自然言語処理 61－1
（1987. 5. 22）

英文科学技術文書の機能語について

松尾文碩，二村祥一
九州大学大型計算機センター

大量書誌的文献を扱う情報検索システムのための自動索引法には，不要語除去法以外に実用に耐える方法がない．このため，情報検索システムの主題検索能力と索引語転置ファイルの大きさは，否定辞書あるいは不要表と呼ばれる不要語の集合によって左右される．本稿では，英文二次文献情報INSPECテープを対象に，既存否定辞書の検索能力と転置ファイルの相対領域量を計測し，それらの限界を示したあと，新しい否定辞書構成法を提示し，この方法のよって転置ファイルの相対領域量を任意の大きさに制御できることを示す．転置ファイルが小さくなるにつれて，当然検索能力は低下するが，その低下は連続的にゆるやかに低下し，破局的に悪化することはない．最後にこの方法の欠点である否定辞書が大きくなることや分野への依存性が強いことなどの対策として一つの科学技術文献用否定辞書をつくり，これが既存の否定辞書よりすぐれていることを示す．

On Function Words of Scientific and Technical Documents Written in English

Fumihiro MATSUO and Shouichi FUTAMURA
Computer Center, Kyushu University 91
Hakozaki, Fukuoka, 812 Japan

There is no practical automatic indexing method other than stop word deletion for large amount of documents. Consequently, the retrieval power for subject and the size of inverted file for index in a practical information retrieval system depend on the set of the stop words called a negative dictionary. In this paper, first we measured the retrieval power and the relative size of inverted file of the existing negative dictionaries to show their limits. Second the authors proposed a method for constructing negative dictionaries. By using this method we can control the size of inverted file as will. As the size decreases, the retrieval power gradually drops and never falls catastrophically. Last the authors built a negative dictionary to eliminate the defects of the dictionaries made by the above method. This negative dictionary is superior to the best of the existing dictionaries in the relative size of inverted file and the same in the retrieval power.

1. まえがき

書誌的文献(bibliographic items)に対し，計算機によって，その文書(documents)の内容を分析し，各文献に内容同定語(content identifiers)を付与することを自動索引(automatic indexing)という[1]．計算機による情報検索(information retrieval)を目的とする自動索引では，内容同定語には索引語(index terms)やキーワード(keywords)などと呼ばれる単一語(single terms)を語いとすることが一般的である．この場合，句による文書指定には，文献探索時に単一語から句への事後結合(post-coordination)を行う．索引語を語いとする言語を索引言語(index language)と呼ぶ．

最も単純な自動索引は，文書テキストあるいは抄録，標題に現われる語をすべて索引語とすることである．しかし，文章の大半を占める機能語(function words)には文書内容の識別力がほとんどないので，これらを除去する必要がある．内容識別力をもたない語を，ここでは不要語(stop words, fluff words)と呼ぶ．不要語の辞書を，否定辞書(negative dictionary)あるいは不要表(stop list)などと呼んでいる．否定辞書にない単一語を索引語とする自動索引を不要語除去法と呼ぶことにする．現在，大量の書誌的文献に対する自動索引には，不要語除去法以外に実用に耐える方式がないので，実用情報検索システムでは，この方式を一般的に採用している．

さて，大量文献に対する情報検索システムでは，単一語あるいは事後結合句による探索を高速にするため，索引語についての転置索引(inverted index)がディスクなどの二次記憶上に作られる．ここでは，これを索引語転置ファイルと呼ぶ．一般に，否定辞書を用いずに全出現単語を索引言語の語いとすると，索引語転置ファイルの大きさは文書ファイル本体を超えるほど大きくなってしまうので，索引言語の文書内容記述力の低下をできるだけ押え，かつ転置ファイルをできるだけ小さくするように否定辞書を作る必要がある．

これまで，否定辞書は不要語と考えられるものを随意的に選択することによって作られており，また索引言語の内容記述力の数量的評価がほとんどなされていなかったこともあって，検索能力と転置ファイルの大きさの関連を調べた研究は皆無に等しい．

本稿では，まずINSPECテープ[2]のように，非統制索引語(uncontrolled index terms)を人手によって事前結合した句(pre-coordinated clauses)をもつ書誌的文献に関して索引言語の検索能力と索引語転置ファイルの相対領域量を示す測度を定義する．次いで，これによって幾つかの代表的な既存否定辞書をINSPECテープによって評価し，それらの限界を示す．この限界を超える否定辞書として，著者らは非統制事前結合索引句の語いを利用した統計的性質に基づく否定辞書を提唱した[3]．この方法では検索能力と索引語転置ファイルの大きさの間の制御が可能であり，INSPECテープに対しては検索能力を既存否定辞書によるものと同一水準にした場合，転置ファイルを最良の既存否定辞書より20%程度小さくできる．しかし，この方法による否定辞書では一般に否定辞書が大きく，またINSPECテープのような非統制事前結合索引句をもたない文献集合に対しては否定辞書を用意することができないなどの欠点がある．本稿では，上記の方法によって得られた否定辞書をもとに約1,700語からなる否定辞書を作成した．この否定辞書は，すべての科学技術文書に有効であると考えられる．

既存の否定辞書を含め，これらの否定辞書を，本稿で定義した二つの測度に基づいて比較･評価する．

2. 否定辞書の評価法

本稿では，否定辞書をそれが作る索引言語の検索能力と索引語転置ファイルの大きさによって評価する．この評価のために，1973年から1982年までの10年間に配布されたINSPECテープを用いる．INSPECテープは，英国IEE(the Institution of Electrical Engineers)が集積･配布している代表的な科学技術書誌的文献である．

