文献検索における検索出力関数

橋本　寛

1　はじめに

本論文で取り扱う問題の発端はつぎのとおりである。

文献集合に対して検索をおこない，各文献についてその文献が与えられた検索質問にどの程度適合しているかを示す度合（適合度とよぶ）が計算されたとする。このとき，全文献について適合度が計算されているわけであるから，文献を出力する場合はその適合度の大きいものから順に出力すればよい。

しかし，適合度のどのレベルまで出力すればよいかは，機械的にはなかなか決めがたい。単純に，適合度に関する閾値を設けて，その値以上の適合度をもつ文献だけをとり出すという考え方が一般的だが，その閾値以上の適合度をもつ文献が存在しない場合は，文献が全くとり出せないことになる。したがって，そのようなときは閾値を少し低くしてみる必要がある。また逆に，ある閾値以上の文献が多すぎる場合は，閾値を高くして適合度の大きい上位の文献を適当な個数だけ出力すれば十分である。

このように適合度の閾値を自動的に変化させる簡便な方法として，本論文では，検索出力関数なるものを考え，これについて議論をおこなっている。この検索出力関数の特別な場合として，これまで用いられている検索結果の代表的な出力方式を説明することができる。

また，検索結果として何個の文献を出力するかを決定する検索出力関数の問題は，ある集合のクラスタ化の問題とも関係があり，本論文では，この点

についても考察をおこなっている。すなわち，クラスタ化すべき要素の間に距離のようなもの（以下，擬似距離または距離とよぶ）を考えるならば，ある一定の値未満の距離のものは同じ類に属するとしてクラスタ化することができる。このとき，距離に関する閾値を大きくすれば，生ずる類の個数は少なくなる。したがって，距離に関する閾値と生成される類の個数との関係は，適合度に関する閾値と出力される文献の個数との関係と形式的に同一である。

さらに，本論文では，検索出力関数がある種の選抜方式とも関係することを示し，その点についても若干の考察をおこなっている。

検索出力関数は以上のような点において興味があるが，筆者らの研究〔1，2〕以外には明確な形で議論された例はないようにおもわれる。

2　適合度と文献累積曲線

ある検索システムに蓄積されている文献と，それに対して与えられた検索質問との一致の程度を適合度とよぶことはすでに述べた。通常，文献検索システムでは全文献について，与えられた検索質問に対する適合度が計算されると考えることができる。従来の検索システムでは適合度が0，1の2値のみをとる検索方式のものが多かった。しかし，最近は適合度が連続値で与えられる検索方式〔1，3〕も増えてきている。その代表的なものは文献ベクトルと質問ベクトルとの間で適当に定められた相関を計算する方式であろう。ここでは，適合度が連続的な数値で与えられ，しかもその値が0と1の間の単位区間に含まれるものと仮定しよう。この単位区間以外の適合度を与える検索方式においても適当な正規化で適合度が単位区間におさまるようにすることは可能であるし，2値的な検索の場合もこの特別な場合として取り扱うことができるので，この仮定は妥当であろう。

ここで，ある検索をおこなったとき，ある値 r 以上の適合度をもつ文献の個数を m(r) で示し，m(r) の描く曲線を文献累積曲線または累積曲線とよぶ

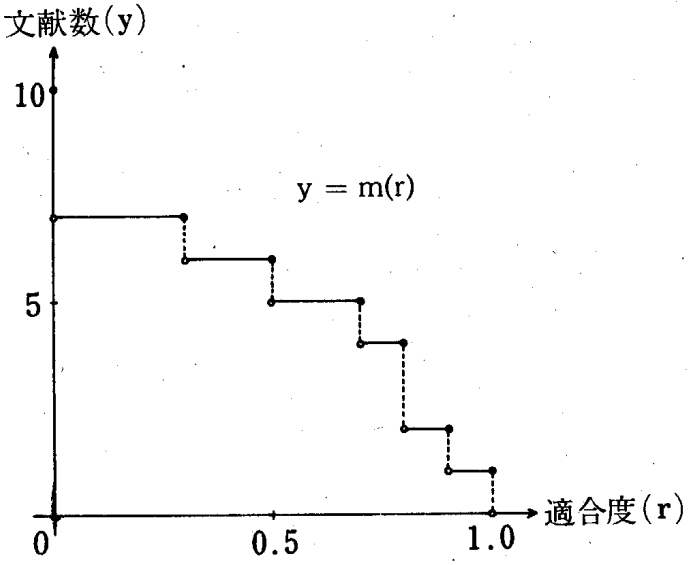
ことにする。

たとえば，ある与えられた検索質問に対し，10個の文献について検索をおこなったとき，その結果が表1のように示されたとする。表1は計算された適合度の大きい順に文献を並べたものである。○印の文献は与えられた検索質問に適合する文献（適合文献）である。このときの文献累積曲線を描くと，図1のようになる。

