

英文科学技術文献情報に対する不要語除去法による自動索引†

二村 祥一†† 松尾 文碩††

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、この方法に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。

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いますので、索引言語の文書内容記述力の低下をできるだけ抑え、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値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

† Automatic Indexing by Stop Word Removal on Scientific and Technical Documents Written in English by SHOUICHI FUTAMURA and FUMIHIRO MATSUO (Computer Center, Kyushu University).

†† 九州大学大型計算機センター

対しては検索能力を既存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 (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 $f_i^X(w)$, $f_o^X(w)$, $f_f^X(w)$ で、相対頻度 (確率) をそれぞれ $q_i^X(w)$, $q_o^X(w)$, $q_f^X(w)$ で表す。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、自由索引句に現れる語集合をそれぞれ W_i^X , W_o^X , W_f^X で表す。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 $f_i^X(W)$ は、

$$f_i^X(W) = \sum_{w \in W} f_i^X(w)$$

によって定義する。同様に、 $f_o^X(W)$ と $f_f^X(W)$ を定義する。また同様に、語集合 W についての相対頻度 $q_i^X(W)$, $q_o^X(W)$, $q_f^X(W)$ 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、分野による違いを見るために、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

表 1 INSPEC テープ (1973~1982 年) からの三つの文献集合の検索項目
Table 1 Three sets of bibliographic items from INSPEC tapes issued from 1973 to 1982.

		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_i $	103, 572	65, 029	56, 139
	Freq. $f(W_i)$	11, 012, 493	4, 747, 226	2, 842, 323
Abstract	No. words $ W_o $	274, 185	154, 231	127, 836
	Freq. $f(W_o)$	87, 354, 577	37, 606, 323	23, 391, 467
Free-indexing terms	No. words $ W_f $	166, 549	89, 319	77, 569
	Freq. $f(W_f)$	19, 763, 849	7, 841, 776	4, 340, 930
Total	No. words	303, 411	169, 433	139, 460
	Freq.	118, 130, 919	50, 195, 325	30, 574, 720

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。

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、索引語を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献の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、文献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²⁾。

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 の標題と抄録に現れる語の集合 ($W_t^X \cup W_c^X$)、 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_t^X は、

$$D_t^X \propto f_t^X(W^X - S) + f_c^X(W^X - S),$$

あるいは、

$$D_t^X \propto g_t^X(W^X - S) + g_c^X(W^X - S).$$

そこで、次式で定義される d_t^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。

$$d_t^X = (f_t^X(W^X - S) + f_c^X(W^X - S)) / (f_t^X(W^X) + f_c^X(W^X)).$$

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 'divide and conquer' を事後結合によって求めた 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説をおく。

[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、

$$R^X = W_t^X.$$

$w \in W_t^X$ 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、

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

この仮定からは、最良の索引語選択法は W_t^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

表 2 抄録における表現と自由索引句における表現
Table 2 Expressions in abstract and free-indexing terms.

Documents	Abstract	Free-indexing terms
1	a divide-and-conquer strategy	divide and conquer strategy
2	a similar divide-and-conquer technique	divide and conquer technique
3	the 'divide and conquer' principle	(none)
4	this divide-and-conquer strategy	(none)
5	the 'divide and conquer' technique	(none)
6	a divide-and-conquer strategy	divide and conquer
7	divide and conquer	divide and conquer
8	the 'divide and conquer' principle	(none)
9	the divide-and-conquer method	(none)
10	a divide and conquer technique	divide and conquer
11	the divide and conquer paradigm	divide and conquer paradigm

るが、定冠詞をもつ句が自由索引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 索引言語の検索能力

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、

$$P_S^X = 1 - q^X(S)/q^X(W^X)$$

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、 $W^X \subseteq S$ のときで、この場合 $P_S^X \leq 0$ となる。 $S \subseteq W^X$ とすると、 P_S^X は、

$$P_S^X = (q^X(W^X) - q^X(S)) / q^X(W^X) \\ = q^X(W^X - S) / q^X(W^X)$$

であるので、検索語の仮定から、 P_S^X はほぼ $w \in (W^X - S) \cap W^X$ によって文献が求まる確率と $w \in W^X \cap W^X$ によって求まる確率の比と考えることができる。つまり、 $q^X(S)/q^X(W^X)$ を不要語の影響による検索能力の低下とみるのである。 $w \in W^X - W^X$ 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⁴⁾ (Zipf's law) 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、 $\sum_{r=1}^{83} 0.1/r \approx 1/2$ 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’、‘CON-

表 3 DIALOG の否定辞書 S_1
Table 3 Stop list of DIALOG (S_1).