INSPECテープの書誌項目の一つに自由索引句(free-indexing terms)がある．これは，非統制索引語を人手によって事前結合した句(pre-coordinated clauses)である．ただし，各文献に付与された自由索引句だけでその文献の内容を同定することができないので，この項目は主題検索の面では補助的な役割しか果たさない．本稿では，自由索引句の語いを検索能力の評価に用いる．

2.1 評価のための文献集合

本稿では，単一語あるいは単に語というのは英字と数字を字母とする(非空な)文字列のことをいう．また，英字は大文字と小文字の区別をせず，すべて大文字と考える．したがって，文書テキスト中の

'CAN'T' は'CAN' と'T' に, '0.12' は'0' と'12' に, 'SYSTEM/370' は'SYSTEM' と'370' に分離した語とみなす.

書誌的文献を単に文献と呼ぶ. 文献集合Xが与えられたときに, 標題, 抄録, 自由索引句における語wの生起頻度をそれぞれ$f_t^X(w)$, $f_a^X(w)$, $f_i^X(w)$で, 相対頻度(確率)をそれぞれ$g_t^X(w)$, $g_a^X(w)$, $g_i^X(w)$で表わす. 語の生起頻度は, 各文献の同一項目内に語が複数回生起する場合は, その回数だけ加算する. また, 文献集合Xにおいて標題, 抄録, 自由索引句に現われる語集合をそれぞれW_t^X, W_a^X, W_i^Xで表わす. 語集合Wが与えられたときに, Xの標題におけるWの生起頻度$f_t^X(W)$は,

$$f_t^X(W)=\Sigma_{w\in W} f_t^X(w)$$

によって定義する. 同様にして, $f_a^X(W)$, $f_i^X(W)$を定義する. また同様に, 語集合Wについての相対頻度$g_t^X(W)$, $g_a^X(W)$, $g_i^X(W)$を定義する. 生起頻度, 相対生起頻度, 語集合のそれぞれにおいて項目を特に指定しないときは下添字を省略し, 文献集合を特に指定しないときは上添字を省略する. 語wについて, $|w|$はwの長さを表わす. Wが語の集合の場合, $|W|$は集合の大きさを示す.

ここでは, 10年分のINSPECテープから, 分野による違いをみるために, 文献集合A, B, Cを作った. 集合A, B, Cは, 互いに素ではなく, 20%の文献が複数の集合に属している. 表1に文献集合A, B, Cの語集合の大きさと語の生起頻度を示す.

表1 INSPECテープ(1973~1982年)からの三つの文献集合の語集合の大きさと語の生起頻度

	A	B	C
No. items	973,735	500,741	323,336
Field	Physics	Electrical Eng. and Electronics	Control Eng. and Computer Sci.
Title			
No. words $\|W_t\|$	103,572	65,029	56,139
Freq. $f(W_t)$	11,012,493	4,747,226	2,842,323
Abstract			
No. words $\|W_a\|$	274,185	154,231	127,836
Freq. $f(W_a)$	87,354,577	37,606,323	23,391,467
Free-indexing terms			
No. words $\|W_i\|$	166,549	89,319	77,569
Freq. $f(W_i)$	19,763,849	7,841,776	4,340,930
Total			
No. words $\|W_r\|$	303,411	169,433	139,460
Freq. $f(W_r)$	118,130,919	50,195,325	30,574,720

2.2 索引語転置ファイルの大きさ

不要語除去法では, 各文献に対して, 通常, 標題と抄録から取られた索引語が付与される. 文書に抄録がある場合, 文書テキストから索引語を選んだとしても, 文書内容の識別力がほとんど向上しないことが知られている[1]. そこで, 本稿では, 索引語は標題と抄録から取ることを前提として議論を進める.

索引語転置ファイルの要素は, 基本的には, 索引語とその語を標題あるいは抄録に含む文書参照番号リストとの対である. 隣接演算(adjacency operations)を転置ファイルによって実行する場合は, 文書参照番号にその文献の標題と抄録における索引語の生起位置情報が加わる. ここでは, 前者の転置ファイルを基本型と呼び, 後者を隣接演算型と呼ぶ.

既存の大部分の実用情報検索システムでは, 隣接演算型を採用している. 隣接演算型の索引語転置ファイルの大きさは, 直感的に予想できるように, 標題と抄録に現われる全索引語の生起頻度に比例するものと考えられる. このことは, 実験によっても確かめられている[3].

そこで次のような仮説をおく.

[索引語転置ファイルの大きさについての仮説]

いま, Xを文献集合, W^XをXの標題と抄録に現われる語の集合($W_t^X \cup W_a^X$), Sを否定辞書, すなわち不要語の集合とすると, 索引語転置ファイルの大きさD_S^Xは,

$$D_S^X \propto f_t^X(W^X\text{-}S)+f_a^X(W^X\text{-}S),$$

あるいは,

$$D_S^X \propto g_t^X(W^X\text{-}S)+g_a^X(W^X\text{-}S).$$

そこで, 次式で定義されるd_S^Xを索引語転置ファイルの相対領域量(relative size of inverted file for index terms)と呼ぶ.

$$d_S^X=\frac{f_t^X(W^X\text{-}S)+f_a^X(W^X\text{-}S)}{f_t^X(W^X)+f_a^X(W^X)}$$

2.3 検索語

文献探索時に, 検索者が指定する単一語を検索語と呼ぶことにする. 検索者が不特定多数の場合, 検索語集合とその要素の相対頻度とをあらかじめ知ることはできない. そこで, 本稿ではこれを近似する手段として自由索引句を利用する.

