検索実験における評価指標としての平均精度の性質

岸 田 和 明[†,††]

最近の情報検索の実験においては各手法やシステムの検索性能の評価指標として平均精度，あるいは平均精度の平均が利用されている．しかし，評価指標としての平均精度の統計学的な性質についてはあまり知られてはいない．本研究の目的は，平均精度を数学的に定義して，その基本的な性質を整理するとともに，検索実験における検索課題が無限母集団からの無作為標本であると仮定した場合に，2つの手法間の性能を平均精度を用いて検定する際の諸問題について議論することにある．特に検定の問題に関しては，第1に，適合判定の変動が平均精度を用いた手法間の差の検定に与える影響を調べる．これは統計学分野で開発された測定誤差モデルを導入することによって行う．第2に，近年のテストコレクションの一般的な作成方法である pooling に起因する適合文書の未発見が性能比較に与える影響を議論する．これらの分析における具体例として，NTCIR-1 の結果の一部を利用して，数値的な計算を試みる．

Property of Average Precision as Performance Measure for Retrieval Experiment

Kazuaki Kishida[†,††]

Average precision is often used for evaluating methods or models at retrieval experiments. However, statistical properties of average precision or mean average precision have not yet been known sufficiently. The purpose of this paper is (1) to define mathematically average precision and to analyze its properties from the mathematical formula, and (2) to discuss some issues on statistical test for determining a difference of retrieval performance between two systems by using mean average precision as an evaluation measure. To do this, first, a mathematical model of measurement error developed in statistical science is introduced for estimating the degree to which the variation of relevance judgments change the result of statistical test by average precision. Second, we examine the effect of discovering relevant documents that were not found due to adopting pooling method for developing test collection. A part of results at NTCIR-1 Workshop is used for showing some examples in a real setting.

1. はじめに

情報検索の手法やシステムは一般的には応答速度や必要な資源などさまざまな観点から評価されうる．それに対してテストコレクションを使った検索実験においては特に検索の有効性（effectiveness）または性能（performance）に焦点が当てられ，検索課題（topic）に対する正解文書（あるいは適合文書）をいかに「上手に」検索できるかという観点からの評価が中心に据えられている．

このための標準的な評価指標は再現率（recall）と精度（precision）である．前者は「すべての適合文書のうち検索されたものの割合」，後者は「検索された文書のうち適合しているものの割合」と定義される．しかし，これらの指標は本来的にはブール演算に基づく伝統的な検索方法に関する評価のために開発されたものである．そのため，最近の主流である，文書をその適合度の順で出力する手法を評価する場合には，これらの指標は直接的には適用できず，若干の工夫を加えなければならない．

その代表的な指標が平均精度（average precision）である．これは簡単にいえば「各適合文書が検索された時点での精度の平均」[1)]である（より詳しい定義は後述）．そして，平均精度自体は検索課題ごとに計算されるが，検索実験では複数の検索課題が用意されるため，課題ごとの平均精度をさらに平均したもの（mean average precision：MAP）が最終的には各手法の性能

† 駿河台大学文化情報学部
Faculty of Cultural Information Resources, Surugadai University
†† 国立情報学研究所
National Institute of Informatics

比較に用いられる．その他の指標として R-precision などもあるが[1)]，TREC☆や NTCIR☆☆をはじめとする最近の検索実験では平均精度が中心的な役割を果たしている．

しかし，情報検索研究の初期に提案された再現率や精度とは異なり，平均精度の歴史は浅く，評価指標としての信頼性や頑健性に関する研究は最近ようやく本格化したばかりである[1),2)]．むしろその特徴や性質についてはほとんど知られていないといってよく，適合判定の変動に対する感度（sensitivity）や多段階の適合判定への拡張についての研究が急務となっている．

本稿の目的は 2 つの検索手法の性能を比較評価するための平均精度による統計的検定の問題を議論し，評価指標としての平均精度の性質に対する理解を深めることにある．特に，(a) 適合判定の変動と，(b) テストコレクションを作成する際に見落とされた適合文書の存在の 2 つの要因が平均精度に与える影響に焦点を当てる．前者 (a) については統計学分野で開発された測定誤差に関するモデルを導入する．そして，このモデルに基づくシミュレーションを現実のテストコレクションのデータを使って実行し，上記 (a) と (b) に関する実際の影響の大きさを試算する．

以下，まず 2 章において，検定の問題を議論するための準備として，平均精度を数学的に定義し，それを利用していくつかの基本的性質を明らかにする．次に 3 章では平均精度による統計的検定について，対標本による検定方法を議論したうえで，適合判定が確率的に変動する場合の影響を分析するための測定誤差モデルを導入する．さらにこの章では pooling の方法をテストコレクションでの適合判定に用いた結果見落とされた適合文書の影響について論じる．これらの議論に基づいて，4 章では日本語テストコレクション NTCIR-1 の具体的なデータを用いた数値計算例を示す．以上の結果を 5 章で考察し，ある有意水準で手法間の差を結論するために必要な平均精度の差の大きさについて考える．

2. 平均精度の定義と基本的性質

2.1 平均精度の数学的定義

意外なことに平均精度を求める数式はこれまでの先行研究の中でほとんど明示的に示されていない．そこで新たに平均精度を数学的に表現する必要がある．まず N を出力文書の総数，x_i を出力順第 i 位の文書の適合/不適合の状態を示す変数とする（$i = 1, \ldots, N$）．適合/不適合の判定は 2 値とし，適合ならば $x_i = 1$，不適合ならば $x_i = 0$ とおく．以上の記号を使うと，ある 1 つの検索実行についての平均精度 ν は

$$\nu = \frac{1}{\sum_{i=1}^{N} x_i} \sum_{i=1}^{N} \frac{x_i}{i} \left(1 + \sum_{k=1}^{i-1} x_k \right) \tag{1}$$

で定義できる．

たとえば $N = 4$ の検索結果が「1010」であったとする．これは第 1 位と第 3 位の文書が適合であり（$x_i = 1$），第 2 位と第 4 位が不適合であること（$x_i = 0$）を意味する．この場合の平均精度は普通に計算すれば $(1 + 2/3)/2 = 5/6$ であるが，式 (1) を使うと，

$$\frac{1}{2} \times \left[\frac{1}{1} \times 1 + \frac{0}{2} \times (1+1) + \frac{1}{3} \times (1+1+0) + \frac{0}{4}(1+1+0+1) \right] = \frac{1}{2} \times \frac{3+2}{3} = \frac{5}{6}$$

となって確かに一致する．

式 (1) の別の表現として，より単純に，

$$\nu = \frac{1}{\sum_{i=1}^{N} x_i} \sum_{i=1}^{N} \left(\frac{x_i}{i} \sum_{k=1}^{i} x_k \right)$$

を考えることもできる．しかしこの場合には，x_i の 2 乗の項が数式内に出てくるため，x_i を確率変数に拡張する場合などにおいて，式 (1) のほうが計算に便利である．よって本稿では式 (1) を採用する．

なお式 (1) 中の $\sum_{i=1}^{N} x_i$ は N 件中に含まれる適合文書の総数を意味し，本稿ではこれを R と表記することがある．すなわち，

$$R = \sum_{i=1}^{N} x_i \tag{2}$$

である．ただし，この R については，テストコレクションによる検索実験の場合には

$$R \neq \sum_{i=1}^{N} x_i \tag{3}$$

となることがある．TREC や NTCIR のようなテストコレクションの場合，後述するように，複数の手法による検索結果の一部を併合することによって適合文書を洗い出している．また，各手法の評価はたとえば上位文書 1,000 件のように一律に固定してなされる（すなわち $N = 1000$）．この状況ではある手法による実行結果の上位 N 件中にすべての適合文書が含まれるとは限らない．つまり，他の検索手法で見つかった適

☆ Text REtrieval Conference, http://trec.nist.gov/
☆☆ NII/NACSIS Test Collection for Information Retrieval, http://research.nii.ac.jp/~ntcadm/

表 1 平均精度の最小値
Table 1 Minimum values of average precision.

