化を行った．この符号の前に 14 を値とするプレフィックスを置き，他の符号と区別する．**図2** に QOC の圧縮形式を，**表1** と **表2** に高順位単語とデリミタの符号を示す．

3.2 1バイトまたは2バイトの整数符号

言語名のように取りうる項目値の異なり数が小さい書誌事項については，単純にそれぞれの値を1バイトあるいは2バイトの整数で符号化した．

3.3 文字を単位とする符号

QOC では圧縮効果が低い書誌事項，あるいは取り

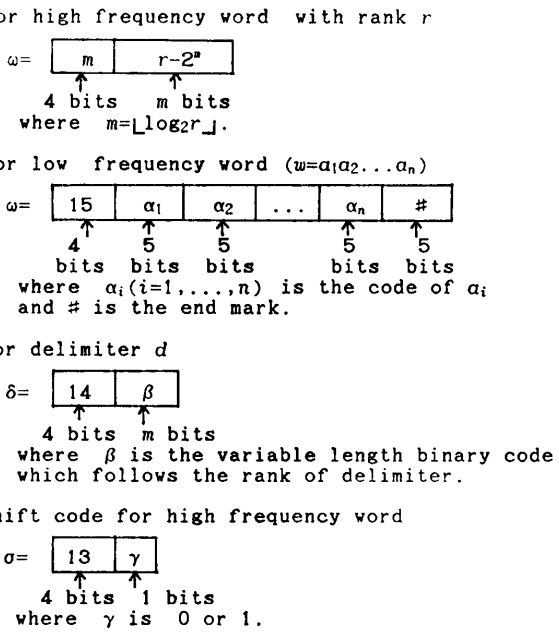

図 2 QOC の形式
Fig. 2 Format of QOC.

うる項目値の異なり数が大きいものについては，文字を単位に1文字を4ビットあるいは5ビットで符号化する．英大文字，英小文字，数字，特殊記号を含む書誌事項については1文字を5ビットで符号化する．この場合も，5ビットのシフトコードによって4種類の字種を切り替える．出現文字が，数字と幾つかの特殊記号に限られている書誌事項に対しては，1文字を4ビットで符号化する．

4. データ圧縮プログラム

QOC 以外は，符号の要素数がたかだか 256 であるので，それらについてのプログラムの仕様と実現については特に述べることはない．そこで，QOC を実現するプログラムについてその仕様と構成ならびに実現法について述べる．

4.1 QOC プログラムの仕様と構成

3.1 節からわかるように，QOC プログラムでは次の(1)～(3)が指定されていなければならない．

1) 順位符号の要素数：この数を QOC の位数(order)と呼ぶ．この数の最大は 8,192 である．
2) デリミタ：順位符号の要素である単語を切り出すために必要である．
3) 英字は大文字だけかそうでないかの別：英小文字を含む場合，英小文字列以外の単語に対する処理が付加される．

表 1 高順位単語の順位符号
Table 1 Rank code for the high frequency words.

word	rank	code
THE	1	0000
OF	2	00010
AND	3	00011
A	4	001000
TO	5	001001
IN	6	001010
IS	7	001011
FOR	8	0011000
SYSTEM	9	0011001
ARE	10	0011010
WITH	11	0011011
COTROL	12	0011100
SYSTEMS	13	0011101
COMPUTER	14	0011110
ON	15	0011111
AN	16	01000000
DATA	17	01000001
BY	18	01000010
BE	19	01000011
THIS	20	01000100

表 2 デリミタの符号
Table 2 Code for the delimiters.

delimiter	rank	code
,B	1	11100
.B	2	1110100
-	3	1110101
B(	4	111011000
/	5	111011001
)B	6	111011010
B	7	111011011
)	8	1110111000
'	9	1110111001
.	10	1110111010
;B	11	1110111011
:B	12	11101111000
B'	13	11101111001
B,	14	11101111010
'B	15	11101111011
B.	16	111011111000