AN
AND
BY
FOR
FROM
OF
THE
TO
WITH

表 4 S_1 による索引語転置ファイルの相対領域量と検索能力

Table 4 Relative size of inverted file for index terms and retrieval power by S_1 .

	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

TROL’、‘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 以下にすればなくなる。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 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。

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

Table 5 Negative dictionary based on parts of speech and negative dictionary of common function words.

The negative dictionary S_2 consists of the words with asterisk of which the parts of speech are article, preposition, conjunction, pronoun, or auxiliary verb. The negative dictionary S_3 , composed of the words with dagger is the stop list in Ref. 6).

*† 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	† WHICH
*† BECAUSE	† HEREUPON	*† OFF	*† THEMSELVES	† WHILE
† BECOME	*† HERS	† OFTEN	† THEN	*† WHO
† BECOMES	*† HERSELF	*† ON	† THENCE	*† WHOEVER
† BECOMING	*† HIM	*† ONCE	† THERE	† WHOLE
*† BEEN	*† HIMSELF	*† ONE	† THEREAFTER	*† WHOM
*† BEFORE	*† HIS	* ONESELF	† THEREBY	*† WHOSE
† BEFOREHAND	† HOW	*† ONLY	† THEREFORE	† WHY
*† BEHIND	*† HOWEVER	*† ONTO	† THEREIN	*† WILL
*† BEING	*† I	* OPPOSITE	† THEREUPON	*† WITH
*† BELOW	*† IF	*† OR	*† THESE	*† WITHIN
* BENEATH	* IMMEDIATELY	*† OTHER	*† THEY	*† WITHOUT
*† BESIDE	*† IN	† OTHERS	*† THIS	*† WOULD
*† BESIDES	† INC	† OTHERWISE	*† THOSE	* YE
*† BETWEEN	*† INDEED	* OUGHT	*† THOUGH	*† YET
*† BEYOND	*† INSIDE	*† OUR	*† THROUGH	*† YOU
*† BOTH	* INSTANTLY	*† OURS	*† THROUGHOUT	*† YOUR
*† BUT	*† INTO	*† OURSELVES	† THRU	*† YOURS
*† BY	*† IS	*† OUT	† THUS	*† YOURSELF
*† CAN	*† IT	* OUTSIDE	* TILL	† YOURSELVES
*† CANNOT	*† ITS	*† OVER	*† TO	

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、接統詞、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

表 6 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

Table 6 Relative size of inverted file for index terms and retrieval power by S_2 .

	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 による索引語転置ファイルの相対領域量と検索能力

Table 7 Relative size of inverted file for index terms and retrieval power by S_3 .

	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

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れており、一方 'IS' はほとんど 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、 $f_i(w)/f_c(w)$ 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。

$$S_\theta = \{w \mid r(w) < \theta \wedge w \in W_i \cup W_o\},$$

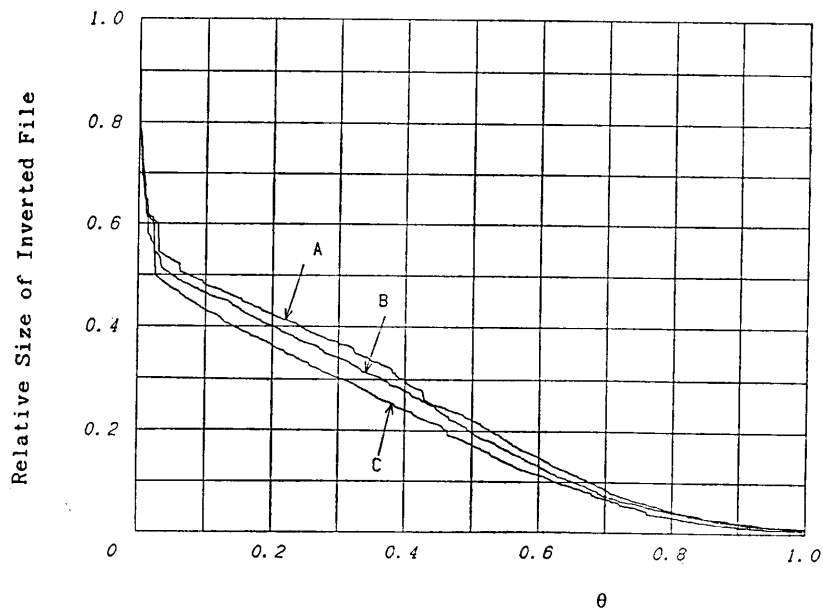


図 1 索引語転置ファイルの相対領域量
Fig. 1 Relative size of inverted file for index terms.

