データベース・システム 34－5
（1983. 2. 17）

オンライン検索のための文献データ圧縮技法

二村　祥一　　松尾　文碩
（九州大学大型計算機センター）

1．まえがき

オンライン文献検索は，大量のディスクスペースを必要とするため，データ圧縮技法を最も必要とする分野のひとつであろう．文献データは抄録，標題など文章テキスト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効率の影響するところが大きい．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［5］．

テキスト圧縮技法のうち最も本質的で効果的なのは，文字や単語をその生起頻度に応じて，それらを可変長符号で表わす方法である．この方法では，ハフマン符号［3］は最適であることが証明されている［4］．文字を要素としたハフマン符号によるテキストの圧縮では，2倍程度の圧縮効果しかなく，それ以上の圧縮効果を上げるには，単語やN－ｇｒａｍの文字列を要素とした圧縮方法をとらざるをえない．しかし，ハフマン符号は，単語のように要素数が非常に大きい場合は，そのまま使えず別の圧縮技法と組み合わせる必要があり，符号化・復号化に要する計算量が大きいという欠点がある．

筆者らは，ＱＯＣ（Ｑｕａｓｉ－Ｏｐｔｉｍｕｍ　Ｃｏｄｅ）と呼ぶ単語を圧縮要素とするテキスト圧縮技法を考案した．ＱＯＣは，ハフマン符号のように最適ではないが，ジップの法則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［1］．一方，符号化・復号化に関する計算量においてＱＯＣはハフマン符号に決定的に勝る［1］．ＱＯＣによりＩＮＳＰＥＣテープの圧縮を試みたところ抄録，標題のような文章情報のみならず，誌名，著者の所属機関などに対してもこの技法が有効であることが確認できた．すなわち，これらについては約1／4にデータを圧縮することができ，ＦＡＣＯＭ　Ｍ－２００上で実現した符号器及び復号器の速度は，それぞれ約5．85マイクロ秒／字，1．48マイクロ秒／字と極めて高速であった．

ここでは，ＱＯＣを含む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．符号化法

2．1　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進数によつて表現した符号である．単語ｗの順位をｒとすると，ｗの符号$\hat{w}$は，4ビットの固定長部分（プレフィックス）とそれに続く可変長の後続部分（マトリックス）からなる．プレフィックスは後続部分の長さを2進数で表示するためのもので，その値は$\lfloor \log_2 r \rfloor$であり，マトリックスの値は$r - 2^{\lfloor \log_2 r \rfloor}$である．

自然言語の単語の数は，事実上無限と言ってよいので，単語を要素とする符号化が可能であるのは高順位単語だけであり，低順位単語及び単語の切り出しに用いる複数のデリミ

タに対しては，別の符号化を行わねばならない．ＱＯＣでは，プレフィックスの値の０から１２までを高順位単語（最高８１９１語）に，１５，１４をそれぞれ低順位単語及びデリミタに用いている．また，小文字をあつかうため１３を高順位単語のシフトコードに用いている．

高順位単語のシフトコードは，

１）単語が英大文字のみからなる；

２）単語が英小文字のみからなる；

３）単語の先頭が英大文字で，あとに英小文字が続く；

のいずれであるかを区別するのに用いる．

低順位単語及びデリミタの符号化法には，種々の方式が考えられるが，ＱＯＣでは次のような単純な方法を採用した．いま，単語ｗを$w = a_1 a_2 \cdots a_n$とするとプレフィックス１５に続く符号を$\alpha_1 \alpha_2 \cdots \alpha_n$#とする．$\alpha_1 \alpha_2 \cdots \alpha_n$は，それぞれ$a_1 a_2 \cdots a_n$の符号で，#はエンドマークである．$\alpha_1 \alpha_2 \cdots \alpha_n$の中には，必要に応じて文字シフトコード（英大文字，英小文字，数字，その他を区別するためのもの）が挿入される．これらはいずれも５ビット長の符号である．デリミタの符号化は次のように行う．まず，単語間の空白１文字は，単語＋空白を符号化したと考え符号化の対象としない．その場合を除いて，おのおののデリミタσに対してσ，σ␣，␣σ（␣は空白）の生起頻度を調べ，その順位に従い可変長２進数０，１００，１０１，１１０００，１１００１，・・・をプレフィックス１４に続くコードとして割り当てる．可変長２進数の長さは，最初に現れる〝０〟により判定する．図１にＱＯＣの圧縮表現を，表１，表２にそれぞれ高順位単語とデリミタの符号を示す．