'B' denotes a blank character

QOC プログラムは符号化および復号化プログラムと，それらが使用する符号表を作成するプログラムから構成されている．後者の表作成プログラムは，上記の(1)～(3)の情報と標本テキストを与えると，順位符号の要素となる単語およびデリミタに関する符号化表と復号化表を作成する．QOC の低頻度単語に対する文字を圧縮単位とする符号表は，標本テキストによらず共通である．

4.2 QOC プログラムの実現法

会話型情報検索システムのためのデータ圧縮プログラムでは，符号化プログラムより復号化プログラムの方の効率が重要である．QOC において，復号化プログラムの速度を決めるのは順位符号の部分である．この部分のプログラムは，IBM 360/370 方式の計算機では，その機械の固有命令を活用した実現が可能であ

る．図3に順位符号のための復号化表の構成を示し，図4にその復号化プログラムを示した．図4では，簡単のため最初の符号の復号だけを示し，続く復号のための繰返し部分を省いている．図3と図4から，順位符号の復号には，IBM 360/370 方式機械の命令が巧妙に使われていることがわかる．もちろん，このプログラムは 31 ビット長の拡張アドレスモードでは動作しない．この理由は，単語へのポインタの最初の1バイトに単語長 −1 が入っているためである．この値は次のポインタとの差によって求まるので，不必要のように思われるかもしれないが，そのようにすると復号化プログラムの速度が低下する．図4のプログラムは改良の余地はほとんどないようである．例えば，3.1 節のジップの法則を仮定すると 10 回に1回の割合で "THE" が復号されるはずである．そこで，"THE" の場合は図3の表を参照しないようにプログラムを変更すると，"THE" であるかどうかの判断のために，かえって復号速度が低下した．

符号化プログラムは，ハッシングを用いた常識的な方法で実現した．

順位符号の符号表は，位数を最大の 8,191 にとった場合でも必要な領域は 250 キロバイト程度であり，一次記憶上に置くことができる．

なお，符号化および復号化プログラムはアセンブラ言語で書かれているが，符号表作成プログラムは Fortran 77 によって作成した．

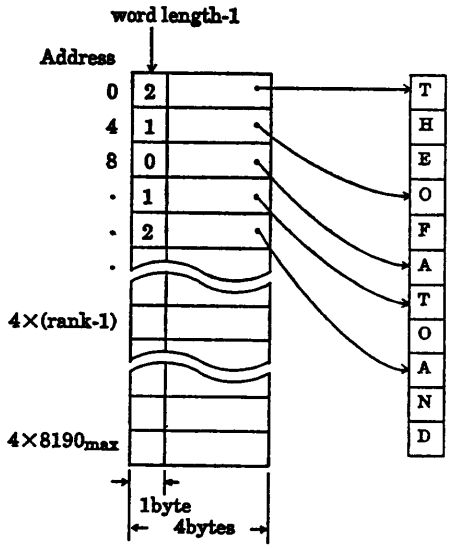

図 3 順位符号復号化表の構成
Fig. 3 Organization of table for decoding of rank code.