N	R					
	5	10	30	50	100	500
10	0.354	1.000				
20	0.161	0.331				
30	0.105	0.206	1.000			
40	0.078	0.149	0.550			
50	0.062	0.117	0.399	1.000		
100	0.030	0.057	0.173	0.312	1.000	
300	0.010	0.019	0.053	0.090	0.191	
400	0.008	0.014	0.040	0.067	0.138	
500	0.006	0.011	0.032	0.053	0.108	1.000
1,000	0.003	0.006	0.016	0.026	0.052	0.307

表 2 無作為出力の場合の平均精度の期待値
Table 2 Expected values of average precision in the case of random output.

N	R					
	5	10	30	50	100	500
10	0.607	1.000				
20	0.353	0.568				
30	0.253	0.402	1.000			
40	0.199	0.313	0.771			
50	0.164	0.257	0.629	1.000		
100	0.090	0.138	0.330	0.521	1.000	
300	0.034	0.050	0.116	0.181	0.345	
500	0.021	0.031	0.071	0.110	0.209	1.000
1,000	0.011	0.016	0.036	0.056	0.106	0.503

合文書がその検索手法での上位 N 件に含まれないことが起こりうる．この場合には式 (3) が成り立つ．

このことは，R の定義自体が状況によって異なる可能性のあることを意味している．したがって，

$$\nu = \frac{1}{R}\sum_{i=1}^{N}\frac{x_i}{i}\left(1+\sum_{k=1}^{i-1}x_k\right) \qquad (1)'$$

と表記しておいて，この式の外側で R を定義する形のほうが一般には便利かもしれない．なお，本稿では，式 (1)' において $R \leq N$ をつねに仮定するものとする．

2.2 平均精度の性質

2.2.1 最大値と最小値

平均精度の最大値は 1.0 である．R 件の適合文書が不適合文書を間に挟むことなく第 1 位から連続すれば，$\nu = 1.0$ となる．

一方，式 (2) を仮定すると最小値は 0.0 ではなく，N と R とに依存し，

$$\nu_{\min} = R^{-1}\sum_{k=1}^{R} k/(N-R+k) \qquad (4)$$

である．この式は，$N = 4$ かつ $R = 2$ での最悪の出力「0011」の平均精度が $(1/3+2/4)/2 = 5/12$ となる事例を考えれば容易に導出できる．

それに対して，テストコレクションによる検索実験において式 (3) が成立する場合，平均精度の最小値は 0.0 となることがある．これは，他の検索手法で適合文書が発見され，$R > 0$ であるにもかかわらず，当該手法による上位 N 件中に適合文書がまったく含まれていない場合である．

参考として，式 (4) を使って，適当な N と R に対して式 (1) の最小値を求めた結果を**表 1** に示す．

2.2.2 無作為出力での期待値

R 件の適合文書を含む N 件の文書集合から無作為に 1 件ずつ文書を抽出して順番に並べ，それに対して平均精度を計算することを考える．つまり，検索システムを何ら用いずに文書を無作為出力したときに，どの程度の平均精度が得られるかを調べてみる．

その期待値を $E_{RAN}(\nu|N,R)$ と表記する．これは，無作為出力を無限回繰り返したときのそれらの平均精度の平均に相当する．この期待値は式 (1) に基づいて理論的に求めることができて，

$$E_{RAN}(\nu|N,R) = \frac{R-1+N^{-1}\sum_{i=1}^{N}\frac{N-R}{i}}{N-1} \qquad (5)$$

となる（証明は付録を参照）．参考として，式 (5) を使って適当な N と R に対して期待値を求めた結果を**表 2** に示す．表 2 が示すように，R が大きい場合，検索システムを使わずに単に文書を無作為に並べるだけでかなり高い平均精度が得られる可能性がある．

2.2.3 平均精度の変化量

次に，適合文書の順位が移動したときの平均精度の変化量について考える．たとえば出力結果が「11000...」から「10100...」へ変わった場合の平均精度の変化の程度を調べるのがここでの目的である．

適合文書が他の適合文書と入れ替わった場合には平均精度は何ら変化しないので，適合文書と不適合文書の順位の交換のみを議論すれば十分である．これらのうちの片方の文書の順位を r と書く．2 つの文書の順位が入れ替わるだけなので適合文書総数 R 自体は変化しないから，この項を省略したうえで，x_r が 0 から 1 に変化したときの変化量（差分）Δ を求めると，式 (1)' を使った簡単な計算から，

$$\Delta = r^{-1}\left(1+\sum_{k=1}^{r-1}x_k\right)+\sum_{t=r+1}^{N}t^{-1}x_t \qquad (6)$$

を得る．式 (6) は不適合から適合への変化による増加分を表す．また逆に，式 (6) は，適合から不適合への

変化による減少分と考えることもできる．そこで，入れ替わった 2 つの文書のそれぞれの順位ごとに式 (6) を計算し，増加分から減少分を差し引いて，最終的に R で割れば，平均精度の変化量が求められる．

式 (6) の右辺第 2 項から，第 r 位で適合または不適合が反転すると，その文書だけでなくそれ以降の第 $r+1$ 位から第 N 位までの間に存在する適合文書も平均精度の大きさの変化に影響することが分かる．これは順位の高い位置での適合/不適合の反転ほど平均精度の変化量が大きくなることを意味している．

したがって，不適合文書が上位に位置するほど平均精度の値の「減点」の度合いは大きい．実際，$N = 10$, $R = 4$ の場合に「1110000001」の平均精度は 0.85,「1011100000」は約 0.80 であり，10 位まで見なければすべての適合文書を発見できない場合（前者）の平均精度が，5 位まででですべて見いだせる場合（後者）のそれを上回る．この原因は後者では第 2 位という高位に不適合文書が出力されてしまっていることにある．この例は式 (6) に示されている平均精度の特徴をよく表している．

2.3 MAP の定義

検索実験では検索課題は 1 つではなく，複数個用意される．そこで平均精度 ν に添え字を付けて，ν_h を第 h 番目の検索課題に対する平均精度とする．また検索課題の総数を L と書く（すなわち $h = 1, \ldots, L$）．MAP は L 個の検索課題に対する ν_h の平均であり

$$\bar{\nu} = L^{-1} \sum_{h=1}^{L} \nu_h \tag{7}$$

で定義される．この統計量については後で詳しく議論する．

3. 平均精度による性能比較のための統計的検定とそれに影響する要因

3.1 検索実験による性能比較のための統計的検定

検索実験で用意される検索課題は，当然，想定されうるすべての検索要求を網羅したものではない．したがって，検索実験によって手法間の性能を統計的に比較する場合にはこれらの検索課題の集合を母集団から抽出された標本としてとらえる必要がある．実際に Tague-Sutcliffe と Blustein [3] はこの線に沿って TREC-3 の実行結果に対して分散分析を試みている．

検索課題の集合を単純無作為抽出による標本とすれば，MAP 式 (7) は平均精度 ν_h の母平均の推定量に相当する．そして，ある 2 つの手法間の性能の統計的な比較とは，結局，それらの母平均に差があるかどうかを統計量 $\bar{\nu}$ によって検定することにほかならない．この 3.1 節では，これ以降の節への準備として，この検定方法についての初等的な統計学の知識について簡単に整理しておく．