表1　計算された適合度の例

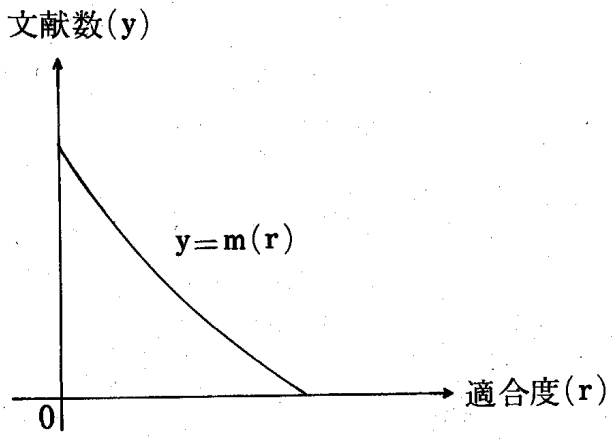
文　献	適合度
○D(1)	1.0
○D(2)	0.9
○D(3)	0.8
D(4)	0.8
○D(5)	0.7
D(6)	0.5
D(7)	0.3
○D(8)	0.0
D(9)	0.0
D(10)	0.0

図1　現実の累積曲線の例



現実の場合には，文献数が有限であるので，累積曲線 m(r) は図1に示すように階段状の曲線となる。しかし，文献数が十分大きければ，図2のような連続的曲線で近似して考えることは許されるであろう。文献数の軸を全文献数で正規化しておけばこのことはより明白である。

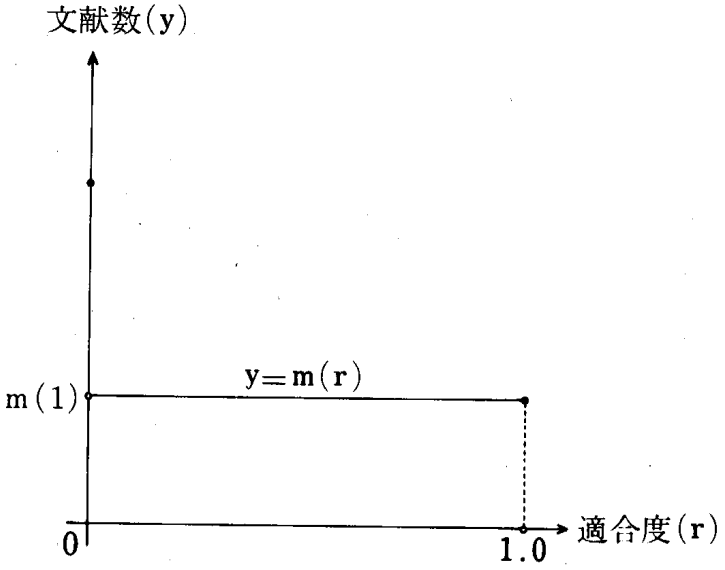
図2　連続な累積曲線



この文献累積曲線は，その定義から適合度 r に関して非増加曲線をなす。

2 値的な場合の文献累積曲線は図 3 のようになる。すなわち，適合度 1 の文献が m(1)個あり，他の文献はすべて適合度 0 である。

図 3　2 値的検索の累積曲線

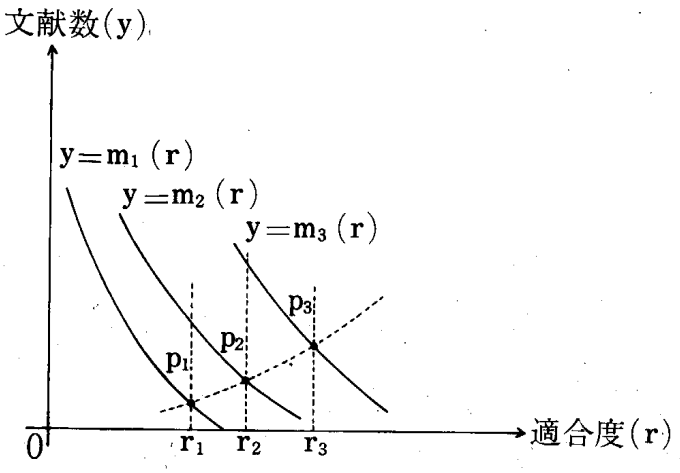


3　検索出力関数

3．1　単調性

蓄積されている文献集合に対し，3 個の検索質問 Q_1, Q_2, Q_3 を与えて検索した結果の文献累積曲線が，それぞれ図 4 に示す $y = m_1(r)$, $y = m_2(r)$, $y = m_3(r)$ で与えられたとする。このとき，文献の適合度に関する出力の閾値

図 4　閾値の切換え



が r_2 であったとする。また，質問 Q_2 に対しては適当な個数の文献が出力されるが，質問 Q_1 に対しては全く文献が出力されず，一方質問 Q_3 に対しては十分すぎる文献が出力されるとする。このようなときは，すでに議論したように，質問 Q_1 に対しては適合度の閾値を少し下げて r_1 とすること，また質問 Q_3 に対しては閾値を少し上げて r_3 とすることは実用上妥当である。

この場合，適合度 r_2 の点においては

$$m_1(r_2) < m_2(r_2) < m_3(r_2)$$

であるから，適合度 r_1, r_3 の点において

$$m_1(r_1) > m_2(r_2)$$

$$\sim$$

$$m_2(r_2) > m_3(r_3)$$

となっては不合理であろう。なぜなら，たとえば質問 Q_1 について考えると，閾値 r_2 においては何も文献が出力されず，そのため閾値を下げてみたわけであるから，下げたときの出力文献数が，適合度に関する閾値 r_2 のときの質問 Q_2 の出力数をこえることは不当である。質問 Q_2 と Q_3 との関係についても同様である。したがって