5. INSPEC テープの圧縮

この章では，1983年に配布された1年分の INSPEC テープの圧縮を行い，本稿の圧縮技法を評価する．分野による違いを見るために，このテープから次の三つの文献集合を作った．

```
* REGISTERS
P1      EQU   5     : pointer to area for encoded information (input)
P2      EQU   6     : pointer to area for decoded information (output)
B1      EQU   7     : base register for decoding table
*                       B1 = address of decoding table  - 4
R0      EQU   0     : work register
R1      EQU   1     : work register
R2      EQU   2     : work register
R3      EQU   3     : work register
R4      EQU   4     : work register
          ...
        L     R1,0(P1)          put input data on R1
        LA    P1,4(P1)          increase P1 to point to next input data
        SR    R3,R3             clear R3
LOOP    EQU   *
        SR    R0,R0             clear R0
        SLDL  R0,4              shift R0 and R1 to obtain prefix of Rank code
        LR    R2,R0             move prefix to R2
        LA    R0,1              put 1 on R0 to prepare R0 for first bit of word rank
        SLDL  R0,0(R2)          shift R0 and R1 to obtain word rank
        LR    R2,R0             move word rank to R2
        SLL   R2,2              multiply R2 by 4 to calculate address
        IC    R3,0(B1,R2)       put word length on R3
        STC   R3,MVCL+1         store word length to MVC operand
        L     R4,0(B1,R2)       put address of word to be decoded on R4
        LA    R4,0(B1,R4)       put absolute word address on R4 for next MVC
MVCL    MVC   0(1,P2),0(R4)     move word to output area
        AR    P2,R3             increase P2 to point to the last character of word
        MVI   1(P2),C' '        put a blank character on output area
        LA    P2,2(P2)          increase P2 to point on the address of next word
          ...
```

図 4 順位符号の復号化プログラム
Fig. 4 Decoding program of rank code.

文献集合名	分野
INSPEC-A	物理学
INSPEC-B	電気工学，電子工学
INSPEC-C	制御工学，計算機科学，情報工学

分野による文献の振分けは，INSPEC テープの分野コードに従った．この三つの文献集合は互いに素ではなく，19% の文献が複数の集合に含まれている．また，1983 年 8 月からは情報工学に対する分野コードが設けられたが，このコードをもつ文献は多くないので，これらは INSPEC-C に含めた．

三つの集合の文献数は，INSPEC-A が 122,103，同-B が 64,628，同-C が 44,299 である．なお，1983 年の INSPEC テープの文献数は 194,182 である．

5.1 圧縮方法

抄録，標題，自由索引句の圧縮には QOC を適用した．位数は最大の 8,191 で，英数字を除くすべての特殊記号をデリミタとした．また，誌名，著者の所属機関名などの圧縮にも QOC を適用した．ただし，誌名，著者の所属機関名などでは，標題や抄録のような文章情報と異なり，出現文字列が限られ，またピリオドで終止する省略形が多い．位数は最大の 8,191 とし，デリミタとしては，空白，“,”，“-”，“'”，“/”，“$” の 6 文字を選択した．ここでは，QOC を文章情報のためのものと，定形データ型書誌情報のためのものと区別し，前者を QOC 1，後者を QOC 2 と呼ぶことにする．QOC 2 の適用対象は，誌名と著者の所属機関名のほか，会議名，会議開催場所，出版社名，出版社所在地，後援機関の書誌事項である．

言語名は，1 バイトに符号化する．言語名の異なり数は 50 以下であった．言語名は，一つの文献が複数の言語を用いて記載されている場合がある．

その他の書誌事項は，文字単位に 1 文字を 4 ビットあるいは 5 ビットで符号化した．書籍番号とページ番号は，使用される文字が数字と幾つかの特殊記号に限られているので，1 文字を 4 ビットで符号化し，それら以外の著者名，CODEN，巻数，ページ数，発行年月日は，1 文字を 5 ビットで符号化した．

ただし，巻数では “Vol.”，“Ser.”，“No.” が，発行年月日では “Jan.”，“Feb.”，…，“Dec.” の文字列が頻繁に現れるので，これらの文字列は 5 ビットに符号化した．また，著者名は “Smith, A. B.” のように姓 “Smith” の後にイニシアルが “A. B.” のような形式で続くことがほとんどである．そこで，イニシアル部分は 5 ビットの特別な符号のあとに，ピリオドを除いたイニシアルの 5 ビット文字列を置くことにした．もちろん，著者名には “Losos, F. J., III” や “Van Auken, B. J., Jr.” のように形式の異なるものがあるが，これらに対しては 5 ビットの特別な符号で状態を切り替えることにした．