まず最も簡単な方法は平均値の差の検定を用いることである．2 つの手法を A と B，それぞれの平均精度の平均（MAP）を $\bar{\nu}_A$ と $\bar{\nu}_B$，平均精度の標本分散を s_A^2 と s_B^2 と書く．たとえば，

$$s_A^2 = (L-1)^{-1} \sum_{h=1}^{L} (\nu_{hA} - \bar{\nu}_A)^2$$

である．ここで ν_{hA} は h 番目の検索課題に対する手法 A による平均精度である．

両者の標本の大きさ（$= L$）が等しければ，帰無仮説「2 つの手法の平均精度の母平均の差は 0 である」の下に

$$t_1 = (\bar{\nu}_A - \bar{\nu}_B) / \sqrt{s_A^2/L + s_B^2/L} \tag{8}$$

は自由度 $2L-2$ の t 分布に従うので，これを利用して検定できる．ただし，正規母集団と等分散の仮定が必要である（なお等分散を仮定しない，式 (8) とは別の計算方法も提案されている）．あるいは s_A^2 と s_B^2 がそれぞれの母分散に等しいと仮定できれば，検定量 t_1 を標準正規分布表と比較すればよい．

別の考え方として，TREC や NTCIR などの検索実験の場合にはつねに同一の検索課題に対する 2 つの平均精度を比較することになるので，これを対標本（paired data）ととらえる場合がある [4]．このときには，検索課題ごとの手法間の差 $\nu_{hA} - \nu_{hB}$ 自体を標本として考えることになる．つまり，$u_h = \nu_{hA} - \nu_{hB}$ のように定義し，データとしては $u_1, \ldots, u_L$ の L 個の数値が得られたものとする．変数 u_h の平均は，

$$L^{-1} \sum_{h=1}^{L} u_h = L^{-1} \sum_{h=1}^{L} (\nu_{hA} - \nu_{hB}) = \bar{\nu}_A - \bar{\nu}_B$$

となり，やはり 2 つの手法間の MAP の差である．一方，標本分散は簡単な計算から $s_A^2 + s_B^2 - 2Co\nu_{AB}$ となる．ここで

$$Co\nu_{AB} = (L-1)^{-1} \sum_{h=1}^{L} (\nu_{hA} - \bar{\nu}_A)(\nu_{hB} - \bar{\nu}_B)$$

である（ν_{hA} と ν_{hB} との共分散）．したがって，帰無仮説「2 つの手法の平均精度の差の母平均は 0」の下に，正規母集団を仮定すれば，

$$t_2 = (\bar{\nu}_A - \bar{\nu}_B) / \sqrt{s_A^2/L + s_B^2/L - 2Co\nu_{AB}/L} \tag{9}$$

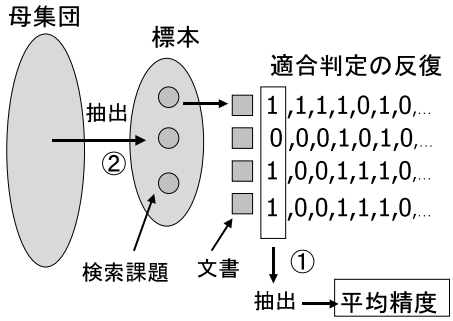

図 1 平均精度の計算における標本抽出
Fig. 1 Sampling for estimating average precision.

が自由度 $L-1$ の t 分布に従うことになる．もし $Co\nu_{AB} > 0$ ならば t_2 は t_1 よりも大きくなるので，同一のデータに対しては，式 (8) より式 (9) を使ったほうが帰無仮説は棄却されやすくなる（有意差が出やすい）．なお，この方法は「対応のある場合の平均値の差の検定」と呼ばれることがある．

3.2 適合判定の変動の影響

平均精度を計算するには適合判定によって各 x_i の値（1 か 0）を決めなければならない．この適合判定は本来，非常に主観的なものであり，判定者や判定状況に依存して変動することが知られている[5]．すなわち，MAP による母平均の推定の誤差には，①適合判定と②標本抽出とに起因する 2 種類の誤差が含まれる可能性がある．

この状況を**図 1** に示す．まず母集団からいくつかの検索課題が抽出される（図中の②）．この抽出の結果として母集団すべてを調査できないことに起因する誤差が生じる．この誤差がいわゆる標本誤差（sampling error）であり，通常の統計学で主に扱われているものである．さらに，抽出された各検索課題に対して，各文書が適合しているかどうかの判定がなされる．もしこの判定が変動すると仮定すれば，いくつかの判定結果の中の 1 つが抽出されて平均精度が求められることになり，ここからも抽出誤差が生じるわけである（図中の①）．

ここでは，前者①の適合判定の変動の影響を調べるために，その変動を x_i の測定の際の誤差ととらえて，統計学分野で開発された測定誤差モデルの適用を試みる☆．まず，標本抽出理論の教科書[6),7)]に沿って，次のモデルを導入する．

$$\nu_{ha} = \theta_h + e_{ha} \tag{10}$$

☆ 適合判定の変動を測定誤差ととらえるのはあくまで技術的な工夫であり，後述するように，判定の変動が常に測定誤差に起因していると考えるわけではない．

ここで，ν_{ha} は，第 h 番目の検索課題に対するある 1 つの検索手法による出力について，第 a 回目の適合判定結果を使って計算された平均精度の値とする．式 (10) はこの値が θ_h と e_{ha} とに分解されることを示している．これらは

θ_h：第 h 番目の検索課題に対する平均精度の真の値
e_{ha}：測定誤差（判定変動による誤差）

である（いずれも検索手法は 1 つに固定）．判定の変動がなければ $\nu_{ha} = \theta_h$ であるが，変動した場合には測定誤差の項 e_{ha} が付加されるというのがこのモデルの要点である．本研究ではこの e_{ha} を適合判定の変動に起因する誤差と考える．

検索課題 h を固定し，適合判定の反復（$a = 1, 2, \ldots$）に対して平均する意味での条件付き期待値 $E_m(\cdot|h)$ を考える．これを式 (10) に適用すると，θ_h は判定変動に影響されない真の値と仮定したので，

$$E_m(\nu_{ha}|h) = \theta_h + E_m(e_{ha}|h) \tag{11}$$

となる．さらに判定変動による誤差を

$$e_{ha} = \beta_h + d_{ha} \tag{12}$$

のように分解する．ここで β_h は判定変動による誤差の平均，すなわち $E_m(e_{ha}|h) = \beta_h$ である．誤差 e_{ha} から β_h を除いた d_{ha} は完全な偶然による変動と仮定できるので，平均すれば 0，すなわち $E_m(d_{ha}|h) = 0$ とおく．その結果，式 (11) は，

$$E_m(\nu_{ha}|h) = \theta_h + \beta_h + E_m(d_{ha}|h) = \mu_h \tag{13}$$

と書ける．ここで $\theta_h + \beta_h = \mu_h$ とおいた．μ_h は，h を固定して適合判定を無限回繰り返し，その結果得られる無限個の ν_h を平均することによって変動要因を除去したものと解釈できる．また，同様の操作で計算される分散を $V_m(\nu_{ha}|h) = \sigma_h^2$ と書く．

次に，h 個の検索課題から成る標本を S と表記する．標本 S が固定されたときの MAP $\bar{\nu}$ は適合判定の変動がなければただ 1 つに確定するが，変動すれば確率的となる．したがって上と同様に MAP に対する期待値 E_m を考える．これは，判定の反復に対する期待値を使って MAP を求めることに相当する．式 (7) による定義と期待値の線型性とを使えば，「検索課題間の適合判定は独立」という仮定の下で，

$$E_m(\bar{\nu}|S) = L^{-1}\sum_{h=1}^{L} E(\nu_{ha}|h) = L^{-1}\sum_{h=1}^{L} \mu_h \tag{14}$$

を得る．検索課題ごとに判定者が異なれば，この仮定は十分に妥当であろう．

式 (14) は標本 S を固定しているので，図 1 の抽出①のみについての期待値に相当する．そこで次に抽出

②に関する期待値を考えなければならない．この期待値を $E_p(\cdot)$ と表記する．これは，母集団から抽出される可能性のあるすべての標本に対してある統計量をそれぞれ計算し，その平均をとる操作を意味する．式 (14) を使えば，

$$E_p[E_m(\bar{\nu}|S)] = E_p\left(L^{-1}\sum_{h=1}^{L}\mu_h\right) = L^{-1}\sum_{h=1}^{L}E_p(\mu_h) \quad (15)$$