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、

$$|S_\theta| / |W_i \cup W_a|$$

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に生じているので、 $\theta = r(\text{'OF'})$ 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、集合AとBの自由索引句には、'50 kHz to 100 kHz'のように、範囲を示すための'TO'の生起頻度が高く

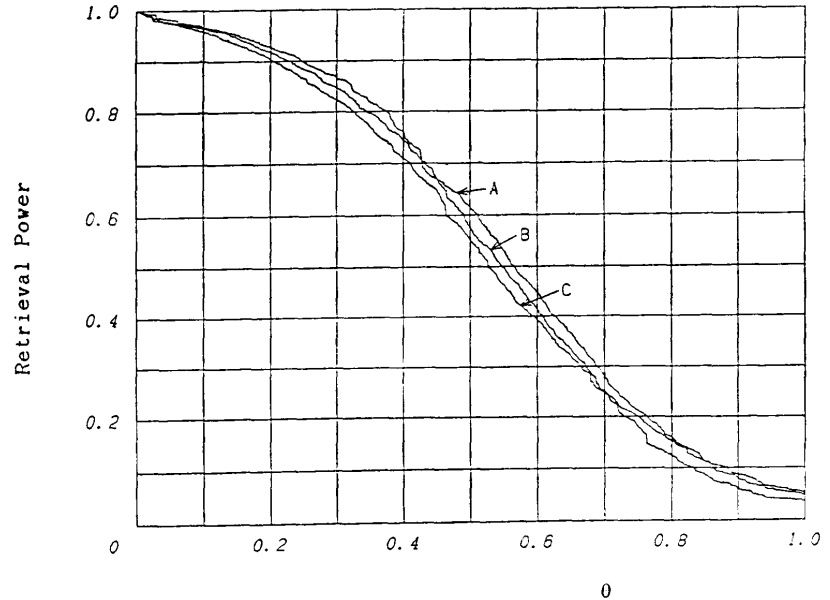


図2 検索能力
Fig. 2 Retrieval power.

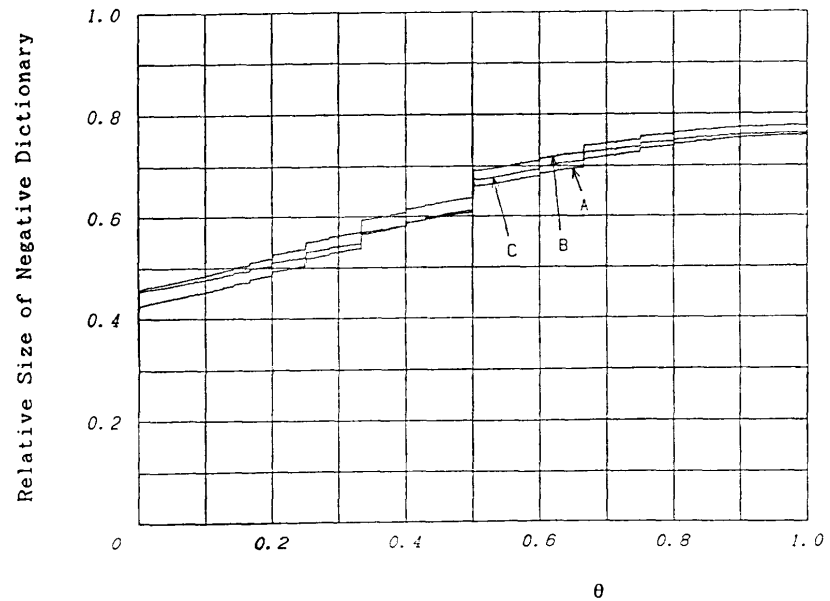


図3 否定辞書の相対大きさ
Fig. 3 Relative size of negative dictionary.

表 8 OF 点付近における諸元
Table 8 Values of parameters at OF-point.

	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 stop words	132,610	75,959	58,686

r (“TO”) は r (“OF”) よりかなり大きいのである。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。

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-I⁷⁾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 $\theta=0.379$ と OF 点の 10 倍程度にしななければならない。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 AIR⁸⁾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

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

Table 9 Modification of Table 8 by addition of the word “TO” to the negative dictionaries in Table 8.