自由索引句は, 索引語が非統制であるだけではなく, 語の用法にも統一性がなく, 文献に対して句を付与する際の基準も索引付与者(indexers)の主観に左右されている. このことを示す例として, 表2に'devide and conquer'を事後結合によって求めた

表2 抄録における表現と自由索引句における表現

Abstract	Free-indexing Terms
a divide-and-conquer strategy	divide and conquer strategy
a similar divide-and-conquer technique	divide and conquer technique
the 'divide and conquer' principle	(none)
this divide-and-conquer strategy	(none)
the 'divide and conquer' technique	(none)
a divide-and-conquer strategy	divide and conquer
divide and conquer	divide and conquer
the 'divide and conquer' principle	(none)
the divide-and-conquer method	(none)
a divide and conquer technique	divide and conquer
the divide and conquer paradigm	divide and conquer paradigm

11文献について，抄録と自由索引句の表現を対比した．このために自由索引句の単語は，検索語に非常に近いものであるとも考えられる．そこで，ここでは次のような仮定をおく．

[検索語の仮定]

文献集合Xに対する検索語の集合をR^Xとすると，

$$R^X = W_i^X.$$

$w \in W_i^X$がXに対して検索語として使われる相対頻度を$p^X(w)$とすると，

$$p^X(w) = g_i^X(w).$$

この仮定からは，最良の索引語選択法はW_i^Xの単語をすべて索引語とすることになる．しかし，この選択法は，自由索引句用法の不統一性によって必ずしも実用的ではない．例えば，'THE'は文献集合A, B, Cのいずれにおいても抄録における生起頻度は最大で，自由索引句における生起頻度は非常に小さい．これは，自由索引句では定冠詞を用いないことを原則にしているが，定冠詞をもつ句が自由索引句項目に誤って若干数含まれていることによる．この選択法では，通常不要語とする'THE'が索引語となってしまう．

2.4 検索言語の検索能力

2.2節と同様，文献集合Xの索引語集合をW^X-Sとする．W^X-Sを語いとする索引言語のXに対する検索能力(retrieval power)P_S^Xを，

$$P_S^X = 1 - \frac{g_i^X(S)}{g_i^X(W^X)}$$

で定義する．この定義では，$P_S^X=1$のとき検索能力が最も大きいと考える．例えば，Sが空集合のときは，$P_S^X=1$である．そして，P_S^Xの値が小さいほど，検索能力が小さいとみる．最小は，$S=W_i^X$のときで，この場合$P_S^X \leqq 0$となる．$S \subseteq W^X$とすると，P_S^Xは，

$$P_S^X = (g_i^X(W^X) - g_i^X(S))/g_i^X(W^X)$$
$$= g_i^X(W^X - S)/g_i^X(W^X)$$

であるので，検索語の仮定から，P_S^Xはほぼ$w \in (W^X - S) \cap W_i^X$によって文献が求まる確率と$w \in W^X \cap W_i^X$によって求まる確率の比と考えることができる．つまり，$g_i^X(S)/g_i^X(W^X)$を不要語の影響による検索能力の低下とみるのである．$w \in W_i^X - W^X$の場合，wによる検索でも文献は求まらないが，ここではこのことと索引言語の能力とを切り離して考えることにする．

3. 既存方式の評価

この章では，これまで提案されている二，三の否定辞書について，前章で定義した索引語転置ファイルの相対領域量と検索能力を用いて評価を行う．

3.1 否定辞書最小法

できるだけ小さい数の不要語で，転置ファイルの大きさを小さくする方法は，いうまでもなく生起頻度をキーにしてwを降順にソートし，上位の単語を不要語とする方法である．ジップの法則(Zipf's law)[4]を仮定すると，r番目に生起頻度の高い単語の相対頻度$p(r)$は，

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

これから，$\sum_{r=1}^{83} 0.1/r \simeq 1/2$となるので最も生起頻度の高い83単語がテキストの半分を占めることになる．INSPECの抄録についても，高頻度単語は同様の性質があるので，生起頻度の高い128個の不要語によって抄録部分の転置ファイルの大きさは半分になる．この方法の欠点は，使用確率の高い検索語が不要語になることである．文献集合Cでは，'SYSTEM', 'CONTROL', 'DATA', 'COMPUTER'の単語は抄録における生起頻度が大きく，順位はそれぞれ11, 20, 21, 23位であるが，自由索引句における生起頻度の順位はもっと高く，それぞれ1, 2, 5, 3位である．不要語集合の大きさを25程度にとっても，検索語の仮定から使用確率の非常に高いこれら4個の検索語が不要語になってしまう．これらの単語は，通常，集合Cの検索語として単独で使われることはないと思われるが，事後結合方式では不要語にできない．例えば，'JOSEPHSON COMPUTER'を探索できるようにするためには，'COMPUTER'も索引語でなければならない．

この欠点は，不要語数を10以下にすれば無くなる．ジップの法則と転置ファイルの大きさについての

仮定から，不要語数が7の場合は，転置ファイルの抄録部分の大きさは25%減少することになる．通常の英文でも最も生起頻度の高い9個の単語がテキストの25%を占めるという報告があり[5]，DIALOGの不要語も9個である．次節でDIALOGの否定辞書について評価する．