２．２　１バイト，２バイトへの符号化

値の異なり数の小さいものについては，おのおのの値を１バイトあるいは２バイトに符号化する．符号化・復合化には変換プログラム（あるいは変換辞書）が必要になる．

２．３　４ビット／字，５ビット／字の圧縮

ＱＯＣによる符号化では圧縮効果が少ない書誌事項，あるいは取りうる値の異なり数が大きいものについては，文字単位に４ビット／字あるいは５ビット／字に符号化する．英大文字，英小文字，数字を含む事項について

・高順位単語（$w = a_1 a_2 \cdots a_n$）

$\hat{w}$ =

m	$r - 2^m$

４ビット　ｍビット

ここでｒは単語ｗの順位，

$m = \lfloor \log_2 r \rfloor$．

・低順位単語（$w = a_1 a_2 \cdots a_n$）

$\hat{w}$ =

１５	α_1	---	α_n	#

４ビット　５ビット　　５ビット

ここでα_iはa_iのコード，

#はエンドマーク．

・デリミタ（σ）

$\hat{\sigma}$ =

１４	β

４ビット　ｍビット

ここでβはデリミタσの順位による可変長２進コード．

・高順位単語シフト（φ）

$\hat{\phi}$ =

１３	δ

４ビット　１ビット

ここでδの値は０／１で文頭など出現位置により意味が異なる．

図１　　ＱＯＣの圧縮表現

表１　高順位単語の符号

単語w	位数r（w）	符号$\hat{w}$
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
CONTROL	12	0011100
ON	13	0011101
AN	14	0011110
SYSTEMS	15	0011111
COMPUTER	16	01000000
DATA	17	01000001
BY	18	01000010
BE	19	01000011
WHICH	20	01000100

表２　デリミタの符号

デリミタσ	位数r（σ）	符号$\hat{\sigma}$
,も	1	11100
.も	2	1110100
−	3	1110101
／も	4	1110111000
も（	5	1110111001
）も	6	1110111010
も	7	1110111011
.も	8	11101111000
)も	9	11101111001
’も	10	11101111010
;も	11	11101111011
も,	12	111011111000
:も	13	111011111001
も.	14	111011111010
も’	15	111011111011
’も	16	1110111111000

はシフトコードを含む５ビット／字の符号化を行い，特に使用される文字が，数字やいくつかの特殊記号に限られている場合は，４ビット／字に符号化する．

３．ＱＯＣによるテキスト圧縮プログラム

ＱＯＣによるテキスト圧縮プログラムは，３つのモジュールＤＩＣＴ＄Ｃ，ＥＮＣＯ＄Ｃ，ＤＥＣＯ＄Ｃから構成される．ＥＮＣＯ＄Ｃ，ＤＥＣＯ＄Ｃはそれぞれ ＱＯＣによる符号化・復号化を行うプログラムである．ＤＩＣＴ＄Ｃは，ＥＮＣＯ＄Ｃ，ＤＥＣＯ＄Ｃが使用する高順位単語辞書，デリミタ辞書を与えられたテキストから作成する．

ＤＩＣＴ＄Ｃでは，テキストを構成する文字種，単語のデリミタ（複数個）を指定できる．文字種は，テキストが英小文字を含むかどうかを区別するためのものである．テキストが英小文字を含むと指定した場合は，例えば〃ＴＨＥ〃，〃ｔｈｅ〃，〃Ｔｈｅ〃を同一の単語とみなして，その生起頻度を調べる．ＤＩＣＴ＄Ｃでは，作業時に作られる単語頻度表，デリミタ頻度表を保存することもできる．

ＥＮＣＯ＄ＣとＤＥＣＯ＄Ｃが使用する高順位単語辞書とデリミタ辞書は，初期設定時に主記憶上に置かれる．この領域の大きさは，ＱＯＣの位数（高順位単語の数）を８１９１に選んだ場合，ＥＮＣＯ＄Ｃでは１６０キロバイト，ＤＥＣＯ＄Ｃで１００キロバイト程度である．