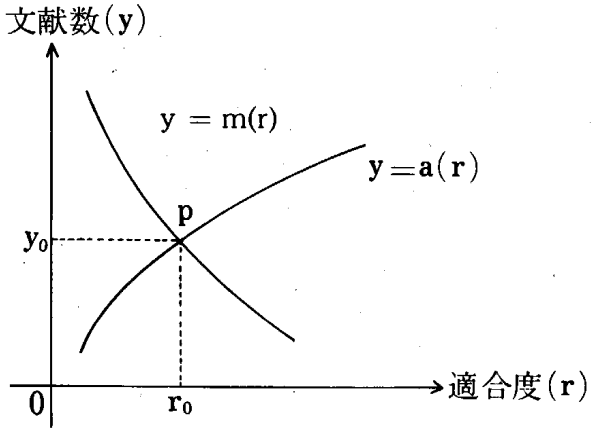
$$m_1(r_1) \leqq m_2(r_2) \leqq m_3(r_3)$$

なることを仮定することは妥当であろう。

ここで，閾値をこのように変化させるための1つの簡便な方法として，3点 $P_1(r_1,m_1(r_1)), P_2(r_2,m_2(r_2)), P_3(r_3,m_3(r_3))$ をとおる曲線を考えておき，この曲線と累積曲線との交点によって，適合度に関する閾値を決定する方法を考えることができる。この曲線を与える関数を検索出力関数と名付ける。この関数は上記の議論によって非減少の関数（広義の単調増加関数）となる。

以上のことをまとめると，ある質問に対して検索をおこなったとき，そのときの出力文献数は図5に示すように文献累積曲線　$m(r)$　と検索出力関数 $a(r)$ の交点Pの縦座標成分 y_0 で与えられ，その交点Pの横座標成分 r_0 以上の適合度をもつ文献が出力されることになる。

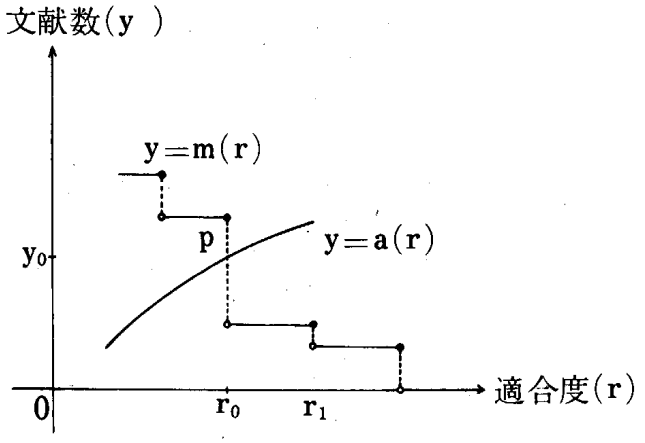
図5　文献累積曲線と検索出力関数との交わり



3．2　実用上の問題

現実の文献集合では，文献数が有限であるので，すでに述べたとおり，文献累積曲線は階段状の関数であり，文献の出力においてこの関数に特有な問題が生ずる。それは，文献累積曲線と検索出力関数が図6のように交わったとき生ずる。すなわち，交点を $P(r_0,y_0)$ とすれば，出力文献数は y_0 であるが，適合度 r_1 以上の文献を出力するときは，出力数は y_0 より小さく，適合度 r_0 以上の文献を出力しようとすると，その数は y_0 より大きくなってしまう。したがって，まず適合度が r_1 以上の文献は無条件で出力し，適合度 r_0 の文献

図6　階段状の累積曲線と検索出力関数



については，適当に定めた順番で，たとえば文献番号の順番で，出力文献数が y_0 に達するまで出力することになる。もちろん，文献番号順で出力しなけ

ればならない特別の理由は何もない。

なお，具体的に文献の出力数を決めるさいに，文献累積曲線および検索出力関数のグラフを描いて，交点を求める必要は全くない。出力数はつぎのようにしてきわめて簡単に決定できる。

まず，各文献をその適合度の大きいものから順に並べる。同じ値の適合度をもつ文献に対しては，一定の順番で，たとえば文献番号の順番で並べるものとする。それを表２の形にまとめる。各適合度について，大きい方から適合度 $r_{(i)}$ における検索出力関数の値 $a(r_{(i)})$ と i を比較して，$a(r_{(i)}) \geqq i$ ならば，

表２　適合度順に並べた文献

文　献	適合度
$D_{(1)}$	$r_{(1)}$
$D_{(2)}$	$r_{(2)}$
⋮	⋮
$D_{(i)}$	$r_{(i)}$
⋮	⋮
$D_{(k)}$	$r_{(k)}$

$r_{(i)} \geqq r_{(i+1)}$　$(i = 1, 2, \cdots, k-1)$

文献 $D_{(i)}$ は出力し，i を１つ増してつぎの文献へ移る。もし，$a(r_{(i)}) < i$ ならば，文献の出力は文献 $D_{(i-1)}$ でやめ，それ以下の文献は出力しない。そのとき，出力文献数は $i-1$ である。このようにして，文献の出力数は容易に決定できる。