3.2 DIALOGの否定辞書

DIALOGの否定辞書S_1を表3に示す．表4にS_1を否

表3 DIALOGの否定辞書S_1

AN
AND
BY
FOR
FROM
OF
THE
TO
WITH

定辞書とした場合の文献集合A, B, Cに対する索引語転置ファイルの相対領域量と検索能力を示した．表4

表4 S_1による索引語転設置ファイルの相対領域量と検索能力

	A	B	C
Relative Size of Inverted File for Index Terms	0.786	0.788	0.785
Retrieval Power	0.987	0.989	0.988

から，転置ファイルの大きさは，25%も減少せず，21~22%程度であることがわかる．検索能力の低下は1.1~1.3%であった．

3.3 品詞情報ならびに機能語による否定辞書

冠詞や前置詞などの品詞をもつ単語を不要語とする方法は，代表的な不要語選択法の一つである．冠詞，前置詞，接続詞，代名詞，助動詞の品詞をもつ223語からなる単語集合をS_2とし，その要素を表5に星印を付して示す．S_2は文語を含んでいるが，口語や古語は含んでいない．先にあげた品詞以外で，その品詞をもつすべての単語を不要語にできるものとしては，感嘆詞くらいであり，その意味では，S_2は科学技術文献に対して品詞情報のみによって不要語を選択したものとしては，要素数が最大に近い．

この数に近い否定辞書としては，英語の共通機能語として知られている文献6の不要表がある．この不要表から'EG'と'IE'を除いた248個の不要語からなる否定辞書をS_3とし，その要素を表5に剣印を付して示した．表6にS_2による索引語転置ファイルの相対領域量と検索能力を，表7にS_3によるものを示した．表6と表7から，S_3の方がS_2よりもわずかではあるが，相対領域量は良く，検索能力は悪いことがわかる．しかし，その差はわずかであり，200~250個の機能語を不要語とした場合の，相対領域量と検索能力の値はこの程度であろう．これらの否定辞書を使うと転置ファイルの大きさは，37~40%減少するが，検索能力は，S_1よりせいぜい1%低下するだけでS_1と大差なく，これらの否定辞書が有効であることがわかる．しかし，転置ファイルをこれ以上小さくするには，機能語と思われる単語を大量に否定辞書に追加しなければならないが，随意的な不要語選択でこのことを行うことは困難である．なお，S_1と異なり，S_2とS_3を使った場合，相対領域量と検索能力の両方に関して，集合Cが最も良く，Aが最も悪い．

4 自由索引句の語いを利用した統計的性質に基づく否定辞書

集合Cでは，抄録と自由索引句とにおける単語'AM'の生起頻度は，それぞれ305と148である．一方，単語'IS'ではこれらの度数は，それぞれ487,263と37である．このことから'AM'はbe動詞としてはあまり出現せず，'振幅増幅'などを意味する略語として現われており，一方'IS'はほとんどはbe動詞として使われていると判断することができよう．そこで，'IS'を不要語として選択し，'AM'を索引語とすることが考えられる．つまり，$f_i(w)/f_a(w)$の値を不要語選択の基準に使う方式である．

いま，否定辞書S_θを次のように定義する．

$$S_\theta=\{w|r(w)<\theta \wedge w\in W_t\cup W_a\},$$

ここで，$r(w)=\begin{cases} f_i(w)/f_a(w) & f_a(w)\neq 0\text{のとき;} \\ 1 & f_a(w)=0\text{のとき.}\end{cases}$

S_θによる転置ファイルの相対領域量と検索能力を図1と図2に示す．

図3に，

$$|S_\theta|\,/\,|W_t\cup W_a|$$

を示した．この値を否定辞書の相対大きさ(relative size of negative dictionary)と呼ぶ．図1からわかるようにこの方法では，θを変化させることによって