	A	B
Relative size of inverted file for index terms	0.528	0.527
Retrieval power	0.981	0.982

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でない場合にも、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

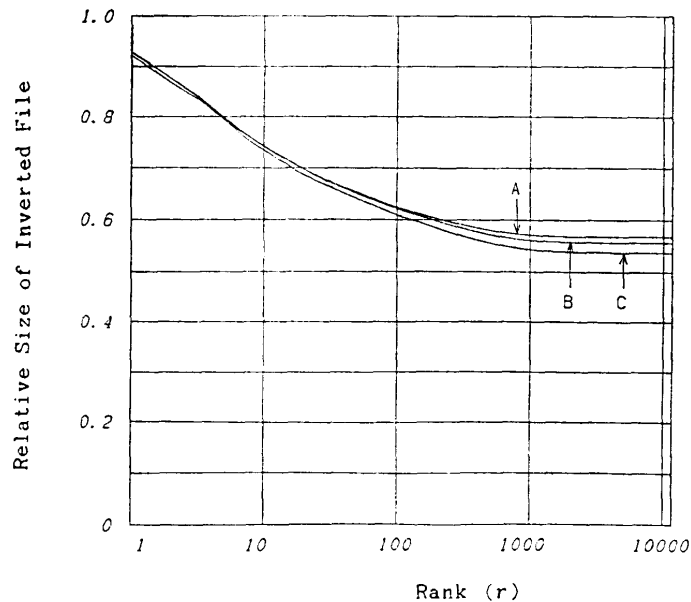


図 4 S^X による索引語転置ファイルの相対領域量
Fig. 4 Relative size of inverted file for index terms by S^X .

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

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

を考える。すると、この大きさ $|S_e|$ は 11, 667 である。次に、 S_e 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^A, S^B, S^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 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

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

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

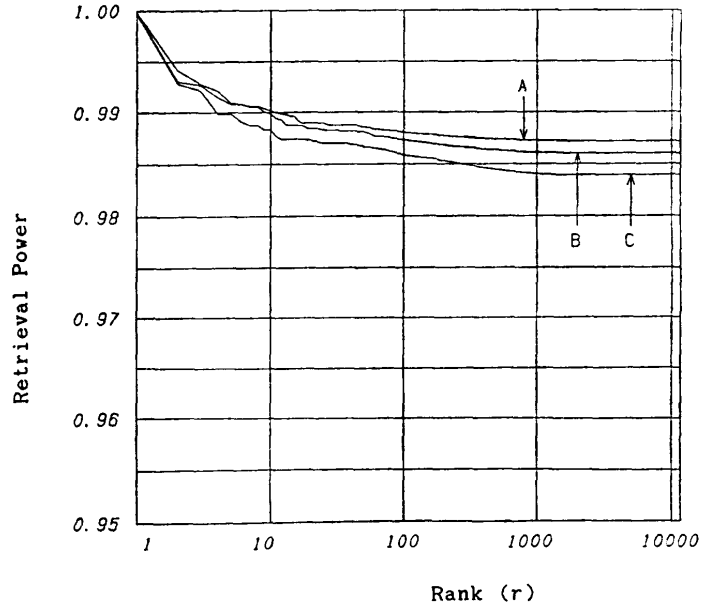


図 5 S_r^X による検索能力
Fig. 5 Retrieval power by S_r^X .

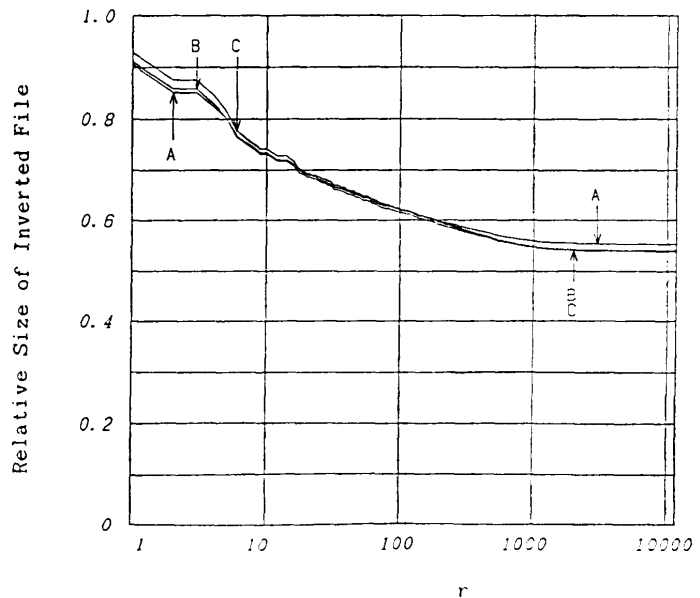


図 6 S_r による索引語転置ファイルの相対領域量
Fig. 6 Relative size of inverted file for index terms by S_r .