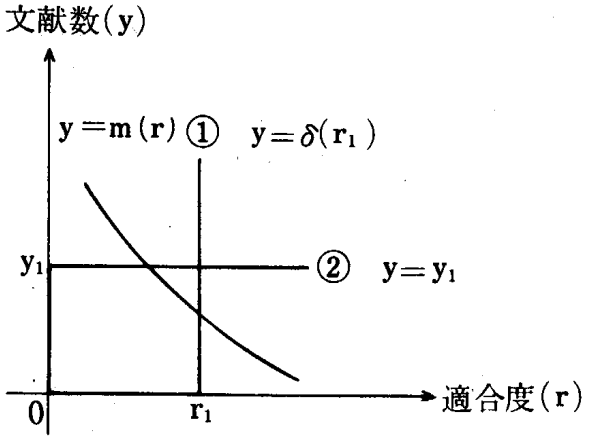
３．３　特殊な検索出力関数

検索出力関数の特別な場合として説明することのできる主要な検索方式について述べる。

(1)　適合度がある値以上の文献を出力するとき

このときの検索出力関数は，適合度に関する閾値 r_1 で立ち上がる曲線となり，図７の①で示される。このときは，すでに述べたように，この閾値 r_1 以上の適合度をもつ文献はすべてとり出され，またこの値以上の適合度をもつ文献が存在しない場合には，何も出力されない。適合度の絶対的な大きさの

図7　特殊な検索出力関数（$y=\delta(r_1)$ と $y=y_1$



みが問題になる。このときの検索出力関数をデルタ型の検索出力関数とよび，$y = \delta(r_1)$ で示す。

(2)　適合度の大きい文献から一定個数の文献をとり出すとき

このときの検索出力関数は $y = y_1$ で与えられ，図7の②で示される。この検索出力関数を用いるときは，検索結果における適合度の相対的な大きさすなわち順位のみによって，その文献が出力されるかどうかが決定される。

(3)　適合度がある値以上のものを一定個数以下でとり出すとき

このときの検索出力関数は，適合度 r_1 で立ちあがり，r_1 以上の適合度に対し，一定値 y_1 をとる階段状の関数となり，図8で示される。これは前記(1)(2)の検索出力関数を組み合わせたものと考えることができる。この関数をステップ型の検索出力関数とよぶことにする。

図8　ステップ型検索出力関数

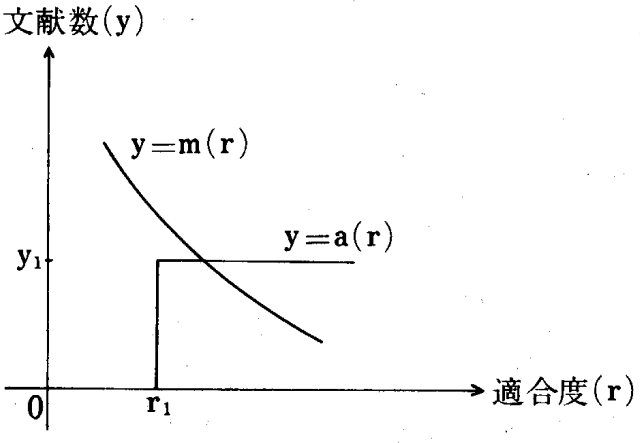
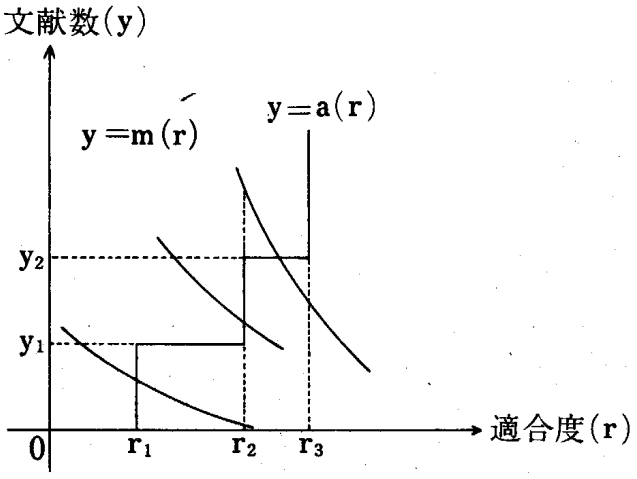


図9　多閾値の検索出力関数



(4)　適合度に複数個の閾値を設けて出力するとき

多閾値の検索出力関数として，図9のような階段状の関数が考えられる。このような検索出力関数に対しても，出力文献数が文献累積曲線との交点によって定まるのはこれまでどおりである。

この図9で示される検索出力関数はつぎのように解釈することができる。

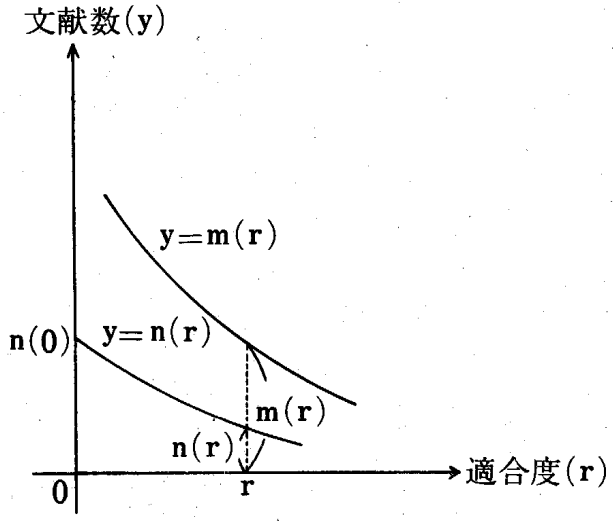
まず，あらかじめ適合度に関する閾値が r_2 に設定されているとする。このとき，適合度 r_2 以上の文献数が定められた一定の個数 y_1 以上で y_2 以下であれば，そのときの適合度 r_2 以上の文献の個数を出力文献数とする。もし，その適合度 r_2 以上の個数が y_1 より小さければ，閾値を下げて r_1 とし，r_1 以上の適合度をもつ文献を y_1 以下の個数で出力する。また，r_3 より大きい適合度をもつ文献の個数が y_2 以下で，かつ適合度が r_2 以上である文献の個数が y_2 以上であれば，y_2 個の文献を出力する。r_3 以上の適合度をもつ文献数が y_2 より大きければ，r_3 以上の適合度をもつ文献はすべて出力する。