表5 品詞情報による否定辞書S_2と共通機能語による否定辞書S_3

*† A	† CO	*† ITSELF	† OWN	* TILL
* ABOARD	* CONCERNING	* LACKING	* PAST	*† TO
*† ABOUT	* CONSIDERING	† LAST	* PENDING	† TOGETHER
*† ABOVE	*† COULD	† LATTER	*† PER	† TOO
* ABREAST	* DARE	† LATTERLY	† PERHAPS	* TOUCHING
*† ACROSS	* DESPITE	† LEAST	* PLUS	*† TOWARD
*† AFTER	* DID	*† LESS	* PROVIDED	† TOWARDS
† AFTERWARDS	* DO	* LEST	* PROVIDING	*† UNDER
† AGAIN	* DOES	* LIKE	† RATHER	* UNDERNEATH
*† AGAINST	*† DOWN	† LTD	* REGARDING	* UNLESS
* ALBEIT	*† DURING	† MANY	* RESPECTING	* UNLIKE
*† ALL	*† EACH	*† MAY	* ROUND	*† UNTIL
† ALMOST	*† EITHER	*† ME	*† SAME	*† UP
† ALONE	† ELSE	† MEANWHILE	* SAVE	*† UPON
*† ALONG	† ELSEWHERE	* MIDST	* SAVING	*† US
* ALONGSIDE	† ENOUGH	*† MIGHT	† SEEM	* VERSUS
† ALREADY	† ETC	* MINE	† SEEMED	† VERY
*† ALSO	† EVEN	* MINUS	† SEEMING	*† VIA
*† ALTHOUGH	† EVER	† MORE	† SEEMS	* WANTING
† ALWAYS	† EVERY	† MOREOVER	* SELF	*† WAS
* AM	* EVERYBODY	† MOST	*† SEVERAL	*† WE
* AMID	*† EVERYONE	† MOSTLY	* SHALL	† WELL
* AMIDST	*† EVERYTHING	† MUCH	*† SHE	*† WERE
*† AMONG	† EVERYWHERE	*† MUST	*† SHOULD	*† WHAT
*† AMONGST	*† EXCEPT	*† MY	*† SINCE	*† WHATEVER
*† AN	* EXCEPTING	*† MYSELF	*† SO	* WHATSOEVER
*† AND	* FAILING	† NAMELY	*† SOME	*† WHEN
*† ANOTHER	*† FEW	* NEAR	* SOMEBODY	*† WHENCE
*† ANY	† FIRST	* NEED	† SOMEHOW	*† WHENEVER
* ANYBODY	*† FOR	*† NEITHER	*† SOMEONE	*† WHERE
† ANYHOW	† FORMER	† NEVER	*† SOMETHING	† WHEREAFTER
*† ANYONE	† FORMERLY	† NEVERTHELESS	† SOMETIME	*† WHEREAS
*† ANYTHING	*† FROM	*† NEXT	† SOMETIMES	† WHEREBY
† ANYWHERE	† FURTHER	† NO	† SOMEWHERE	* WHEREFORE
*† ARE	*† HAD	*† NOBODY	† STILL	† WHEREIN
*† AROUND	*† HAS	*† NONE	*† SUCH	* WHERESOEVER
*† AS	*† HAVE	† NOONE	* SUPPOSING	† WHEREUPON
* ASTRIDE	*† HE	*† NOR	*† THAN	*† WHEREVER
*† AT	† HENCE	† NOT	*† THAT	*† WHETHER
* ATHWART	*† HER	*† NOTHING	*† THE	*† WHICH
* ATOP	† HERE	* NOTWITHSTANDING	* THEE	* WHICHEVER
* BAR	† HEREAFTER	*† NOW	*† THEIR	*† WHILE
*† BE	† HEREBY	† NOWHERE	* THEIRS	† WHITHER
† BECAME	† HEREIN	*† OF	*† THEM	*† WHO
*† BECAUSE	† HEREUPON	*† OFF	*† THEMSELVES	*† WHOEVER
† BECOME	*† HERS	† OFTEN	† THEN	† WHOLE
† BECOMES	*† HERSELF	*† ON	† THENCE	*† WHOM
† BECOMING	*† HIM	*† ONCE	† THERE	*† WHOSE
*† BEEN	*† HIMSELF	*† ONE	† THEREAFTER	† WHY
*† BEFORE	*† HIS	* ONESELF	† THEREBY	*† WILL
† BEFOREHAND	† HOW	*† ONLY	† THEREFORE	*† WITH
*† BEHIND	*† HOWEVER	*† ONTO	† THEREIN	*† WITHIN
*† BEING	*† I	* OPPOSITE	† THEREUPON	*† WITHOUT
*† BELOW	*† IF	*† OR	*† THESE	*† WOULD
* BENEATH	* IMMEDIATELY	*† OTHER	*† THEY	* YE
*† BESIDE	*† IN	† OTHERS	*† THIS	*† YET
*† BESIDES	† INC	† OTHERWISE	*† THOSE	*† YOU
*† BETWEEN	† INDEED	* OUGHT	*† THOUGH	*† YOUR
*† BEYOND	* INSIDE	*† OUR	*† THROUGH	*† YOURS
*† BOTH	* INSTANTLY	*† OURS	*† THROUGHOUT	*† YOURSELF
*† BUT	*† INTO	*† OURSELVES	† THRU	† YOURSELVES
*† BY	*† IS	*† OUT	† THUS	
*† CAN	*† IT	* OUTSIDE	* TILL	
*† CANNOT	*† ITS	*† OVER	*† TO	

星印の単語はS_2属し，剣印の単語はS_3に属している．

転置ファイルを望みの大きさにすることができる．θが大きくなるにつれ，検索能力は図2のように連続的になだらかに低下し，破局的に悪化することはない．

さて，図1はθが0から少し増加すると転置ファイルの大きさが急激に減少し，あとはθの増加に伴ってなだらかに減少する分岐点があることを示している．この$|S_\theta|$の減少は，'OF'が不要語になったため

表6 S_2による索引語転設置ファイルの相対領域量と検索能力

	A	B	C
Relative Size of Inverted File for Index Terms	0.626	0.622	0.609
Retrieval Power	0.978	0.980	0.982

表7 S_3による索引語転設置ファイルの相対領域量と検索能力

	A	B	C
Relative Size of Inverted File for Index Terms	0.619	0.615	0.599
Retrieval Power	0.976	0.980	0.981

に生じているので，θ=r('OF')の点をOF点と呼ぶことにする．文献集合A, B, CのOF点の値は，それぞれ0.0299, 0.0236, 0.0245である．θがこの付近の値であれば，検索能力の低下は非常に小さく，この点