となるが，統計学の初等的な結果から，$E_p(\mu_h)$ は μ_h の母平均に一致する．この母平均を μ と表し，さらに簡略化のため $E_p[E_m(\cdot|S)] = E_{pm}(\cdot)$ と書くと，結局，

$$E_{pm}(\bar{\nu}) = \mu \quad (16)$$

を得る．これは，判定変動を含んだ MAP は母平均 μ の不偏推定量であることを示している（真の値 θ_h の母平均に対する不偏推定量ではない点に注意）．

同様に，$V_p[V_m(\cdot|S)] = V_{pm}(\cdot)$ と表記して，これを計算すると，結果的に，

$$V_{pm}(\bar{\nu}) = L^{-1}(\sigma_d^2 + \sigma_\mu^2) \quad (17)$$

を得る（導出の詳細は付録に示した）．ここで σ_d^2 は検索課題を固定したときの判定変動による分散 σ_h^2 の母平均である．言葉を換えれば，標本平均

$$L^{-1}\sum_{h=1}^{L}\sigma_h^2 \quad (18)$$

によって推定される母集団の統計量である．また，σ_μ^2 は μ_h の母分散である．

適合判定が確率的に変動する場合，その判定結果から平均精度の標本分散 s^2 を単純に計算すると，それには σ_μ^2 に起因する分散だけなく，判定変動による σ_d^2 からの散らばりも含まれることを式 (17) は示している．もし，式 (10) 中の e_{ha}（もしくは d_{ha}）が，本当の意味での測定誤差であるならば，純粋な標本誤差である σ_μ^2 が σ_d^2 によって過大評価されることになる．たとえば，ある判定者が文書の重要な部分を読み落としてしまったなどの誤りだけが判定の変動の原因だったと仮定する．2 つの手法間の差の検定はこのような誤差を除去して σ_μ^2 のみに基づいてなされるべきである．しかし，MAP の分散には測定誤差による散らばりが混入するため，その推定量である標本分散 s^2 もその分，大きくなってしまう．この結果，3.1 節で示した検定量 t_1 または t_2 が見かけ上小さくなり，手法間に有意差が出にくくなる可能性がある．

一方，すでに述べた，適合判定は本来主観的なものであり，確率的にしかとらえられないという主張を受け入れれば，σ_d^2 に対して別の見方ができる．たとえば，判定の際の誤りがまったくなくても，判定者の考え方の違いから，判定結果が本質的に異なる可能性がある．この場合には，e_{ha} は排除されるべきものではなく，そのような適合性の確率的変動も考慮に入れたうえで，手法間を比較する必要がある．この立場では，式 (17) は標本分散 s^2 の過大評価を意味しない．σ_μ^2 のみが考慮される場合よりも検定量 t_1 または t_2 は当然小さくなるが，それは単に，適合判定の確率的な変動によって，検定の精度が低くなることを示しているのにすぎない．

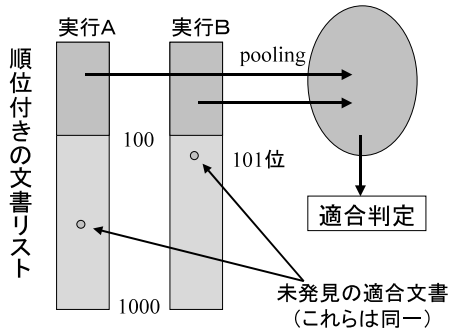

図 2 poolingによる適合判定のしくみ
Fig. 2 System of pooling for relevance judgment.

3.3 poolingの方法論的な問題に起因する平均精度の変動

大規模な文書データベースを検索実験に利用する場合には検索課題に対するすべての適合文書を発見するのは難しく，そのために pooling の方法が採用されることが多い．これは，数多くの検索手法による同一検索課題に対する各検索結果の上位 x 件を併合し，それらに対してのみ適合判定を施す方法である．適合判定に必要な労力と時間が大幅に軽減される反面，あくまで近似的な方法であって，すべての適合文書が発見されていない可能性がある．

たとえばある検索実験プロジェクトにおいて，その参加チームにそれぞれ検索の実行ごとに上位 1,000 件の文書を提出してもらい，そのうちの各上位 100 件を取り出して適合判定を実施したとする．この場合，ある手法による検索実行の 101 位の文書 d が実は適合文書であり，なおかつこの文書が他の手法で 100 位以内に入っていなければ，この適合文書は未発見のまま埋もれることになる．**図 2** はこの状況を示している．

実際には，平均精度の計算は提出された上位 1,000 件に対してなされ，その中で未発見の適合文書は不適合として処理される．したがって，もしこの文書 d が適合していることが事後的に発見されたならば，そ

表 3 第 101 位に適合文書が発見された場合の平均精度の変化量
Table 3 Change of average precision after finding a new relevant document at the rank 101.

適合文書総数	平均精度		
	0.1	0.3	0.5
10	0.00081	−0.01737	−0.03555
50	0.00794	0.00402	0.00010
100	0.00891	0.00693	0.00495

れによって平均精度は変化することになる．その変化量 Δ は，

$$\Delta = (R+1)^{-1}\left[r^{-1}\left(1+\sum_{k=1}^{r-1}x_k\right)+\sum_{t=r+1}^{N}t^{-1}x_t\right] - (R+1)^{-1}\bar{\nu} \quad (19)$$

で与えられる．ここで R と $\bar{\nu}$ はそれぞれ文書 d が発見される以前の適合文書総数と平均精度である．r は適合文書が発見された順位で，ここでは $r = 101$ となる．

式 (19) は以下のように導かれる．まず右辺第 1 項は文書 d の発見による平均精度の増加分であり，式 (6) を直接使っている．一方，右辺第 2 項は，適合文書数が R から $R+1$ に変化した結果として，文書 d に関連しない部分で平均精度が減少する分に相当し，

$$\bar{\nu} - R\bar{\nu}/(R+1) = [R^{-1} - (R+1)^{-1}]R\bar{\nu} = (R^2+R)^{-1}R\bar{\nu} = (R+1)^{-1}\bar{\nu}$$

として得られる．

もし第 $r+1$ 位以下に適合文書が存在しなければ，$1+\sum_{k=1}^{r-1}x_k = 1+R$, $\sum_{t=r+1}^{N}t^{-1}x_t = 0$ であるから，式 (19) はさらに簡単に

$$\Delta = r^{-1} - (R+1)^{-1}\bar{\nu} \quad (20)$$

となる．この式を使って，第 101 位の文書が適合していることが事後的に発見されたときの平均精度の変化量を計算すると**表 3** のようになる．なお，R と $\bar{\nu}$ は適当に選んだ．

表 3 が示すように，場合によっては 101 位の適合文書 d の発見により平均精度が下がることもある．全体的に平均精度の変動は小さく，したがって MAP での検定に大きな影響は与えにくい．これは，すでに上で議論した，上位の文書の適合/不適合の変化が相対的に大きく影響するという平均精度の性質に起因している．

ここでは，表 3 は 101 位に未発見の適合文書が存在した場合だけの数値であるが，上位 100 件を併合する pooling では，この場合の平均精度の変化量が最も大きく，その発見の順位が 102 位，103 位と下がるにつれて，その変化量は減少していく．ただし，平均精度の変化量が負になる場合にはこの限りではない．このことは式 (20) から明らかである．たとえば，平均精度が 0.5，適合文書総数が 10 件の場合に，102 位で適合文書が新たに発見されたとすると，その平均精度の減少分の絶対値は 101 位の場合よりも大きくなる．この減少分の上限値は，式 (20) において $R = 1$ かつ $\bar{\nu} = 1.0$ で，$r \to \infty$ とした場合の -0.5 である．もっとも，このような極端な例が起こる確率は小さいと予想される．