4　平均の検索効率

4．1　適合文献の累積曲線

ある与えられた検索質問Qに対して検索をおこなったとき，適合度 r 以上の文献の中に含まれる適合文献の個数を n(r) で示し，この適合度 r の関数 n(r) の描く曲線を適合文献累積曲線とよぶことにする（図10参照）。

図10　適合文献累積曲線



このとき，適合度 r 以上の文献をとり出したときの検索効率すなわち再現率 $\alpha(r)$ と適合率 $\beta(r)$ はつぎのように定められる。

$$\alpha(r)=\frac{n(r)}{n(o)} \qquad \beta(r)=\frac{n(r)}{m(r)}$$

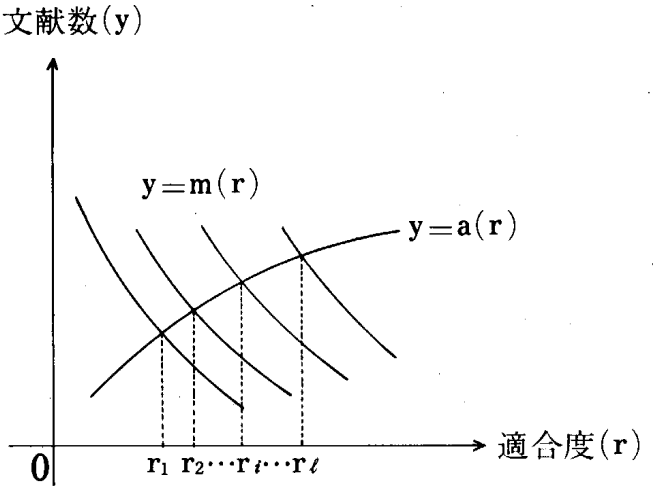
4．2　平均の効率の定義

いま，検索出力関数 y = a(r) を用いて，ℓ 個の質問 $Q_1, Q_2, \cdots, Q_\ell$ に対し，検索をおこなったときの再現率を $\alpha_1(r_1), \alpha_2(r_2), \cdots, \alpha_\ell(r_\ell)$ とし，また適合率を $\beta_1(r_1), \beta_2(r_2), \cdots, \beta_\ell(r_\ell)$ とする。ただし，$r_1, r_2, \cdots, r_\ell$ は各質問に対する文献累積曲線と検索出力関数との交点の横座標成分である。$r_1 \leqq r_2 \leqq \cdots \leqq r_\ell$ と仮定しよう（図 11 参照）。

ここで，検索出力関数 y = a(r) を用いるときの平均の検索効率すなわち平均再現率 $\bar{\alpha}$，平均適合率 $\bar{\beta}$ を次式で定義する。

$$\bar{\alpha}=\frac{1}{\ell}\sum_{i=1}^{\ell}\alpha_i(r_i) \qquad \bar{\beta}=\frac{1}{\ell}\sum_{i=1}^{\ell}\beta_i(r_i)$$