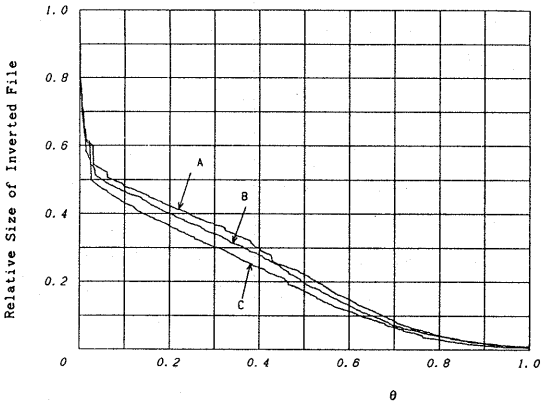

図1 索引語転置ファイルの相対領域量

は，索引語を選択するθの値として一つの規準になる．

表8に，こうして選んだθに対する転置ファイルの相対領域量，検索能力，否定辞書の相対大きさ，不要語数を示した．表8から，OF点に基づく否定辞書では，転置ファイルの大きさを46~50%減少させることができることがわかる．検索能力は，集合Cを除いて，S_2とS_3より良く，S_1とほとんど同じである．集合Cは，転置ファイルの縮小率がAとBより小さいが，検索能力の低下も一番大きい．この原因の一つは，集合Cでは'TO'は不要語であるが，AとBでは不要語ではないということにある．この理由は，

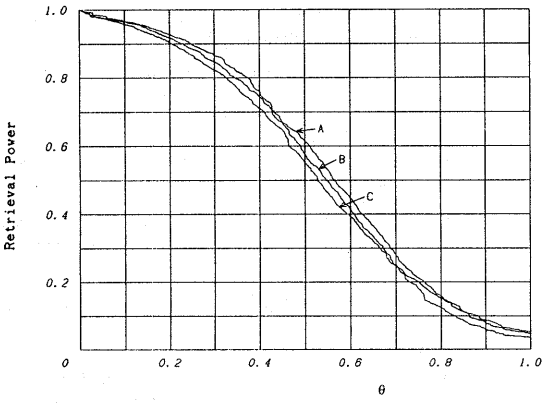

図2 検索能力

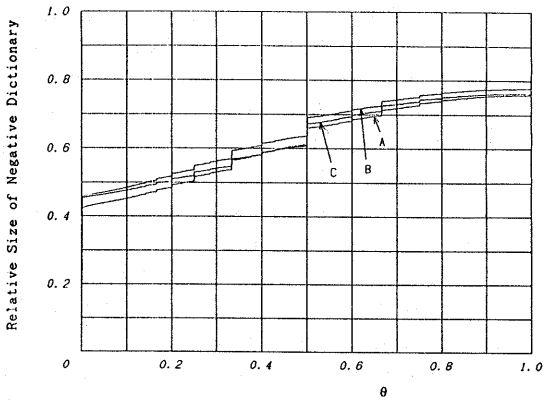

図3 否定辞書の相対大きさ

表8 OF点付近における諸元

	A	B	C
Threshold θ	0.0299	0.0236	0.0245
Relative Size of Inverted File for Index Terms	0.544	0.544	0.502
Retrieval Power	0.986	0.985	0.982
Relative Size of Negative Dictionary	0.461	0.464	0.432
No. of Stopwords	132,610	75,959	58,686

集合AとBの自由索引句には，'50KHz to 100KHz'のように，範囲を示すための'TO'の出現頻度が高く，r('TO')はr('OF')よりかなり大きいためである．そこで，'TO'だけを特別扱いをして不要語とした場合の表8の修正値を表9に示した．

表9 不要語に'TO'を加えた場合の索引語転置ファイルの相対領域量と検索能力

	A	B
Relative Size of Inverted File for Index Terms	0.528	0.527
Retrieval Power	0.981	0.982

OF点に基づく否定辞書は，九州大学大型計算機センター(九大センター)におけるINSPECデータベースの検索サービスに1980年夏から1986年3月まで使用した[3]．このサービスは，1979年秋にFAIRS-I[7]を用いて始まった．当初，3.3節の線に沿った453語の否定辞書を使用していたが，随意的に選んだ機能語を不要語とする方式としては不要語数が大き過ぎ，検索者から不評であり，しかも索引語転置ファイルの領域も3.3節の結果とほとんど差がなかった．いずれにしても，当時の九大センターのディスク事情に対し最初に採用した否定辞書では転置ファイルが大き過ぎ，OF点に基づく否定辞書の開発によって検索サービスの継続が可能になったのである．

このサービスにおいて，INSPECデータベースは，集合A, B, Cの各分野に対応して，INSPEC-A, 同-B, 同-Cが構築された．INSPEC-Bと同-Cは，θとしてOF点値を採用することができたが，転置ファイル用二次記憶領域の不足でINSPEC-Aはθ=0.379とOF点の10倍程度にしなければならなかった．それでも，検索者から索引言語についての不満はほとんど聞かれなかった．INSPEC-AのθがOF点値になったのは，1984年1月にこのデータベースの検索システムがFAIRS-Iからディスク使用効率の高いAIR[8]に替ってからである．このように，この否定辞書を5年半の間，実用に供してきたが，索引言語の文書内容識別力が問題になったことはほとんどない．したがって，OF点に基づく方式は，識別力と転置ファイルの大きさの均衡をとることができ，転置ファイルのための二次記憶領域が十分にない場合にも，それなりの検索能力を与えることができるものとして実用的価値が高いが，次のような問題点がある．

1) 表8または図3からわかるように全出現単語の約半数(43~46%)が不要語となるため，否定辞書が大きい，
2) INSPECテープの自由索引句を使用しているため，このような非統制索引句をもたない文献集合に対しては否定辞書を作ることができない．