なお，表 3 はあくまで 1 つの例示にすぎず，実際には，比較対象の他の手法の平均精度もまた式 (19) に従って変化する．したがって，実際には，未発見の適合文書の影響は状況に依存し，かなり複雑であることに注意しなければならない．

4. 実際のデータを用いた分析

4.1 使用データ

ここでは，上で議論した MAP による統計的検定の計算例を実際のデータを用いて示す．使用するデータは NACSIS（現：国立情報学研究所，NII）による日本語テストコレクション NTCIR-1 を使った検索実験の結果である．今回は，California 大学 Berkeley 校による 2 つの実行結果 BKJJBIDS と BKJJDCFU とを選び，それらの比較評価を検討の材料とする．この事例を通じて，本稿のこれまでの議論に対する理解を深めることがここでの分析の目的である．

BKJJBIDS は索引作成方法として bigram を使用した実行結果であり，一方 BKJJDCFU は辞書との最長一致法を用いている[8]．ともにロジスティック回帰型検索モデルを使っている．ただし前者 BKJJBIDS は検索課題の ⟨narrative⟩ フィールドを使用していない☆．なお以下の計算に使う検索課題は No.31〜No.83 までの 53 件である．すなわち，ここで示す計算例における標本の大きさは 53 である（$L = 53$）．

4.2 適合判定の変動の影響についての分析

4.2.1 シミュレーションによる解法と手順

3.2 節で議論した測定誤差モデルを実際に活用するにはさらに μ_h と σ_h^2 を求めるためのモデルが必要である．これらは上で定義したように，検索課題を固定して，各文書に対する適合判定を無限回繰り返した場合の平均精度の平均と分散である．

本稿では 2 値での適合判定を仮定しているので，第 i

☆ すなわち BKJJBIDS はいわば「短い検索質問」，BKJJDCFU は「長い検索質問」に対する実行結果ということになる．

位の文書の x_i に対して「1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, ...」のようなデータが判定の繰返しによって得られることになる（図 1 参照）．この繰返しが相互に独立ならば，これをベルヌーイ試行としてとらえることが可能である．すなわち，x_i が適合と判定される確率を p_i とおき，適合判定を確率的事象としてモデル化する（$i = 1, \ldots, N$）．なお，本稿では便宜上，p_i を「適合確率」と呼ぶ．

この結果，2 項分布に関する理論より，確率変数 x_i の期待値と分散はそれぞれ p_i と $p_i(1 - p_i)$ で与えられる．しかし，式 (1) に示したように平均精度の計算式は複雑なので，ベルヌーイ試行でモデル化しても $\mu_h = E_m(\nu_{ha}|h)$ の値を解析的に正確に計算することは難しい．これは，式 (1) には R^{-1} の項があり，さらに式 (2) に示すように，この R 自体に確率変数 x_i が含まれるために，μ_h が確率変数の比の形式になってしまうからである．つまり，式 (1) の R^{-1} の項を除いた部分を ν_1 と表記して，その期待値を計算してみると，

$$
\begin{aligned}
E(\nu_1) &= E\left[\sum_{i=1}^{N} i^{-1} x_i \left(1 + \sum_{k=1}^{i-1} x_k\right)\right] \\
&= \sum_{i=1}^{N} i^{-1} E(x_i + x_i x_1 + x_i x_2 + \ldots + x_i x_{i-1}) \\
&= \sum_{i=1}^{N} i^{-1} [E(x_i) + E(x_i x_1) + \ldots + E(x_i x_{i-1})] \\
&= \sum_{i=1}^{N} i^{-1} (p_i + p_i p_1 + \ldots + p_i p_{i-1}) \qquad (21)
\end{aligned}
$$

である（なお $i \neq j$ のとき x_i と x_j に対する判定の独立性，すなわち $E(x_i x_j) = E(x_i)E(x_j)$ を仮定）．一方，式 (2) より，R についても，

$$
E(R) = \sum_{i=1}^{N} p_i \qquad (22)
$$

であり，平均精度自体の期待値は式 (21) と式 (22) との比の形式になる．この形式は，統計学では比推定と呼ばれ，その標本分布は複雑であり，解析的に解くことは難しい（特に分散 σ_h^2 の計算が難しい）．

したがって，もし何らかの方法で p_i の値を決めることができれば，むしろ乱数を使ったシミュレーションによって μ_h と σ_h^2 とを推計したほうが簡便である．その具体的な手順を以下に示す．

① L 件の検索課題ごとにそれぞれ全文書に対して何らかの方法で適合確率 p_i（0.0〜1.0）を設定する（実際には各課題に対応した適合判定ファイルに含まれるすべての文書に対して適合確率を設定する）．

② ある 1 つの方法（たとえば BKJJBIDS）を選び，その L 件の検索課題に対する実行結果を用意する．

③ 1 つの検索課題に対して，次の (1)〜(2) の操作を M 回繰り返す．

(1) 実行結果の第 1 位から第 N 位までの文書に対して以下の操作を行う．(a) 文書ごとに 0.0〜0.1 の一様乱数を発生させ，(b) その乱数の値が p_i を超えなければ適合（$x_i = 1$），そうでなければ不適合（$x_i = 0$）と設定する．

(2) 上記 (1) の結果から平均精度を計算する．

④ 上記③の操作により，1 件の検索課題に対して M 個の平均精度が得られるので，その平均と分散を計算する．この平均が μ_h，分散が σ_h^2 に相当する．

⑤ 上記③〜④の作業を L 件の検索課題に対して行う．

⑥ 最終的に，μ_h（$h = 1, \ldots, L$）から，通常の標本分散を計算する方法によって σ_μ^2 を推計し，一方，σ_h^2（$h = 1, \ldots, L$）を平均して σ_d^2 の推計値とする．

4.2.2 実際の計算例

実際の計算を試みるには適合確率 p_i の設定が必要である．これに関しては，理想的には，たとえば 10 人の判定者を用意し，その 10 個の判定結果のうち「適合」とされた割合を適合確率の推計値とするのが合理的である（もちろん判定者の数は多いほどよい）．しかし，そのためには，最低でも，延べ数で 1,000 文書 × 53 課題 × 10 回 = 530,000 回もの判定が必要になり，実現は難しい．そこで，今回はテストコレクションとして利用可能な範囲のデータを使って近似的に適合確率を設定することとした．

NTCIR-1 の適合判定は 2 人の判定者によって 3 段階（正解，部分的正解，不正解）でなされている[9]．そして実際にはそれらの判定結果がすり合わされ，さらに 3 段階の判定を適合/不適合の 2 値に圧縮して平均精度が計算されている．もし 1 人のみの判定者が 2 値で判定をしていただけならば，本稿での適合確率の推計は不可能であるが，合併して圧縮される以前の元の判定結果を使えば，以下のように適合確率の近似的な推計が可能になる．

元の判定結果の場合には 1 件の文書に対する 2 人

の判定者の判定結果のパターンは組合せで 6 通りある．本稿ではそのパターンごとに**表 4** のように適合確率を設定する．この基本的な考え方は次のとおりである．まずはじめに「両者とも正解」ならば 1.0,「両者とも不正解」ならば 0.0 とする.「正解—不正解」ならば，それらの中間をとって $(1.0+0.0)/2=0.5$ とする.「部分的正解—不正解」はこの 0.5 よりも低く設定されなければならず，なおかつ「正解—不正解」の設定方法に従えば，その 2 倍は「両者とも部分的正解」の適合確率となるべきである．したがって，その候補はおおよそ 0.1〜0.4 となるが，今回は「部分的正解」を「正解」にかなり近いと仮定して「部分的正解—不正解」を 0.4 にした．そうすると,「両者とも部分的正解」が 0.8 となり，この値と「両者とも正解」1.0 から「正解—部分的正解」が高目の 0.9 になるからである．

適合確率として表 4 を使い，4.2.1 項で説明した手順に従って計算した結果を**表 5** に示す．この計算では，文書数は $N=1,000$ であり，反復回数については $M=100,000$ とした．また表 5 には，シミュレーションの結果だけでなく，比較のため，2 人の判定者それぞれについて，①「正解」と「部分的正解」の両方を適合文書とした場合と，②「正解」のみを適合文書とした場合との両方の判定結果によって計算される MAP とその分散（標本分散）も示した．表 5 中で「判定者 1」と表記されているのは，各検索課題で「第 1 判定者」とされた人々による判定結果から計算された数値であり，同様に「判定者 2」は「第 2 判定者」からの計算結果を意味している．

表 4 適合確率の設定例
Table 4 Example of relevance probabilities.