図11　多数の文献累積曲線と検索出力関数



検索出力関数の具体的な形を仮定して，そのパラメータを変化させることにより，平均再現率対平均適合率の曲線が得られる。

たとえば，a(r) = δ(r) として，平均再現率対平均適合率の曲線を描くと図 12 のような曲線が得られる。しかし，1 個の質問に対して再現率対適合率の

曲線を描くと図13のようになり，一般には単調ではない。図13は表1について計算し，描いたものである。

図12　平均再現率対平均適合率曲線

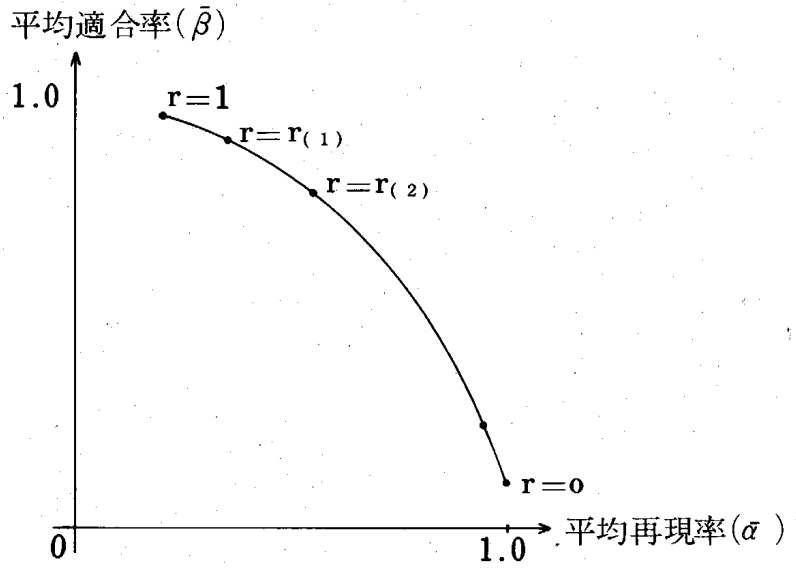
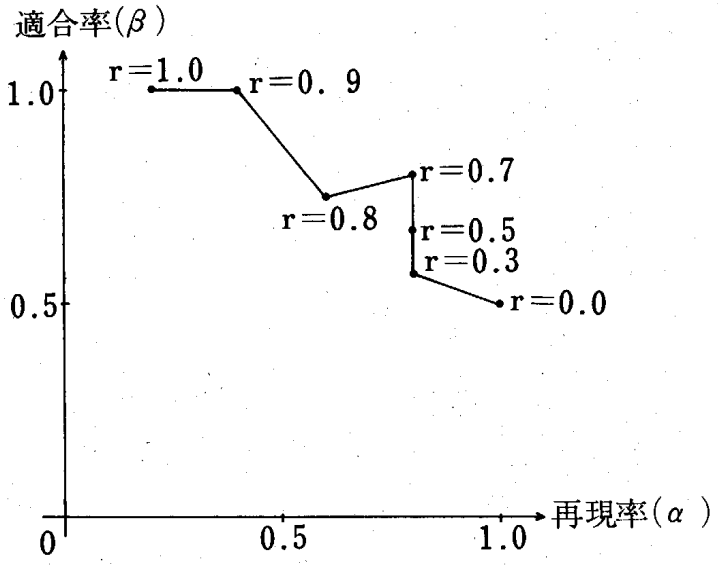


図13　現実の再現率対適合率曲線

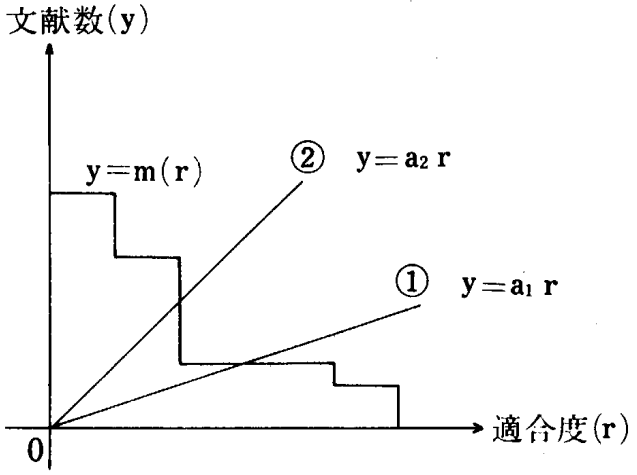


5　検索出力関数を用いた実験

線形の検索出力関数 y = ar を用いて，パラメータ a を変化させながら，出力文献数を決定し，そのときの a に対する平均の検索効率を計算した。検索方式は Fuzzy 集合の理論を利用したキーフレーズ間の関連を考慮する手法

〔1〕であって，適合度は0と1の間の数値で与えられる。キーフレーズとは文献および検索質問の内容を表現するために用いられる単語の系列である。検索の対象とした文献数は350個である。文献数が有限であるので，文献累積曲線は階段状となる。このときの文献累積曲線と検索出力関数との関係は図14のようになる。同図②のような場合には，交点の閾値に相当する同じ適合度をもつ文献に対しては，文献番号の順に，交点できまる文献数まで出力する。

図14　線形の検索出力関数



平均の効率を求めるために21個の検索質問について，線形の検索出力関数とデルタ型の検索出力関数とを用いて検索をおこなった。その結果を図15に示す。

図15　デルタ型および線形の検索出力関数に対する効率

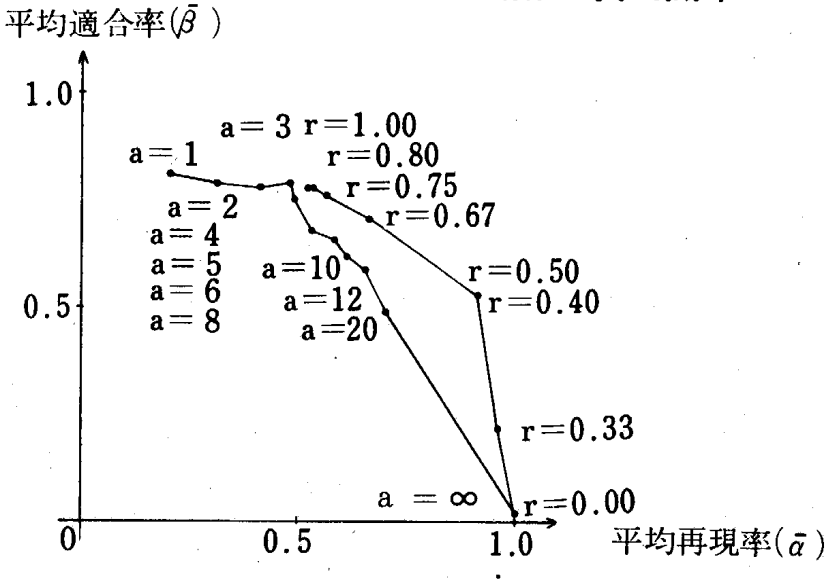


図15によれば，デルタ型の検索出力関数の方がよい結果を与えている。しかし，デルタ型の検索出力関数では出力文献数に関する制御ができない。これに対し，線形の検索出力関数ではそれがある程度可能である。