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、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、 $|S_r|/r$ の値は比較的变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

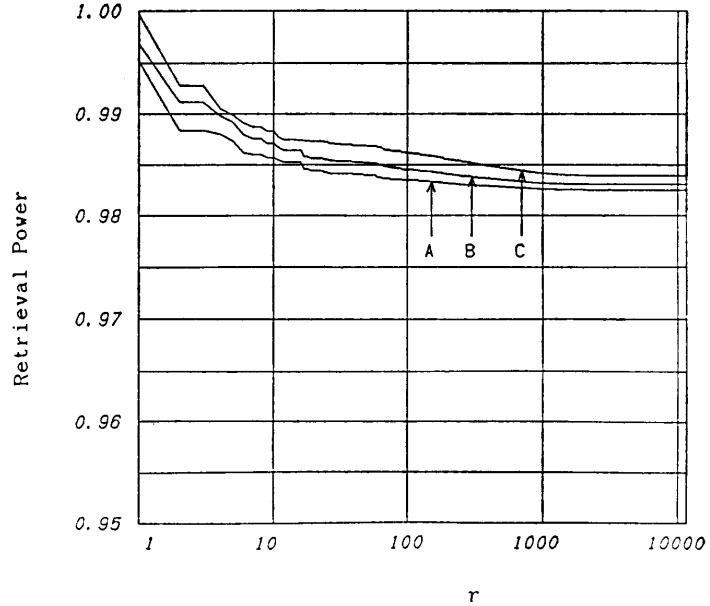


図7 S_r による検索能力
Fig. 7 Retrieval power by S_r .

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%, 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%

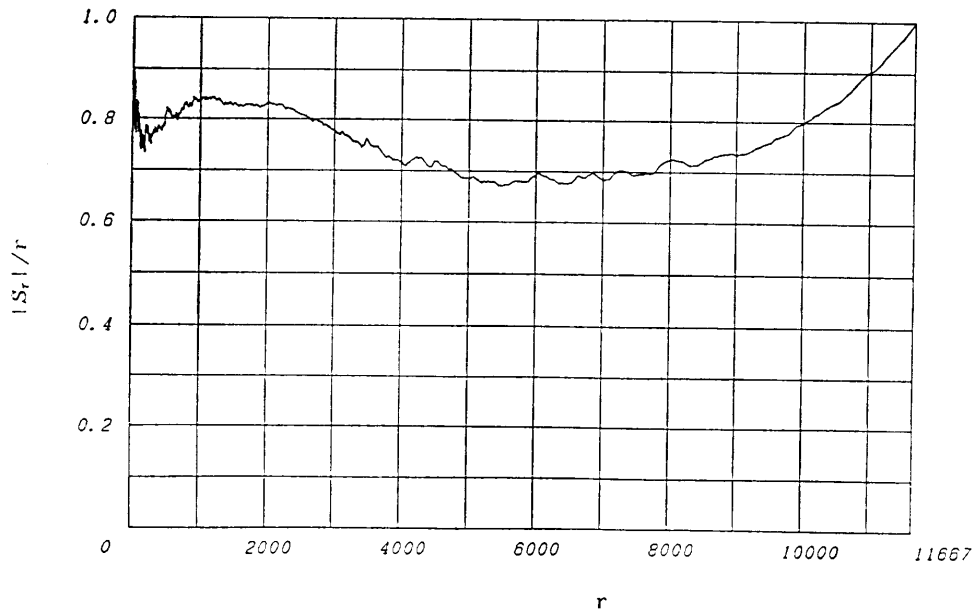


図8 共通不要語数の割合 $|S_r|/r$
Fig. 8 Ratio of number of stop words to rank $|S_r|/r$.

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

Table 10 Relative size of inverted file for
index terms and retrieval power
by common negative dictionary.

	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

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 10~12% 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。

謝辞 本稿をまとめるに当たり、九州大学工学部情報工学科の牛島和夫教授から御助言をいただいた。同教

授に謝意を表する。

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, Vol. 54, No. 2, 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, Cambridge, Massachusetts (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, 九大工学集報, Vol. 54, No. 3, pp. 183-187 (1981).

(昭和 61 年 8 月 29 日受付)

(昭和 62 年 4 月 15 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電子工学専攻修了。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。