(1)の量の問題は，(2)に比べて本質的ではない．量の問題の一つは，否定辞書が大きいと，索引語か不要語かの判定のための計算量が増加することである．例えば，FAIRS-Iでは非空の否定辞書Sがあるとき，索引語転置ファイルの構築時間は，$O(\log|S|)$である．しかし，この問題は辞書の構成技法によって克服することができる．実際，上記のFAIRS-Iによる検索サービスでは，否定辞書の大きさ$|S|$が数千から数十万のとき，$O(1)$の方法を開発した[9]．しかしながら，否定辞書が大きいことは，辞書の維持・管理面からやはり好ましいことではない．次章で，これらの問題に対する一つの解答を示す．

5. 共通否定辞書

文献A, B, Cに対して，表8に示したOF点に基づく否定辞書を，それぞれS^A, S^B, S^Cとする．表8に示したように，$|S^A|=132{,}610$, $|S^B|=75{,}959$, $|S^C|=58{,}686$.

いま，これらの共通部分

$$S_c=S^A\cap S^B\cap S^C$$

を考える．すると，この大きさ$|S_c|$は11,667である．次にS_cの要素のうち文献集合Xの標題と抄録における最も頻度の高いr個の語によって否定辞書S_r^Xを作る．図4は，順位rとS_r^A, S_r^B, S_r^Cによる索引語転置ファイルの相対領域量の関係を示したものである．また，図5にrと検索能力の関係を示す．

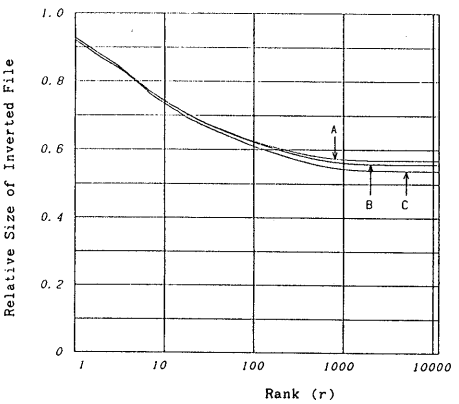

図4 S_r^Xによる索引語転置ファイルの相対領域量

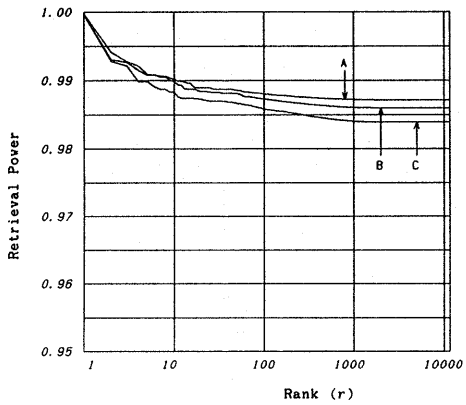

図5 S_rXによる検索能力

図4から，集合A, B, Cとも，rが2,000を超えると，もはや相対領域量は減少しなくなることがわかる．最小値は，集合A, B, Cについて，それぞれ0.568, 0.557, 0.538であるので，転置ファイルの大きさはOF点によるものより，それぞれ4.4%, 2.4%, 7.2%増加することになる．r=2,000のときは，これより0.2%程度増加する．図5も，同様に，rが2,000を超えると検索能力の低下がほとんどなくなることを示している．最小値は，集合A, B, Cでそれぞれ0.9872, 0.9860, 0.9839である．したがって，検索能力はOF点によるものよりわずかではあるが，0.1~0.2%程度向上する．この場合は，r=2,000のときと最小値との差はほとんどない．

さて，共通否定辞書として

$$S_r=(S_r^A\cap S_r^B\cap S_r^C)\cup\{\text{'TO'}\}$$

を考えてみよう．図6にrとS_rによる索引語転置ファイルの相対領域量の関係を示し，図7にrと検索能力

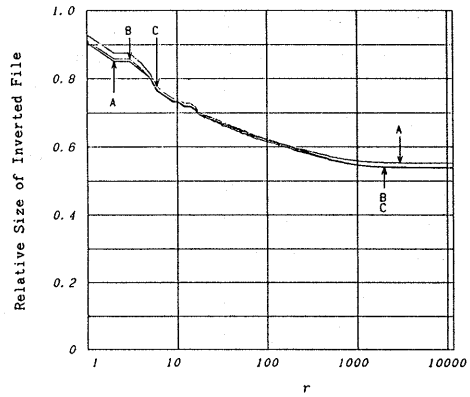

図6 S_rによる索引語転置ファイルの相対領域量

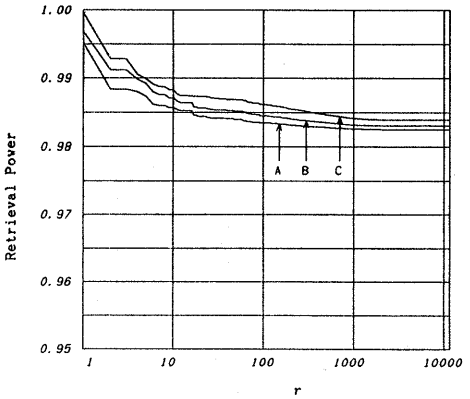

図7 S_rによる検索能力

の関係を示した．図6と図7も，図4と図5同様，rが2,000を超えると相対領域量と検索能力の低下が止まることを示している．相対領域量の最小値は，集合A, B, Cで，それぞれ0.552, 0.540, 0.538であり，集合AとBは，図4の最小値に比べ約3%減少している．これは，'TO'を不要語としたための効果である．r>5のとき，S_r^Cは'TO'を含んでいるので最小値は変わらない．r=2,000のときは，最小値より0.3~0.4%大きくなる．r=2,000のときで図5と図7の値を比較すると，集合AとBについては図7の方が3%程度領域量が小さくなっているが，集合Cでは0.2%程度大きくなる．同様に，検索能力についても図7の最小値は，集合AとBに関し，それぞれ0.9825, 0.9830であり，わずかながら0.3~0.5%程度減少する．集合Cについては，当然最小値に変化はない．この場合も，r=2,000のときと最小値との差はほとんどない．