なお，当然のことながら，検索出力関数だけで検索効率を改善することには限界がある。検索効率を上げるためには適合度が適確となるような手法を採用しなければならない。

6　クラスタ化との関係

検索出力関数の問題はクラスタ化の問題とある面で関係している。ある与えられた対象の集合をクラスタ化しようとするとき，文献検索と同じような事情に出会う〔4〕。もちろん，その本質は相当異なっているが，表面上は同様の事情であるようにみえる。

クラスタ化すべき対象の間に擬似的距離を考え，この距離は0と1の間の数値で与えられるものとする。距離を利用してクラスタ化をおこなう手法は多数提案されている〔3，5，6〕。しかし，ここではつぎのような単純なクラスタ化の手法を考えよう。すなわち，ある一定値未満の距離にある対象どうしは同じ類に属するものとする手法である。いいかえれば，距離的に近いものは同じ類に属し，遠く離れているものは別の類に属するという考え方である。

たとえば，クラスタ化すべき対象間の距離的関係が表3のように与えられたとする。かりに，0.41未満の距離にある対象は同じ類とすると，図16に示すごとく，3個の類が生ずる。この距離に関する閾値をdとしたときの類の個数，すなわち相互間の距離がd未満である対象は同類とするときの類の個数をg(d)で示せば，dを変化させたときの類の個数 $y = g(d)$ は図17のようになる。

表3　対象間の距離

	E_1	E_2	E_3	E_4	E_5	E_6
E_1		0.16	0.56	0.61	0.96	0.8i
E_2			0.43	0.59	0.97	0.86
E_3				0.45	0.83	0.86
E_4					0.40	0.43
E_5						0.32
E_6						

図16　距離によるクラスタ化

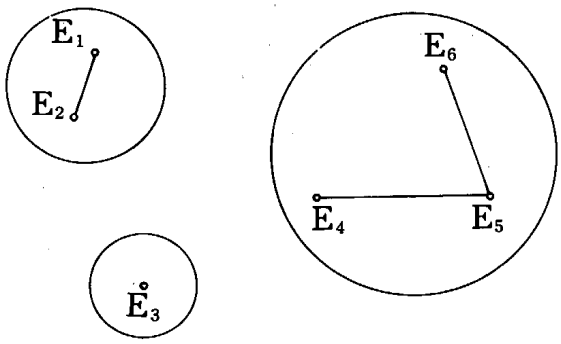


図17　類の個数の変化

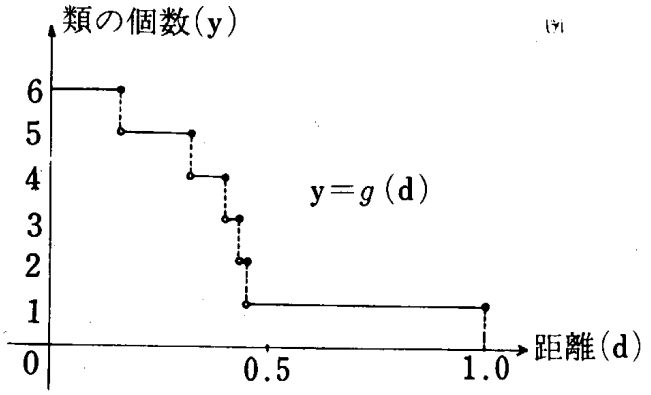


図17からもあきらかなように，距離に関する閾値を大きくすれば，生ずる類の個数は小さく，閾値を小さくすれば，すなわち比較的近いものでもある値以上離れていれば別の類とするときは，生ずる類の個数は大きい。

このとき，距離に関する閾値と類の個数との関係は，文献累積曲線と同じ非増加の階段状の曲線で示される。したがって，検索出力関数に相当するものを考えれば，生ずる類の個数に制限をつけながら，クラスタ化をおこなうことが可能となる。

出力される文献数に対応するものは生ずる類の個数であり，全文献に対応するものはクラスタ化すべき要素の個数である。なお，ある閾値でクラスタ化したとき，類の個数が多すぎた場合は，何らかの方法で小さい類を合併して，全体の類の個数が制限内におさまるようにする。

7　選抜方式との関係

これまで述べてきた検索出力関数は，入学試験等の選抜方式とある種の対応があり，興味深い。まず，3．3で述べた(1)の方式であれば，ある得点以上の受験者はすべて合格となり，これは資格試験等に採用されている。(2)の方式であれば，順位が一定の範囲にあれば合格となり，通常の競争試験はこの方式であろう。(3)の方式であれば，順位および点数によって合否が決まる。これはいわゆる水準点のある場合の合否判定方式と考えることができる。

以上が，比較的よく採用されている方式であるが，他の検索出力関数も選抜方式と結びつけて考えることができる。従来は，人手で処理する事情と手順の簡単さのために，上記のような選抜方式が採用されているが，機械で処理する場合には，伝統的な方式にとらわれる必要はない。他の方式についても検討してみる余地はあるとおもわれる。