5.2 QOC の符号表

QOC 1 の符号表は，INSPEC-A, 同-B, 同-C ごとに作った．QOC 2 の符号表は 3 種の文献集合に共通なものを一つ作成した．QOC 1 の符号表を分野別に作成したのは，分野により使われる単語が異なるためであり，QOC 2 の符号表を共通にしたのは，誌名や著者の所属機関名などにおいては，単語の異なり数が小さく，分野による違いがあまりないためである．**表 3** に高頻度単語を，**表 4** に高頻度デリミタを示した．

5.3 3 文献集合の圧縮

INSPEC テープからの 3 文献集合 INSPEC-A, 同-B, 同-C に対して，圧縮テストを行った結果を**表 5** に示す．ここで，圧縮率とは 8 ビット/字の符号で表現されたもとのデータの大きさを圧縮されたデータ量で割った値のこととする．QOC 1 の各項目の圧縮率は，文献集合によって異なり，INSPEC-C が最も良く，INSPEC-A が最も悪い．QOC 2 の各項目の圧縮率は，3 文献集合で QOC 1 ほどの違いはない．QOC 1 の 3 項目全体での圧縮率は，INSPEC-A, 同-B, 同-C がそれぞれ 3.58, 3.88, 4.03 であり，QOC 2 の 7 項目全体での圧縮率は，それぞれ 3.14, 3.23, 3.25 であった．それ以外の項目の圧縮率は，文献集合による違いはほとんどない．1 文字を 5 ビットで符号化した項目のうち著者名，巻数，発行年月日は，頻繁に現れる文字列を 5 ビットに符号化していることにより他の項目に比べて圧縮率が高い．

全項目での圧縮率は，3.31〜3.64 であった．符号化時間と復号化時間は，各項目とも 3 文献集合でほぼ同じであり，3 文献集合および全項目にわたっての平均符号化時間および復号化時間は，FACOM M-382 計算機で 1 文字当り，それぞれ 3.90 マイクロ秒，1.09 マイクロ秒であった．

5.4 検　　討

QOC は，抄録だけでなく標題や自由索引句，さらにジップの法則が明らかに成り立たない誌名や著者の所属機関のような書誌事項についても有効であることがわかった．

QOC 1 の 3 項目の圧縮率は，3 文献集合とも標題，抄録に比べて自由索引句が低い．分野による違いが

表 3 INSPEC テープ 8301-8324 における高頻度単語
Table 3 High frequency words of INSPEC tapes 8301-8324.

rank	Word from title, abstracts, and free-indexing terms						Word from journal title, author affiliation, etc.	
	INSPEC-A		INSPEC-B		INSPEC-C			
1	THE	6.48%	THE	5.96%	THE	5.93%	OF	4.74%
2	OF	4.41	OF	4.13	OF	4.08	USA	3.39
3	AND	2.22	AND	2.46	AND	2.67	(USA)	2.62
4	IN	1.86	A	2.20	A	2.49	PHYS.	2.62
5	A	1.81	IN	1.63	TO	1.75	UNIV.	2.59
6	TO	1.39	TO	1.52	IN	1.57	&	2.00
7	SUB	1.30	IS	1.30	IS	1.52	J.	1.51
8	IS	1.17	FOR	1.20	FOR	1.37	DEPT.	1.51
9	FOR	0.96	ARE	0.82	SYSTEM	1.00	INST.	1.47
10	WITH	0.74	WITH	0.76	ARE	0.89	SCI.	1.18
11	ARE	0.72	BY	0.59	WITH	0.68	(GB)	0.97
12	SUP	0.71	ON	0.57	CONTROL	0.66	IEEE	0.97
13	2	0.71	SYSTEM	0.52	SYSTEMS	0.66	THE	0.93
14	BY	0.62	AN	0.49	COMPUTER	0.61	RES.	0.78
15	ON	0.55	SUB	0.47	ON	0.57	CONFERENCE	0.75
16	THAT	0.45	AS	0.44	AN	0.56	(USSR)	0.73
17	AT	0.42	BE	0.41	DATA	0.55	ON	0.72
18	1	0.41	THAT	0.38	BY	0.53	PROCEEDINGS	0.72
19	FROM	0.39	POWER	0.37	BE	0.47	CA	0.68
20	AN	0.39	THIS	0.36	THIS	0.47	LAB.	0.67