パターン	パラメータ
正解—正解	1.0
正解—部分的正解	0.9
正解—不正解	0.5
部分的正解—部分的正解	0.8
部分的正解—不正解	0.4
不正解—不正解	0.0

μ_h の標本平均，すなわち測定誤差モデルによって判定の変動部分が除去された MAP は，BKJJBIDS で 0.27973，BKJJDCFU で 0.32588 であり，いずれも確率設定の基礎となった 4 つの判定結果による MAP の平均に近い値になっている．これはもちろん，式 (13) から容易に予想できることであり，偶然変動 d_{ha} が平均の操作によって除去された結果として解釈できる．

一方，μ_h の標本分散による σ_μ^2 の推定値 $\hat{\sigma}_\mu^2$ は 0.05171 と 0.04558 であり，いずれも 4 つの判定結果の標本分散よりも小さい．そして判定変動 σ_h^2 の標本平均としての σ_d^2 の推定値 $\hat{\sigma}_d^2$ はそれぞれ 0.00188 と 0.00299 であった．この結果は，すでに議論したように，判定結果から直接計算される MAP には判定変動による分散が含まれていることを意味している．

ただし，σ_d^2 の影響は σ_μ^2 に比べてそれほど大きくはない．実際，$\hat{\sigma}_d^2+\hat{\sigma}_\mu^2$ はそれぞれ 0.05359 と 0.04857 であり，これらに対する $\hat{\sigma}_d^2$ の構成比は 3.5%と 6.2%にすぎない．

次に，3.1 節の議論に基づいて，ここでの 2 つの検索実行の間での検索性能の有意差を調べるための検定を試みる．結果を**表 6** に示す．表 6 では，式 (8) による検定量 t_1 と式 (9) による検定量 t_2 をそれぞれ計算し，標準正規分布と t 分布のそれぞれの両側確率を示してある．また，測定誤差モデルによるシミュレーションに関しては，変動要因 $\hat{\sigma}_d^2$ を除いて $\hat{\sigma}_\mu^2$ のみで t_1 と t_2 を計算した結果と，除かずに計算した結果とを表示した．なお，測定誤差モデルにおける $u_h=\nu_{hA}-\nu_{hB}$ の標本分散は，表 5 と同様の方法を使って，$M=100,000$ の反復で計算してある．これは式 (10) における ν_{ha} を u_{ha} に置換してそのまま測定誤差モデルを適用した結果である．

まず，検定量 t_1 と t_2 とを比較すると，対標本とし

表 5 適合判定の変動の影響を調べるためのシミュレーションの結果
Table 5 Results of simulation for examining variations of relevance judgments.

手法	統計量	測定誤差モデルでのシミュレーション	判定者 1		判定者 2	
			正解+部分的正解	正解のみ	正解+部分的正解	正解のみ
BKJJ BIDS	標本平均（MAP）	0.27973	0.29042	0.27685	0.28930	0.27984
	標本抽出の分散 $\hat{\sigma}_\mu^2$	0.05171	—	—	—	—
	判定変動の分散 $\hat{\sigma}_d^2$	0.00188	—	—	—	—
	標本分散	—	0.05593	0.05231	0.05288	0.06297
BKJJ DCFU	標本平均（MAP）	0.32588	0.34901	0.33558	0.33726	0.31579
	標本抽出の分散 $\hat{\sigma}_\mu^2$	0.04558	—	—	—	—
	判定変動の分散 $\hat{\sigma}_d^2$	0.00299	—	—	—	—
	標本分散	—	0.05401	0.05545	0.04893	0.05346

表 6 手法間の性能比較のための MAP の差の検定結果（53 件の検索課題による結果）
Table 6 Results of statistical tests for comparing MAP between two methods.

統計量 手法 A：BKJJDCFU 手法 B：BKJJBIDS	判定者 1		判定者 2		測定誤差モデル	
	正解＋部分	正解のみ	正解＋部分	正解のみ	変動要因除去	変動要因含む
$\bar{\nu}_A - \bar{\nu}_B$（MAP の差）	0.0586	0.0587	0.0480	0.0360	0.0462	
統計量 t_1：式 (8)	1.2866	1.3025	1.0583	0.7671	1.0773	1.0513
両側確率 $P(t_1 < \|z\|)$：標準正規分布	0.1982	0.1928	0.2899	0.4430	0.2814	0.2931
両側確率 $P(t_1 < \|t\|)$：t 分布^{*1}	0.2011	0.1956	0.2924	0.4448	0.2838	0.2956
$\nu_{hA} - \nu_{hB}$ の標本分散	0.0422	0.0337	0.0446	0.0349	0.0330^{*3}	0.0378^{*3}
統計量 t_2：式 (9)	2.0763	2.3304	1.6531	1.4006	1.8503	1.7272
両側確率 $P(t_2 < \|z\|)$：標準正規分布	0.0379	0.0198	0.0983	0.1613	0.0643	0.0841
両側確率 $P(t_2 < \|t\|)$：t 分布^{*2}	0.0428	0.0237	0.1043	0.1673	0.0700	0.0901

注：∗1 自由度は 104，∗2 自由度は 52，∗3 この数値は 4.2.1 で示したシミュレーションによって求めた．なおこの場合 $\hat{\sigma}_d^2 = 0.00487$.

てとらえた場合の t_2 のほうが大きく，両側確率が小さいことが分かる．測定誤差モデルによるシミュレーションでは 2 つの実行結果の MAP の差は約 4.6%であるが☆，統計量 t_1 の場合には両側確率が正規分布と t 分布の両方ともに 30%に近く，有意水準 5%からはかなり遠い．それに対して，t_2 の場合には両側確率が格段に小さくなり，シミュレーションの結果の場合には両側確率は 10%以下になる．特に，シミュレーションで判定の変動要因を除いた場合には，t_1 の場合で t 分布の両側確率は 0.2838 であるのに対して，t_2 の場合には 0.0700 で約 1/4 である．ちなみに標本分散を固定すれば，t_1 の場合には MAP の差が 7.8%まで開くとようやく両側確率が約 7%になる．

次に，シミュレーションの結果において変動要因を除いた場合と含めた場合の差に着目する．変動要因を除いた場合，両側確率が約 1〜2%小さくなる．特に，対標本を仮定した t_2 の場合にその影響は大きく，t 分布の両側検定で，変動要因を含んだ場合には 0.0901 であるのに対して，除去した場合には 0.07 となっている．これは，$\nu_{hA} - \nu_{hB} = u_h$ を変数と見なした場合には，$\hat{\sigma}_d^2$ の影響が相対的に大きくなるからである．表 6 の注に示したように，u_h を変数とした場合 $\hat{\sigma}_d^2$ は 0.00487 であり，一方 $\hat{\sigma}_d^2 + \hat{\sigma}_\mu^2$ は 0.0378 なので，変動要因の構成比は約 12%となっている．表 5 の結果に比べれば，対標本の場合の変動要因の影響は大きいことが分かる．