ある得点 t 以上の受験者数を f(t) で示し，検索出力関数に相当する関数を h(t) で示せば，合格の最低点は $y = f(t)$ と $y = h(t)$ の交点 $P(t_0,y_0)$ の横座標成分 t_0 で与えられる(図18参照)。同様に合格者数は交点の縦座標成分 y_0 で与えられる。得点 t を正規化して $0 \leqq t \leqq 1$ とすれば，これまで考えてきた文献検索と形式的に同一である。

図18　得点の分布

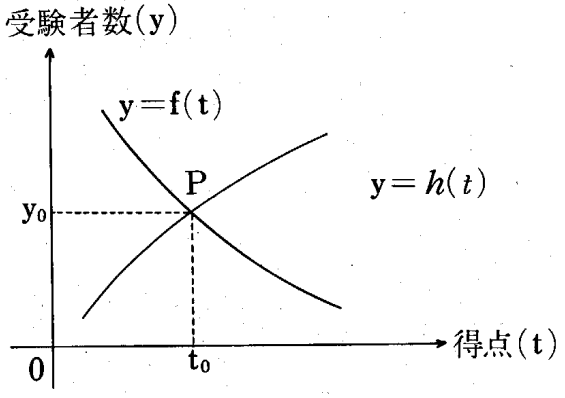
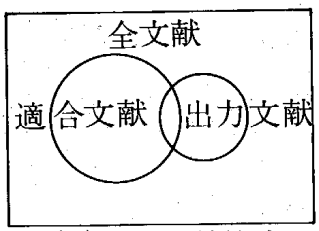
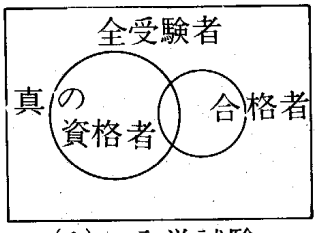


図19　文献検索と入学試験との対応



(a)　文献検索



(b)　入学試験

文献検索と入学試験の関係についてさらに考えてみると，図19に示すごとく，まず文献に受験者，したがって文献数に受験者数が対応する。さらに適合度に対応するものは受験者の得点である。適合文献には真に入学する資格のある受験者が対応し，入学試験の目的は真に入学する資格のある受験者を選抜することであるから，検索効率に対応するものは入学試験の信頼度である。ただし，この信頼度は，検索効率の場合とことなり，算出するのは実際上は相当困難である。このように，文献検索と入学試験との間にはある種の対応関係が成立する。

選抜方式とクラスタ化との間にも，当然関係がある。たとえば，入学試験等において，合格者の最低点を決定するときに得点の分布を見て，募集人員にとらわれずに，得点に差があるところに最低点を設定することがしばしばおこなわれる。これと同様の事情はクラスタ化においてもみられる。たとえば，すでに述べた距離を利用するクラスタ化をおこなっているさいに，距離に関する閾値をごくわずか変化させたとき，類の個数が大きく変化する場合がある。このような不安定なことは，一般に，好ましくなく，閾値を設定しなおす必要がある。このため，クラスタ化においては，閾値が少しぐらい変化しても類の個数が変化しない箇所に設定されることがある〔4〕。

検索出力関数または選抜方式と類似の事情は，一般に，ある与えられた集合（またはパタン）の中から，一定の部分集合をとり出すさいに生ずる。そのようなときには，ここでの議論と同様の議論をおこなうことが可能である。しかし，形式的に同様にとりあつかうことができても，その問題の本質は，それぞれに固有の事情があって，相当異なっている。

8　ま　と　め

文献検索の結果において，出力すべき文献数を決定するための簡便な方法として，検索出力関数を提案し，それに関し若干の議論をおこなった。この検索出力関数は文献検索の本質をなすものではないが，実際上は必要であろ

う。

また，検索出力関数とクラスタ化との関連についてもふれ，クラスタ化と文献検索との間にある種の対応が成立することを示した。さらに，選抜方式と検索出力関数との間にも対応関係があることを示した。

今後の課題として，多量の文献をとり扱う場合の大きな問題である探索時間の短縮と関連して，検索出力関数を考察することがある。

普通，探索の高速化のためには多段探索〔3〕などの手法が採用されているが，このとき考慮されるのは適合度に関する閾値だけである。すなわち，ある閾値以上の適合度をもつ可能性のある文献だけについて，適合度を計算することにより，時間の短縮をはかるものであって，これはデルタ型の検索出力関数を用いた探索の高速化であると考えることもできる。

謝　辞

本研究の一部は，筆者が名古屋大学大学院工学研究科に在学中に，なされたものである。ご指導いただいた福村晃夫教授および当時の福村研究室の諸氏に深謝する。

文　献

〔1〕橋本，阿部，福村：〝キーフレーズ間の関連を考慮した情報検索の一手法〟電子通信学会オートマトン・インホメーション理論研究会資料（1972年3月）

〔2〕橋本，阿部，福村：〝文献検索における2種類の出力関数について〟昭和48年度電子通信学会全国大会　講演論文集6（昭和48年3月）

〔3〕G. Salton: *"Automatic Information Organization and Retrieval"*, McGraw-Hill (1968)

〔4〕C.C. Gotlieb, S.Kumar: "Semantic Clustering of Index Terms", *J.ACM*, Vol. 15, No.4 PP493～513, (Oct. 1968)

〔5〕坂井　編："パターン認識の理論"，共立出版（1967）

〔6〕奥野　他："多変量解析法"，日科技連（1971）