表 4 INSPEC テープ 8301-8324 における高頻度デリミタ
Table 4 High frequency delimiters of INSPEC tapes 8301-8324.

rank	Delimiter from title, abstract, and free-indexing terms						Delimiter from journal title, author affiliation, etc.	
	INSPEC-A		INSPEC-B		INSPEC-C			
1	,B	33.89%	,B	44.51%	,B	49.62%	,B	68.21%
2	-	12.96	-	15.47	.B	15.83	$	13.43
3	/	12.88	.B	13.40	-	13.50	B$	11.45
4	.B	9.64	/	6.76	B(	3.09	-	4.48
5	B	3.75	B(	3.00	/	2.68	'	0.77
6	B(	3.41	.	2.20	)B	1.91	/	0.52
7	B/	3.32	B	1.95	B	1.64	B'	0.46
8	.	2.61	)B	1.94	)	1.28	B	0.25
9	/B	2.27	/B	1.28	'	1.25	B-	0.20
10	)B	2.10	)	1.02	.	1.25	-B	0.09
11	B,	1.80	B,	0.91	;B	1.17	'B	0.05
12	)	1.42	B/	0.80	:B	1.06	B,	0.05
13	B-	1.36	;B	0.78	B'	0.81	/B	0.01
14	B.	0.86	B-	0.64	B,	0.80	,	0.01
15	(	0.81	B.	0.63	'B	0.70	B/	0.00
16	B)	0.76	'	0.61	B.	0.70	$B	0.00

'B' denotes a blank character

最も顕著であるのは，この QOC 1 である．特に，INSPEC-A の圧縮率が他の二つの集合に比べて悪い．順位符号部分のエントロピー，圧縮効率，平均単語長等を**表 6** に示す．これから，圧縮効率は分野による差がほとんどないが，平均単語長は INSPEC-A が他のものに比べて短いことがわかる．INSPEC-A の圧縮率が悪いことは，ここに主な原因があるように思われる．

QOC 2 の 7 項目の圧縮率は，2.61～4.14 と項目による差が大きい．この中では，会議名と出版社名は，高頻度単語が占める割合が大きく，圧縮率が高い．著者の所属機関名は，低頻度の単語の割合が大きく，圧縮率は低い．会議開催場所と後援機関は，高頻度の単語の割合が大きいにもかかわらず，圧縮率は低い．これは，デリミタの生起が大きいためである．

整数符号を用いて符号化したのは，1 バイトで表現可能な言語名だけに限ったが，QOC 2 で符号化した会議名，会議開催場所，出版社名，出版社所在地，後援機関も，値の異なり数が小さいため，固定長の整数による符号化が可能である．INSPEC テープ 8301～8324 におけるこれらの書誌事項値の異なり数と生起回数を**表 7** に示す．これら五つの書誌事項の符号表の大きさは 50 キロバイト程度である．

著者名の圧縮率は良くない．著者名は，生起頻度の

表 5 INSPEC テープ 8301-8324 に対するデータ圧縮の効果
Table 5 Results of compression for INSPEC tapes 8301-8324.