ＤＩＣＴ￥ＣはＦｏｒｔｒａｎ７７で，また，ＥＮＣＯ＄Ｃ，ＤＥＣＯ＄Ｃは高速化を

図るためアセンブラ言語で記述し，ＩＢＭ３６０／３７０アーキテクチャの性能を十分に引き出すようにしている．

４．ＩＮＳＰＥＣテープの圧縮実験

ＩＮＳＰＥＣテープ［２，６］８００１～８０２４（文献数１７２，３５１件）を対象に圧縮実験を行った．ＩＮＳＰＥＣの文献データはＡ（物理学），Ｂ（電気・電子工学），Ｃ（計算機・制御工学）の３つの分野に分類できる．ＩＮＳＰＥＣテープ８００１～８０２４での分野別の件数は，Ａ，Ｂ，Ｃそれぞれ１０８，１４１件，５５，２８４件，３６，２３２件であった．

４．１　ＩＮＳＰＥＣテープの圧縮方法

抄録，標題，自由索引句の圧縮にはＱＯＣを適用する．ＱＯＣでは単語の切り出しのためのデリミタを指定できるが，ここでは英数字を除くすべての特殊記号を登録する．

誌名，著者の所属機関などの圧縮にはＱＯＣを適用する．ただし，誌名，著者の所属機関などの場合は，抄録などのような文章とは違って，使用される単語が限られ，又省略形が多く用いられている．このため抄録，標題，自由索引句とは別に扱う．デリミタとして〝ち〟，〝，〟，〝，〟，〝－〟，〝’〟，〝／〟，〝＄〟の６文字を登録する．次の書誌事項もこのクラスに含める．

会議名，会議開催場所，出版社名，
出版社所在地，後援機関

言語名は，言語単位に１バイトに符号化する．一次文献の記載言語名の異なり数は５０以下である．言語名では一つの論文が複数の言語を用いて記載されている場合があり，それを区別する必要がある．

その他の書誌事項は文字単位に４ビット／字あるいは５ビット／字に符号化する．書籍番号のように使用される文字が，数字やいくつかの特殊記号に限られている場合は，４ビット／字に符号化し，それ以外は５ビット／字に符号化する．

・４ビット／字に符号化する書誌事項
　書籍番号，分冊番号，ページ数
・５ビット／字に符号化する書誌事項
　著者名，ＣＯＤＥＮ，レポート番号，
　雑誌のボリューム番号，参照ページ，
　発行年月日

ここで，雑誌のボリューム番号では，〝Ｖｏｌ．〟〝Ｓｅｒ．〟，〝Ｎｏ．〟が，発行年月日では〝Ｊａｎ．〟，〝Ｆｅｂ．〟，・・・が固定的に使用されているので，これらについては，文字列全体を５ビット／字に符号化する．著者名については，ほとんどの場合〝Ｓｍｉｔｈ，　Ａ．Ｂ．〟のようにファミリー名〝Ｓｍｉｔｈ〟に続いてイニシャル〝Ａ．Ｂ．〟が設定されており，これを配慮して符号化した．

４．２．辞書の作成

本実験では，符号化・復号化のために次の２種類の辞書を作成した．

１）抄録，標題，自由索引句のための高順位単語辞書とデリミタ辞書
２）誌名，著者の所属機関などのための高順位単語辞書とデリミタ辞書

辞書１は分野別に，辞書２は全分野のデータを用いて作成した．辞書１を分野別に作成したのは，分野により使用される単語が異なるためであり，辞書２で全分野のデータを用いたのは，誌名，著者の所属機関などにおい

