英文科学技術文献のための否定辞書

1H-7

二村 祥一　松尾 文碩
(九州大学大型計算機センター)

1. はじめに

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.そのため実用の情報検索システムでは,この方式が一般的に採用されている.自動索引のための否定辞書(不要語リスト)として,筆者らは語の統計的性質に基づく否定辞書構築法を開発し,INSPECデータベースの情報検索に使用してきた.ここでは,INSPECテープの三つの分野に対するこの方式の否定辞書の共通部分から選んだ約1,700語からなる共通否定辞書を提示し,それを評価する.この否定辞書は,すべての英文科学技術文献に有効である.

2. サンプル文献集合

ここでは,科学技術文献に共通の否定辞書を作るために,サンプルデータとして1973年から1982年までのINSPECテープを用いた.INSPECテープは英国IEEが集積・配布している科学技術分野における代表的な二次文献情報である.INSPECテープの文献は,三つの分野A(物理学),B(電気・電子工学),C(計算機科学,制御工学,情報工学)に分けられる.分野A,B,Cの文献集合は互いに素ではなく,約20%の共通部分をもつ.INSPECテープの書誌的事項の一つに自由索引句がある.自由索引句は,非統制索引語を人手によって結合した句であり,それだけで文献の内容を同定することができないが,ここでは,この項目に出現する語とその生起頻度を否定辞書の評価に用いる.表1に文献集合A,B,Cにおける文献数と標題,抄録,自由索引句における単語数を示す.

表1. INSPECテープ(1973~1982年)からの三つの文献集合の検索項目

	A	B	C
文献数	973,735	500,741	323,336
標題			
異なり単語数	103,572	65,029	56,139
延べ単語数	11,012,493	4,747,226	2,842,323
抄録			
異なり単語数	274,185	154,231	127,836
延べ単語数	87,354,577	37,606,323	23,391,467
自由索引句			
異なり単語数	166,549	89,319	77,569
延べ単語数	19,763,849	7,841,776	4,340,930
全体			
異なり単語数	303,411	169,433	139,460
延べ単語数	118,130,919	50,195,325	30,574,720

3. 否定辞書の評価基準

不要語除去法では,索引語は,通常,標題と抄録に出現する単語から不要語を除去することによって得られる.ここでも,索引語をこのようにして選ぶことを前提とする.

まず,幾つかの定義を行う.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.また,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.$|W|$は語集合の大きさを示す.

3.1 索引語転置ファイルの相対領域量d_S^X

大部分の実用情報検索システムでは,隣接演算[1](adjacency operation)型を採用している.隣接演算型の索引語転置ファイルの領域量は,直感的に予想できるように,全索引語の生起頻度に比例する.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,

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

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と呼ぶことにする.

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

3.2 索引言語の検索能力 P_S^X

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の主観に左右されている.この

Negative Dictionary for Scientific and Technological Documents Written in English
Shouichi Futamura, Fumihiro Matsuo
Kyushu University

ために自由索引句の単語は検索語に近いものであると考えられる.そこで,次のような仮定をおく.

仮定1) 文献集合Xに対する検索語の集合をR^Xとすると,$R^X=W_i^X$.

仮定2) $w\in W_i^X$がXに対して索引語として使われる相対頻度を$p^X(w)$とすると,$p^X(w)=g_i^X(w)$.

これらの仮定のもとに,索引言語の検索能力を次のように定義する.

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

4. 語の統計的性質に基づく否定辞書

ここでは,不要語の選択の基準に$f_i^X(w)$, $f_a^X(w)$の値を使う.否定辞書S_θを次のように定義する.

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

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

この方法では,θを変化させることにより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,検索能力は連続的になだらかに低下し,破局的に悪化することがない[2].$\theta=r$('OF')の点をOF点と呼ぶ.表2-aにOF点に基づく否定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.θがこの付近の値であれば,検索能力の低下を非常に小さくおさえ,転置ファイルの大きさを不要語がない場合に比べて45~50%減少させることができる.

しかし,この方式による否定辞書には,1)全出現単語の約半数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などの問題点がある.

次に,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

文献集合A, B, Cに対してOF点に基づく否定辞書を,それぞれS^A, S^B, S^Cとする.$|S^A|=132{,}610$, $|S^B|=75{,}959$, $|S^C|=58{,}686$である.いま,これらの共通部分$S_c=S^A\cap S^B\cap S^C$を考える.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.

さて,共通否定辞書として

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

を考える.単語'TO'は特別あつかいしてS_rに加えた.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'TO'の出現が高く,r('TO')はr('OF')よりかなり大きいためOF点では不要語とはならないためである.

rが2,000を超えると転置ファイルの相対領域量と検索能力の低下が止まる.rが2,000のときを,rが最大の11,667のときと比べると,相対領域量では0.3~0.4%大きくなり,また検索能力では0.3~0.5%程度減少する.

一般に,$|S_r|\leqq r$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1にrと$|S_r|/r$の関係を示す.$r=2{,}000$の付近では$|S_r|/r$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.この辞書の大きさ$|S_{2000}|$は,1,667である.表2-bに共通否定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.

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.

表2. 転置ファイルの相対領域量と検索能力,不要語数

a) OF点に基づく否定辞書

	A	B	C
転置ファイルの相対領域量	0.544	0.544	0.502
検索能力	0.986	0.985	0.982
不要語数	132,610	75,959	58,686

b) 共通否定辞書

	A	B	C
転置ファイルの相対領域量	0.554	0.541	0.540
検索能力	0.983	0.983	0.984
不要語数	2,000		

参考文献

1) Salton, G. and McGill, M.J.: Introduction to Modern Information Retrieval, p.448, McGraw-Hill, New York(1983).

2) 松尾ほか: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,54, 2, pp.99-105(1981).

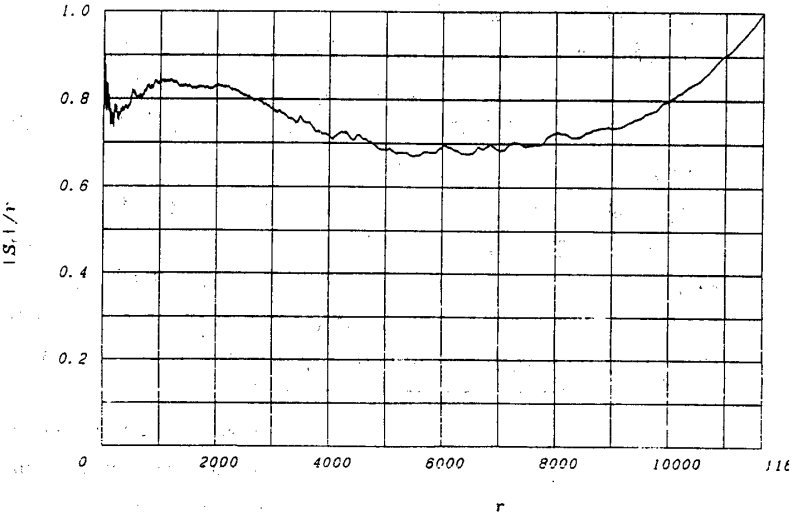

図1. 共通不要語数の割合$|S_r|/r$