compression method	item	INSPEC-A data size (Mega-bytes)	INSPEC-A comp-ression ratio	INSPEC-A encoding time (μs/char)	INSPEC-A decoding time (μs/char)	INSPEC-B data size (Mega-bytes)	INSPEC-B comp-ression ratio	INSPEC-B encoding time (μs/char)	INSPEC-B decoding time (μs/char)	INSPEC-C data size (Mega-bytes)	INSPEC-C comp-ression ratio	INSPEC-C encoding time (μs/char)	INSPEC-C decoding time (μs/char)
QOC1	title	9.31	3.64	4.61	1.24	4.25	3.95	4.57	1.22	2.60	4.09	4.42	1.15
	abstract	76.65	3.64	3.90	0.96	35.21	3.93	3.96	0.93	22.99	4.07	3.70	0.86
	free-indexing terms	22.94	3.36	3.59	1.00	10.19	3.69	3.60	0.97	6.05	3.87	3.35	0.90
QOC2	journal title	3.76	3.23	5.91	1.78	2.33	3.52	5.34	1.53	1.71	3.55	4.98	1.37
	author affiliation	6.14	3.00	4.83	1.47	2.40	2.95	5.02	1.50	1.50	2.98	4.79	1.37
	conference title	0.95	4.14	3.90	1.07	0.28	4.13	4.01	1.11	0.08	3.82	3.94	1.04
	location of conference	0.39	3.10	6.36	2.49	0.31	3.00	6.68	2.43	0.21	3.07	6.18	2.19
	publisher	0.13	3.53	6.42	2.80	0.15	3.86	6.79	2.87	0.13	3.85	6.24	2.52
	place of publication	0.17	3.49	6.08	2.44	0.23	3.22	6.77	2.49	0.19	3.14	6.36	2.26
	sponsors of conference	0.21	2.61	5.51	1.84	0.15	2.94	6.24	2.15	0.11	3.14	5.75	1.99
1byte per value code	language	0.07	6.85	3.01	2.13	0.08	6.68	2.49	1.70	0.05	6.64	2.26	1.41
4bits per character code	ISBN	0.08	1.86	2.16	1.86	0.04	1.86	1.90	1.61	0.05	1.86	1.72	1.47
	page numbers	0.69	1.57	3.69	3.32	0.36	1.40	3.04	2.67	0.25	1.35	2.61	2.31
5bits per character code	author	4.14	1.70	1.68	1.34	1.71	1.69	1.71	1.37	0.89	1.68	1.75	1.39
	CODEN	0.66	1.33	4.27	3.61	0.30	1.33	3.58	3.10	0.19	1.32	3.19	2.72
	vol/issue no's	1.37	2.23	2.41	2.02	0.60	2.21	2.13	1.81	0.37	2.30	1.93	1.64
	no. of page	0.09	1.16	3.89	3.06	0.11	1.15	3.55	2.85	0.10	1.17	2.97	2.32
	publishing date	1.07	1.67	2.63	2.59	0.51	1.63	2.42	2.38	0.34	1.59	2.21	2.16
total or average		128.82	3.31	3.91	1.11	59.21	3.52	3.97	1.09	37.81	3.64	3.75	1.01

表 6 順位符号部分のエントロピー，圧縮効率，平均単語長
Table 6 Entropy, efficiency, and average word length about the words coded by rank code.

item		INSPEC-A	INSPEC-B	INSPEC-C
Proportion of the words coded by rank code (%)	title	94.53	95.20	94.49
	abstract	95.74	96.25	96.14
	free-indexing terms	92.45	93.76	92.65
Entropy (bits/word)	title	9.65	9.93	9.66
	abstract	9.54	9.67	9.55
	free-indexing terms	10.89	10.95	10.63
Efficiency (%)	title	89.50	89.54	89.56
	abstract	92.24	92.30	92.31
	free-indexing terms	86.21	85.84	85.38
Average word length +1	title	6.47	6.87	6.98
	abstract	5.96	6.18	6.27
	free-indexing terms	7.59	7.98	8.32