なお，変動要因を除去した場合の t_2 の t 分布での両側確率は 0.0700 であるが，これが 0.05（すなわち有意水準 5%）以下になるには，標本分散を固定すると，MAP の差が約 5.01%必要である．それに対して，変動要因を含めた場合には，有意水準 5%以下を達成するには約 5.37%の開きが必要になる．この数字だけを見れば，やはり変動要因の影響はさほど大きくないといえるかもしれない．

ところで，3.1 節で述べたように，t 検定を正しく適用するには正規母集団の仮定が必要である．そこで，標本での平均精度の値の経験的な分布が正規分布になっているかどうかを調べてみる．ここでは，測定誤差モデルを使って計算された 2 つの手法の間の差のみについてプロットを試みる（**図 3**）．この図は 53 件の検索課題ごとの $u_h = \nu_{hA} - \nu_{hB}$ に関する $\mu_h = E_m(u_{ha}|h)$ の値を昇順に並べて通し番号 $n = 1, \ldots, 53$ を与え，第 n 位の μ_h については x 座標を μ_h，y 座標を $n/53$ としてプロットしたものである．図中の正規分布の曲線は，表 5 に示されている u_h の平均（$= \bar{\nu}_A - \bar{\nu}_B$）と標本分散とを用いて，$\mu_h$ を変数とする累積正規分布を計算して描いた．ただし測定誤差を除去した場合を使ってある．図からは標本中の μ_h の分布が正規分布に一致しないまでも比較的近いことが読み取れる．

4.3 pooling による適合判定の過程で発見されなかった適合文書の影響の分析

ここでは，3.3 節で議論した pooling による適合判定の結果として未発見のまま「埋もれてしまった」適合文書の影響について，実際のデータを用いた分析を試みる．その主目的は，2 つの手法間の MAP による差の検定において，もし埋もれていた適合文書が新たに発見されたならば，発見以前と比較して検定結果がどのように変化するかを調べることにある．

実際にこのような適合文書を見いだすことは難しいため，ここではきわめて仮想的な計算を試みる．すなわち，2 つの実行結果のうち，MAP の値が低い BKJJBIDS における第 101 位の文書の適合確率を 1.0 に強制的に置き換えて，平均精度を計算してみる．4.2

☆ 4.6%とは平均精度の最大値 1.0 を 100%としたときのパーセント表示であり，実際には 0.046 を意味する．以下，同様．

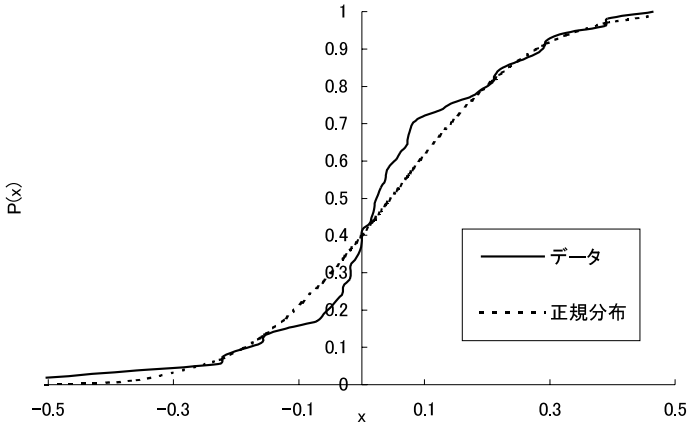

図 3 標本データの経験的分布：$u_h = \nu_{hA} - \nu_{hB}$
Fig. 3 Empirical distribution of sample data.

節で見たように，シミュレーションでは，BKJJDCFU の MAP は BKJJBIDS のそれに比べて約 4.6%ほど高く，差の検定の結果でもかなりの有意差がある．ここで，BKJJBIDS のほうに未発見の適合文書が数多く含まれていたと仮定し，その結果 MAP の大きさが逆転するのか，あるいはその差がどの程度縮まるのかを 4.2.1 項で説明したシミュレーションによって調べてみる．なお，3.3 節で議論したように,「101 位」とは，上位 100 件の併合という仮定の下では，適合文書の新規発見による平均精度の変化量が正となる場合において，最もその量が大きくなる順位である．したがって，101 位を強制的に適合に読み替えるということは,「未発見」がその実行結果にとって不利になるという意味での「最悪」の場合を分析しようとしていることになる（もちろん，表 3 について議論したように，状況によってはその変化量が負になることもありうる）．

53 件の検索課題に対して BKJJBIDS による第 101 位の文書を調べたところ，そのうちの 51 件で適合確率が 0.0 であった．最も極端な設定として，これらの 51 文書の適合確率をすべて 1.0 に置き換える．そのうえで，上と同様に反復回数 10 万回のシミュレーションを試みた結果，μ_h の平均（MAP）は 0.26803 となり，表 5 の結果と比較して約 0.012 の減少（約 4%）となった．3.3 節で議論したように，下位で適合文書が発見された場合，逆に平均精度が減少する場合があり，このシミュレーションの結果でも MAP は小さくなった．

BKJJBIDS において適合確率を置き換えた検索課題 51 件については適合文書総数がそれぞれ 1 件増えるので，他方の BKJJDCFU では R を $R+1$ に強制的に置き換えて MAP を再計算しなければならない．この操作を加えて，上と同様に反復回数 10 万回のシミュレーションを試みたところ MAP は 0.30619 となった．これは表 5 の結果と比較して，約 0.020 の減少（約 6%）である．

ここでは，適合文書数を 1 件増やした以外には，BKJJDCFU での計算手順に何ら変更を加えていない．もし BKJJBIDS で適合確率を 1.0 に置き換えた文書が BKJJDCFU の上位 1,000 件に含まれていたとしても，その適合確率は 0.0 のまま計算してある．したがって，BKJJDCFU の MAP の値が減少するのは当然である．しかし，すでに述べたように BKJJBIDS も同時に減少しており，MAP の逆転は起こらなかった．また，BKJJBIDS の減少によって両者の差はそれほど縮まっていない．やはりこの計算例でも 3.3 節での議論と同様に 101 位での適合文書の発見は平均精度の計算結果に大きな影響を及ぼさない．

ただし，MAP が逆転しないまでも，両者の差が縮小しているのは事実であり，その検定結果への影響を調べる必要がある．ここでの計算結果では両者の MAP の差は 0.0382 であり，表 6 の結果 0.0462 と比較して，0.008 の減少である．これは 0.0462 に対して約 17%の減少に相当する．この減少幅が検定量 t_2 に与える影響を見るために，未発見の適合文書を上と同様に設定して $\nu_{hA} - \nu_{hB}$ に対するシミュレーションを試みたところ，$\hat{\sigma}_\mu^2$ は 0.0294 となった．当然，これも表 6 の 0.0330 より小さい（約 11%の減少）．結果的に，判定の変動要因を除去して計算した検定量 t_2 は 1.6209 となり，これを表 6 の結果 1.8503 と比べれば，約 12%の減少である．両側確率は t 分布の場合に約 0.11 となり，0.07 に比べて約 0.04 増加する．このことは，検

定結果における有意差が適合文書の新発見に対しては比較的敏感に反応することを意味する．つまり，有意水準 5%には達しないまでも確率 0.07 程度の有意差があると考えていたものが，実はそれは確率 0.11 程度の差でしかなかった，というケースをこの例は示している．

もちろん，ここでの設定例はかなり極端であり，現実的には，検索課題 53 件中の 51 件において片方の実行結果の 101 位に未発見の適合文書が存在し，なおかつそれらが比較対象の他方の実行結果の上位 1,000 件にまったく含まれないという状況はほとんど考えれらない．しかし，可能性としては否定できないわけであり，テストコレクションを製作する際に何らかの注意が払われるべきかもしれない．