さて，予想されるように，一般に$|S_r|\leq r$である．そこで，$|S_r|/r$は，S_r^A, S_r^B, S_r^Cの共通度を示す一つの尺度と考えることができる．図8に，rと$|S_r|/r$の関係を示す．r=2,000付近では，$|S_r|/r$の値は比較的変化がなく平坦であるものの，r=2,000で極大となる．そこで，S_{2000}を共通否定辞書に選ぶことにする．この辞書は，科学技術文献に対する共通の否定辞書と考えることができる．この辞書の大きさ$|S_{2000}|$は，1,667である．表10に，共通辞書による転置ファイルの相対領域量と検索能力を示す．これから共通辞書では，文献集合による性能差が小さいことがわかる．前章のOF点に基づく否定辞書とこの共通辞書を比較すると，ファイル領域量の大きさでは，集合AとCに対しては共通辞書の方が大きく，それぞれ1.8%,

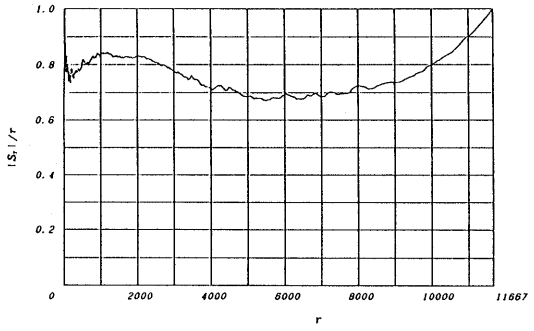

図8 共通不要語数の割合$|S_r|/r$

表10 共通否定辞書による索引語転置ファイルの相対領域量と検索能力

	A	B	C
Relative Size of Inverted File for Index Terms	0.554	0.541	0.540
Retrieval Power	0.983	0.983	0.984

7.5%増加するが，集合Bに対しては0.5%減少する．検索能力は，集合AとBについては，共通辞書の方が0.2~0.3%低下するが，集合Cに対しては0.2%増加する．

更に，この共通否定辞書と3.2節のDIALOGの否定辞書および3.3節の文献6の不要表とを比較してみると，まずDIALOGの否定辞書に対しては，検索能力はわずか0.5%程度劣るだけで，ほとんど差がない．そして，転置ファイルの大きさを，それより30%小さくすることができる．次に，文献6の不要表に対しては，検索能力はやはり差があまりないが，0.2~0.6%向上し，転置ファイルの大きさは，それより10~20%小さい．

九大センターでは，1986年4月にINSPECデータベースの検索サービスにおける否定辞書を，OF点に基づく否定辞書からこの共通否定辞書に変更した．

6. むすび

索引語転置ファイル生成における不要語の選択法に関し，従来からの方法では転置ファイルのディスク領域量改善の点で限界があることをINSPECテープの解析を通じて示し，自由索引句を利用した単語生起頻度の統計的性質に基づく方法が有効であることを示した．この方法では，転置ファイルの大きさを自在に小さくすることができ，これによる検索能力の低下は連続的であり，破局的に悪化することはない．また，この方式は5年半にわたって実用に供されていて，実用的有効性が立証されている．ただし，この方式には，全出現語の約半数が不要語になるため，否定辞書が大きく，かつ自由索引句のような非統制索引句をもたない書誌的文献に対して適用できないなどの欠点がある．しかし，これらの欠点はこの方式をもとに作成した1,667語の共通否定辞書によって取り除くことができる．この否定辞書は，科学技術文献に共通の不要語集合あるいは機能語集合と考えることができよう．

参考文献

1) Salton, G. and McGill, M.J. : Introduction to Modern Information Retrieval, p.448, McGraw-Hill, New York(1983).
2) 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).
3) 松尾文碩，二村祥一，高木利久，吉田将: INSPECデータベース転置ファイル生成における不要語選択法，九大工学集報，第54巻 第2号，pp.99-105(1981).
4) Shannon, C.E. : Prediction and Entropy of Printed English, Bell Syst. Tech. J., Vol.30, No.1, pp.50-65(1951).
5) Dewey, G. : The Relative Frequency of English Speech Sounds, p.187, Harvard University Press(1923).
6) Van Rijsbergen, C.J. : Information Retrieval (2nd Ed.), p.208, Butterworths, London(1979).
7) 計算機マニュアル FACOM OS Ⅳ FAIRS-I解説書，富士通(株) (1978).
8) Matsuo, F., Futamura, S., and Shinohara, T. : Efficient Storage and Retrieval of Very Large Document Databases, Proc. 2nd Int'l Conf. on Data Engineering, pp.456-463(1986).
9) 松尾文碩，二村祥一，高木利久，吉田将: 高速検索のための単語辞書索引の一構成法，第54巻 第3号，pp.183-187(1981).