高い姓に対し QOC を適用することも考えられる．1983年の1年分の INSPEC テープを調べたところ姓の異なり数は 116,298 であった．姓の分布はかなり平坦で，高頻度の 1,024, 2,048, 4,096, 8,192 の姓をとると，それらが生起頻度で全体に占める割合はそれぞれ 25%, 32%, 41%, 51% である．これは文章情報において高頻度単語が占める割合に比べると非常に低い．抄録では，高頻度の 1,024, 2,048, 4,096, 8,192 の単語でそれぞれ 70%, 80%, 90%, 95% を占める．高頻度の 4,096 個の姓を符号表に置き，圧縮率の改善を試みたところ，圧縮率は 10% 程度しか改善されなかった．このため，著者名は QOC による符号化を行っていない．著者名については，音節などの部分文字列を単位とする符号化も考えられる．

ここで採用した符号化法による INSPEC テープに対する平均符号化時間および復号化時間は，FACOM M-382 計算機で1文字当りそれぞれ 3.90 マイクロ秒，1.09 マイクロ秒であった．特に，オンライン文献検索などへの応用において重要な復号化時間は，1文献当り INSPEC-A, 同-B, 同-C で，それぞれ 1.71 ミリ秒，1.00 ミリ秒，0.86 ミリ秒であった．

テキストが英大文字だけの場合の圧縮率や符号化・復号化時間への影響を見るために，3文献集合の英小文字を英大文字に変換して圧縮したところ，QOC 1 の部分では，INSPEC-A, 同-B, 同-C の圧縮率は，それぞれ 3.73, 4.03, 4.16 であった．したがって，3～4% 圧縮率が向上することがわかった．同様に，QOC 2 では，英大文字だけの場合の圧縮率は，INSPEC-A, 同-B, 同-C で，それぞれ 4.40, 4.46, 4.46 であった．したがって，QOC 2 では，37～40% も圧縮率が向上する．全項目での平均圧縮率は，INSPEC-A, 同-B, 同-C でそれぞれ 3.54, 3.77, 3.87 であったので，INSPEC-C では 6%，同-A と B では 7% 圧縮率が向上することがわかる．英大文字だけの場合の全項目についての平均符号化時間および復号化時間は，大文字と小文字混在の場合と同様に分野による違いはほとんどなく，1文字当りそれぞれ 3.04 マイクロ秒，0.79 マイクロ秒であった．したがって，符号化時間で 28%，復号化時間で 38% 速くなることがわかった．このように，大文字，小文字混在による影響は，データ圧縮率より，符号化・復号化時間の方が大きい．

表 7 INSPEC テープ 8301-8324 での会議名，出版社名などの項目値の異なり数と生起回数
Table 7 Numbers of different values and value occurrences of conference title, publisher, etc. for INSPEC tapes 8301-8324.

item	no. of different values	no. of occurrences
conference title	375	15,302
location of conference	457	42,431
publisher	288	27,662
place of publication	179	27,665
sponsors of conference	271	30,492

6. 圧縮プログラムの実用

本稿の圧縮プログラムは，会話型情報検索システム AIR において実用に供されている．AIR は，九州大学大型計算機センターにおいて著者らが開発したシステムで，1983年11月から同センターの FACOM M-382 計算機システムにおいて使用されるようになった．現在，AIR によって情報検索サービスが行われている文献データベースは，INSPEC テープ，JICST 科学技術文献ファイルの情報工学関係，東京理科大学の RAMBIOS である．このうち，データ圧縮を行っているのは，INSPEC テープだけであり，残りの二つ

のデータベースについてはデータ量が小さいため，データ圧縮を実施していない．

同センターが，FAIRS-I[9]によって INSPEC テープの検索サービスを開始したのは 1979 年である．FAIRS-I は，データ圧縮機能を持たず，また圧縮プログラムを組み込むこともできなかった．このような状況のもとでは，同センターのディスクの事情は INSPEC テープの 3〜5 年分以上のデータをディスク上に置くことを許さず，この検索サービスは遡及度の点で問題があった．AIR の開発により，本稿の圧縮プログラムの使用が可能になり，また同センターのデータベース用ディスクがその後 IBM 3330 型から IBM 3380型に変わったこともあり，1969年からのINSPECの全データをディスクにおくことが可能になった．復号化プログラムは高速なため，オンライン検索時のデータ復号による文献表示の遅延を検索者に全く感じさせない．また，符号化プログラムも十分高速であり，データベース構築およびデータ追加時に符号化プログラムによって作業が遅れることは全くない．