ては用語が統一されており分野による違いがあまりないことによる．

表3 に高順位単語を，表4に高順位のデリミタをまとめる．それぞれ抄録，標題，自由索引句については分野別に，誌名，著者の所属機関などは全分野のデータを用いた．

INSPECテープではギリシャ文字は，αはALPHA，ωはOMEGAのように，特殊記号%，∫，∞はPERCENT，INTEGRAL，INFINITYのように置

表3　高順位単語

位数	INSPEC－Aの抄録，標題，自由索引句	%	INSPEC－Bの抄録，標題，自由索引句	%	INSPEC－Cの抄録，標題，自由索引句	%	INSPECの誌名，著者の所属機関など	%
1	THE	6.58	THE	6.10	THE	6.16	OF	4.62
2	OF	4.58	OF	4.34	OF	4.53	USA	3.33
3	AND	2.27	AND	2.52	AND	2.66	UNIV.	2.68
4	IN	1.89	A	2.28	A	2.64	PHYS.	2.66
5	A	1.84	IN	1.65	TO	1.72	(USA)	2.66
6	TO	1.40	TO	1.50	IN	1.60	AND	2.53
7	SUB	1.31	IS	1.34	IS	1.56	J.	1.61
8	IS	1.22	FOR	1.24	FOR	1.46	INST.	1.58
9	FOR	0.99	ARE	0.85	SYSTEM	0.99	DEPT.	1.48
10	WITH	0.75	WITH	0.81	ARE	0.92	SCI.	1.20
11	ARE	0.75	BY	0.61	WITH	0.73	IEEE	1.06
12	SUP	0.75	ON	0.58	CONTROL	0.69	(GB)	1.01
13	2	0.71	SYSTEM	0.53	ON	0.60	(USSR)	0.85
14	BY	0.63	AN	0.51	AN	0.59	RES.	0.80
15	ON	0.56	SUB	0.51	SYSTEMS	0.59	ENGLAND	0.76
16	THAT	0.44	AS	0.44	COMPUTER	0.58	THE	0.75
17	AT	0.43	BE	0.42	DATA	0.57	LAB.	0.73
18	AN	0.40	POWER	0.40	BY	0.56	GERMANY	0.65
19	FROM	0.40	WHICH	0.38	BE	0.49	ON	0.65
20	1	0.38	THAT	0.38	WHICH	0.46	CONFERENCE	0.63
21	3	0.37	THIS	0.34	THAT	0.45	NEW	0.62
22	AS	0.37	AT	0.33	AS	0.45	ENGNG.	0.61
23	BE	0.34	2	0.32	THIS	0.42	CHEM.	0.58
24	ENERGY	0.30	FROM	0.30	IT	0.30	JAPAN	0.57
25	WHICH	0.29	HIGH	0.28	TIME	0.29	SOC.	0.57
26	0	0.29	SYSTEMS	0.27	DESIGN	0.28	CA	0.56
27	TEMPERATURE	0.29	IT	0.25	CAN	0.27	(NETHERLANDS)	0.55
28	BEEN	0.27	CIRCUIT	0.25	METHOD	0.26	(GERMANY)	0.53
29	THIS	0.26	CONTROL	0.24	MODEL	0.26	PROCEEDINGS	0.52
30	FIELD	0.25	CAN	0.23	FROM	0.26	YORK	0.52
31	WAS	0.24	FREQUENCY	0.23	INFORMATION	0.26	TECH.	0.48
32	X	0.24	DATA	0.23	ANALYSIS	0.25	INTERNATIONAL	0.48
33	ELECTRON	0.24	BEEN	0.22	PROGRAM	0.24	DE	0.47
34	MODEL	0.24	METHOD	0.22	OR	0.23	USSR	0.45
35	IT	0.24	HAS	0.22	HAS	0.22	APPL.	0.44
36	4	0.22	1	0.22	PROBLEM	0.22	FRANCE	0.43
37	WERE	0.22	CURRENT	0.22	AT	0.21	STATE	0.41
38	TWO	0.21	DESIGN	0.21	USED	0.21	TECHNOL.	0.41
39	SURFACE	0.21	OR	0.20	TWO	0.21	ENG.	0.38
40	RESULTS	0.20	USED	0.20	PROCESSING	0.20	ACAD.	0.37
41	HAVE	0.20	LASER	0.20	SOFTWARE	0.19	TRANS.	0.37
42	METHOD	0.20	TIME	0.20	S	0.19	SOV.	0.35
43	HIGH	0.20	VOLTAGE	0.20	USE	0.19	LETT.	0.35
44	SYSTEM	0.20	DIGITAL	0.20	USING	0.18	NY	0.34
45	HAS	0.19	TWO	0.19	PROBLEMS	0.18	REV.	0.33

表 4　高順位デリミタ

位数	INSPEC－Aの抄録，標題，自由索引句	%	INSPEC－Bの抄録，標題，自由索引句	%	INSPEC－Cの抄録，標題，自由索引句	%	INSPECの誌名，著者の所属機関など	%
1	,b	33.89	,b	44.09	,b	47.06	,b	67.72
2	/	14.82	.b	14.43	.b	16.91	¥	13.21
3	-	10.58	-	13.09	-	11.51	b¥	11.03
4	.b	10.39	/	7.79	/	3.71	-	4.95
5	b	3.63	b(	2.85	b(	3.25	¥b	1.48
6	b(	3.26	.	2.75	)b	2.19	'	0.68
7	.	2.99	b	2.22	b	1.96	b'	0.34
8	b/	2.92	)b	1.86	.	1.78	/	0.23
9	/b	2.34	/b	1.53	)	1.31	-b	0.10
10	)b	2.07	)	0.99	'	1.18	b	0.09
11	b,	1.77	b,	0.98	;b	1.11	'b	0.05
12	)	1.40	;b	0.82	b,	1.04	b,	0.05
13	b-	1.23	b.	0.67	:b	1.04	b-	0.04
14	b+	0.83	b-	0.65	b.	0.92	,	0.00
15	b.	0.82	b/	0.64	b'	0.82	/b	0.00
16	(	0.81	'	0.60	'b	0.70	b/	0.00

き換えられる．また，下添及び上添文字列は，／ＳＵＢ．．．／，／ＳＵＰ．．．／によって表現される［６］．例えば，$\log_2 n$はＩＮＳＰＥＣテープ上では，ＬＯＧ／ＳＵＢ２／Ｎとなる．

高順位単語について作成した辞書をみるとＩＮＳＰＥＣ－Ｃでは　計算機・制御工学で特徴的な単語ＳＹＳＴＥＭ，ＣＯＮＴＲＯＬ，ＣＯＭＰＵＴＥＲ，ＤＡＴＡがそれぞれ第９，１２，１６，１７位にある．ＳＵＢ，ＳＵＰは第５７，１５５である．ＩＮＳＰＥＣ－Ａでは物理学で特徴的な単語ＥＮＥＲＧＹ，ＴＥＭＰＡＲＡＴＵＲＥ，ＦＩＥＬＤ，ＥＬＥＣＴＲＯＮがそれぞれ第２４，２７，３０，３３位にある．ＳＵＢ，ＳＵＰは第７位，第１２位と非常に高位に位置づけられている．

誌名，著者の所属機関などについては〝ＵＮＩＶ．〟，〝ＰＨＹＳ．〟，〝Ｊ．〟などの省略形，ＵＳＡ，ＥＮＧＬＡＮＤ，ＧＥＲＭＡＮＹなどの国名，地方名が多く用いられている．また，ＩＮＳＰＥＣでは誌名の最後に国名が括弧を伴って入っているため，〝（ＵＳＡ）〟，〝（ＧＢ）〟などが高順位に位置している．

デリミタで高順位を占めるものは，〝，b〟，〝．b〟，〝－〟でＩＮＳＰＥＣ－Ｂ，Ｃではこれだけで７０％を占める．ＩＮＳＰＥＣ－Ａでは〝／〟の出現が大きい．誌名，著者の所属機関などでは値の複数生起を表わす〝＄〟がよく使用されている．

４．３　ＩＮＳＰＥＣテープの圧縮

ＩＮＳＰＥＣテープから分野Ｃの文献を抽出し圧縮実験を行った結果を表５に示す．圧縮率はＱＯＣ１（標題，抄録，自由索引句）によるものは全体では４．０８，ＱＯＣ２（誌名，著者の所属機関など）によるものは全体で３．８１　，一律に１バイトに符号化した言語名は６．６１，書籍番号，ページ数など４ビット／字に符号化したものは１．７～１．９であり（分冊番号が１．２９と低いのは，この値は〝Ｉ〟，〝ＩＶ〟など文字列の長さが短いものが多く文字列の終了を表わ

すコードが相対的に多く出現するためである），著者名，ボリューム番号，発行年月日など5ビット／字で符号化したものは1．2～2．3の圧縮率を持つ．ボリューム番号，発行年月日のようにキーワードを持つ項目に対しては圧縮率が高い．全体としての圧縮率は3．67であり，平均符号化時間及び復号化時間はＦＡＣＯＭ　Ｍ－２００でそれぞれ5．76μｓ／字，1．69μｓ／字であった．

4．4　実験結果の検討

ＱＯＣによる符号化は，抄録だけでなく標題や自由索引句，更にジップの法則が明らかに成り立たない誌名，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る．値を一律に1バイトに符号化したのは今回は言語名に限ったが，ＱＯＣ２で管理している会議名，会議開催場所，出版社所在地，出版社名，後援機関については値の異なり数が小さいため2バイトへの符号化が可能である．ＩＮＳＰＥＣテープ８００１～８０２４でのこれらの書誌事項の生起回数と異なり数は次のようであった．