なお，FAIRS-I には 1986 年 12 月から QOC に基づくデータ圧縮機能が組み込まれている．

7. むすび

ここでは，QOCと呼ぶ単語を圧縮要素とする符号化法と，これを中心とした二次文献データの圧縮法とそのプログラムについて述べ，続いて 1 年分の INSPEC テープを対象に，この技法の性能を評価した．その結果，文章情報の圧縮率は 3.36〜4.09，二次文献データ全体での圧縮率は 3.31〜3.64 と非常に良好であった．また，オンライン文献検索などへの応用において重要となる復号化時間が FACOM M-382 計算機で 1 文字当り 1.09 マイクロ秒であり，1 文献当り 0.86 ミリ秒〜1.71 ミリ秒であったので，この技法は完全に実用に耐えることがわかった．

この圧縮プログラムは，九州大学大型計算機センターにおいて，会話型情報検索システム AIR による INSPEC テープの検索サービスに使用されている．このデータ圧縮によって，1969年からの全データをディスクに置くことが可能になり，遡及度が大幅に改善された．データ圧縮に伴う速度面への悪影響は，このデータベースの検索および維持において全く認められない．

本稿では，INSPEC テープだけを対象としたが，ここで述べた技法は他の英文二次文献情報に対しても有効であると考えられる．今後は，JICST 科学技術文献ファイルや JAPAN-MARC を対象にして和文二次文献情報の圧縮技法を同様な手法で開発する計画である．

参考文献

1) Radue, J. E.: Text Compression Techniques, *Quaestiones Informaticae*, Vol. 1, No. 1, pp. 30-36 (1979).
2) Huffman, D. A.: A Method for the Construction of Minimum Redundancy Codes, *Proc. IRE*, Vol. 40, No. 9, pp. 1098-1101 (1952).
3) Gilbert, E. N. and Moore, E. F.: Variable-length Binary Encodings, *Bell Syst. Tech. J.*, Vol. 38, No. 4, pp. 933-967 (1959).
4) Shannon, C. E.: Prediction and Entropy of Printed English, *Bell Syst. Tech. J.*, Vol. 30, No. 1, pp. 50-65 (1951).
5) 松尾文碩，二村祥一，吉田　将：準最適テキスト圧縮符号，九大工学集報，Vol. 55, No. 2, pp. 103-106 (1982).
6) 松尾文碩，二村祥一，吉田　将：科学技術論文抄録における単語の統計的性質，九大工学集報，Vol. 54, No. 4, pp. 411-416 (1981).
7) Aitchison, T. M., Martin, M. D. and Smith, J. R.: Developments towards a Computer Based Information Service in Physics, Electrotechnology and Control, *Inform. Storage and Retrieval*, Vol. 4, No. 2, pp. 177-186 (1968).
8) 松尾文碩，二村祥一，吉田　将：英文テキスト圧縮についての一考究，九大工学集報，Vol. 54, No. 4, pp. 407-409 (1981).
9) 計算機マニュアル FACOM OS Ⅳ FAIRS-I 解説書，富士通(株) (1978).

(昭和 60 年 12 月 11 日受付)
(昭和 61 年 12 月 10 日採録)

二村　祥一（正会員）

昭和 23 年生．昭和 48 年九州大学大学院修士課程通信工学専攻修了．九州大学大型計算機センター助手．情報検索システムの研究に従事．電子情報通信学会，人工知能学会各会員．

松尾　文碩（正会員）

昭和 16 年生．昭和 41 年九州大学大学院工学研究科修士課程修了．工学博士．九州大学大型計算機センター助教授．推論系，データベース，情報検索の研究に従事．