	異なり数	生起回数
会議名	255	11412
会議開催場所	383	33863
出版者所在地	148	23527
出版社名	247	23544
後援機関	192	18238

上記の書誌事項をすべて管理するのに必要な記憶量は18キロバイト程度であるので，適用する書誌事項，期間等を限った場合には有効な圧縮技法となる．

著者名の圧縮率は1．79とそれほど良くない．著者名は，〃Ｓｍｉｔｈ，　Ａ．Ｂ．〃のようにファミリー名が先行しているのでこれを辞書に登録し圧縮率を上げることが考えられる．ＩＮＳＰＥＣテープ８００１～８０２４を調べたところファミリー名の異なり数は１０６，２６３であった．ファミリー名の分布は，かなり平坦で高位の１０２４，２０４８，４０９６，８１９２個を選んだ場

表５　圧縮の実験結果

	文字数 (%)	圧縮率	符号化時間 (μｓ／字)	復号化時間 (μｓ／字)
ＱＯＣ１				
標題	7．8	3．92	7．21	2．27
抄録	59．4	4．14	5．62	1．26
自由索引句	13．3	4．01	5．21	1．28
ＱＯＣ２				
誌名	4．7	4．42	6．75	2．44
会議名	0．2	4．66	5．47	1．43
会議開催場所	0．5	3．02	8．05	2．83
出版社所在地	0．5	3．18	7．95	2．91
著者所属機関	4．6	3．59	6．66	2．00
出版社名	0．3	2．93	7．98	3．29
後援機関	0．4	3．08	7．42	2．64
1バイト／値				
言語名	0．2	6．61	4．13	1．94
4ビット／字				
書籍番号	0．1	1．86	4．01	2．56
分冊番号	0．0	1．29	20．69	10．02
ページ数	0．2	1．70	6．95	3．92
5ビット／字				
著者名	2．6	1．79	3．47	2．54
ＣＯＤＥＮ	0．6	1．33	4．36	3．84
レポート番号	0．0	1．22	5．13	3．70
ボリューム番号	1．1	2．29	4．27	2．72
参照ページ	0．7	1．35	6．74	4．94
発行年月日	1．0	1．61	3．51	3．03
		3．67	5．76	1．69

合，生起頻度では，それぞれ２４％，３１％，４０％，５０％を占めるにすぎない．高位の４０９６を辞書に乗せ圧縮率の改善を試みたところ１０％程度改善されるにすぎない．このため，現在ファミリー名の辞書化は行っていない．著者名ではファミリー名の部分文字列への分解も考えられる．

５．あとがき

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２次文献データの圧縮法について述べ，続いてＩＮＳＰＥＣテープを対象に圧縮実験を行い，この技法の性能を評価した．この実験では，文章情報の圧縮率は４以上，２次文献データ全体では圧縮率は３．６７という非常に良好な結果を得た．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－２００で１文字当たり１．６８μｓ程度であり，ＩＮＳＰＥＣのデータをＣＰＵ時間１秒で７６０件程度を復号することができたので，この技法が実用に耐えることがわかった．

九州大学大型計算機センターでは，現在，富士通会話型情報検索システムＦＡＩＲＳ－Ｉを用いてＩＮＳＰＥＣの検索サービスを行っている．我々は，ＡＩＲと呼ぶ効率を重視した情報検索システムを開発しＦＡＩＲＳ－Ｉと置き代える計画である．ＡＩＲでは，ここで述べた圧縮技法を用いる．

参考文献

［１］松尾文碩，二村祥一，吉田将：順最適テキスト圧縮符号，九大工学集報，５５，２，１０３－１０６（１９８２）．

［２］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，同上，５４，４，４１１－４１６（１９８１）．

[3] Huffman, A.D.:A Method for the Construction of Minimum Redundancy Codes, Proc.IRE, 40, 9, 1098-1101 (1952).

[4]Gilbert, E.N. and Moore, E.F.: Variable-length Binary Encodings, Bell Syst. Tech. J., 38, 4, 933-967 (1959).

[5]Radue, J.E.:Text Compression Techniques, Quaestiones Informaticae, 1, 1, 30-36(1979).

[6] INSPEC Tape Services Manual, The Institution of Electrical Engineers.