5. 考　　察

以上の議論と分析結果をまとめると次のようになる．

① 対標本を仮定した場合の検定量は通常の平均値の差の検定量よりも大きくなり，したがって，有意差が出やすい．
② 判定の変動による分散の影響は相対的には小さい．
③ 未発見の適合文書の影響はかなり小さい．
④ しかし，②および③に関して，有意差の程度という点では，手法間の差の検定結果へある程度の影響を及ぼす．

適合判定の変動や未発見の適合文書の影響は手法間での MAP の大小関係を逆転させるほど大きなものではないというのが妥当な結論であろう．もちろん，MAP にほとんど差がないような，性能の接近した手法間では逆転が生じる可能性はある．しかし，それは，TREC や NTCIR のような検索実験における手法の全体的な順位付けに大局的な影響を及ぼすほどのものではない．本稿で取り上げた 2 つの手法間の MAP の差は約 4.6%であったが，この程度の差ならば，順位が逆転する気配はまったくない．

この結果は，複数の判定者間の判定結果の差異の影響を実証的に分析した栗山ら[9]と矛盾しない．また，平均精度が評価指標の主流となる以前における適合判定の変動についての研究結果（Burgin[10]を参照）にもほぼ一致する．その最大の要因は，検索実験では複数の検索課題が用意され，それらの平均として評価指標が算出されることにある．検索課題の数が多くなれば，適合判定の変動が特定の検索手法に有利あるいは不利に働く可能性が小さくなる．結局，変動の影響は全体に及び，結果的に，手法間の順位付けのような大局的な分析には大きな変化をもたらさない（未発見の適合文書に関しては後述）．

むしろ，本稿の議論・分析からの重要な帰結は上記の④かもしれない．すなわち，適合判定の変動にともなう統計量の分散が 2 つの手法間の性能の差に関する検定結果へ与える影響である．たとえば，t 分布の両側確率が本来ならば 0.07 程度の差が判定の変動による誤差分を除去すると実は 0.05 であり，有意水準 5%で差が出てくるというような状況が生じうることが本稿の議論によって明らかとなった．もっとも，情報検索の実験において，「有意水準 5%」が特に絶対的な基準というわけではなく，もともと「5%」や「1%」は 1 つの目安にすぎない．しかし，検索手法の開発・改良において手法間の有意差は重要な手がかりであり，判定の変動がこれに関する検定結果に大きな影響を与えるとすれば，この問題をきちんと考慮する必要がある．

たとえば，ある研究者がテストコレクションを使って検索手法の改良を試みたときに，MAP がどの程度上昇すれば，その改良に意味があるかどうかを知りたいとする．理論的根拠はないが，1 つの目安はやはり「有意水準 5%での差」であろう．対標本を仮定した場合に，手法間の MAP にどの程度の差があれば有意水準 5%に達するかを計算するには，次の式を使えばよい．

$$y = \sqrt{s^2(1-K)/L} \times P_t^{-1}(0.05, L-1) \quad (23)$$

ここで，

s^2：変数 $u_h = \nu_{hA} - \nu_{hB}$ の標本分散
K：s^2 において判定の変動の分散が占める割合
L：検索課題数（標本の大きさ）
$P_t^{-1}(\alpha, F)$：t 分布の逆関数☆，α は両側確率，F は自由度
y：MAP の差

である．K は，判定の変動を測定誤差としたときに，それを除去するための定数である．たとえば 4.3 節の分析ではこの値は約 0.12 であった．

K として 0.00，0.05，0.10，0.15 の 4 つを選び，s^2 を 0.01，0.03，0.05，0.07，0.09 の 5 段階に設定して式 (23) を計算した例を**表 7** に示す．なお，検索課題数は 30，50，100，150 とした．

4.2 節で議論したように判定の変動を完全な誤差と見なす根拠はない．むしろ適合判定が本質的に確率的なものであるとすれば，K を 0.00 に設定するのが妥

☆ この逆関数は市販の表計算ソフトウェアで簡単に計算できる．本稿では Microsoft 社の Excel を使用した．

表 7 有意水準 5%を達成するための MAP の差
Table 7 Differences of MAP at the significance level 0.05.

標本分散	検索課題数 30	50	100	150
(a) 測定誤差の割合：0%				
0.01	0.0374	0.0285	0.0199	0.0162
0.03	0.0647	0.0493	0.0344	0.0280
0.05	0.0835	0.0636	0.0444	0.0361
0.07	0.0988	0.0752	0.0525	0.0427
0.09	0.1121	0.0853	0.0596	0.0485
(b) 測定誤差の割合：5%				
0.01	0.0364	0.0278	0.0194	0.0158
0.03	0.0631	0.0480	0.0335	0.0273
0.05	0.0814	0.0620	0.0433	0.0352
0.07	0.0963	0.0733	0.0512	0.0417
0.09	0.1092	0.0832	0.0581	0.0472
(c) 測定誤差の割合：10%				
0.01	0.0355	0.0270	0.0189	0.0154
0.03	0.0614	0.0467	0.0327	0.0266
0.05	0.0793	0.0603	0.0421	0.0343
0.07	0.0938	0.0714	0.0499	0.0405
0.09	0.1063	0.0809	0.0565	0.0460
(d) 測定誤差の割合：15%				
0.01	0.0345	0.0263	0.0183	0.0149
0.03	0.0597	0.0454	0.0317	0.0258
0.05	0.0770	0.0586	0.0410	0.0333
0.07	0.0911	0.0694	0.0485	0.0394
0.09	0.1033	0.0787	0.0549	0.0447

当であろう．また，本来的な測定誤差が含まれている場合でも，有意差が出ることに対して慎重な態度をとるとすれば，0.00 が最も危険が少ない．

以上，適合判定の変動に関して考察したが，同様に未発見の適合文書についても，4.3 節で設定した，ある特定の検索手法だけに未発見の適合文書が集中するような状況は考えにくく，特定の手法だけではなく全体的に未発見の適合文書が存在するとすれば，結局，その影響は平均の操作で相殺され，相対的な MAP の大きさにはそれほどの効果をもたらさないと考えられる．

ただし，やはり差の検定結果に対する影響については慎重に見てみる必要がある．たとえば，4.3 節の結果では，未発見の適合文書の影響により，MAP の差は約 17%，標本分散は約 11%減少し，それによって検定結果の有意差が変化した．ここでは，これらの数値を手がかりに，

(a) MAP の差が 15%，標本分散が 10%減少
(b) MAP の差が 10%，標本分散が 5%減少

という 2 つの場合を想定してみる．未発見の適合文書を考慮に加えた場合には，式 (23) は，

$$y = (1-Q)^{-1}\sqrt{s^2(1-K)(1-H)/L} \times P_t^{-1}(0.05, L-1) \quad (24)$$

となる．ここで，Q は MAP の差の減少の割合で，上の (a) の場合ならば $Q = 0.15$，H は標本分散の減少の割合で，上の (a) の場合ならば $H = 0.1$ である．

表 8 有意水準 5%を達成するための MAP の差：未発見の適合文書の影響を含む
Table 8 Differences of MAP at the significance level 0.05: including the effect of unknown relevant document.

標本分散	検索課題数 30	50	100	150
(a) MAP の差が 15%，標本分散が 10%減少				
0.01	0.0417	0.0318	0.0222	0.0181
0.03	0.0722	0.0550	0.0384	0.0312
0.05	0.0932	0.0710	0.0496	0.0403
0.07	0.1103	0.0840	0.0586	0.0477
0.09	0.1251	0.0952	0.0665	0.0541
(b) MAP の差が 10%，標本分散が 5%減少				
0.01	0.0405	0.0308	0.0215	0.0175
0.03	0.0701	0.0534	0.0373	0.0303
0.05	0.0905	0.0689	0.0481	0.0391
0.07	0.1070	0.0815	0.0569	0.0463
0.09	0.1214	0.0924	0.0645	0.0525

注：測定誤差はすべて 0.0 と仮定

未発見の適合文書の存在は通常知られることはないので，「もし存在した場合」の危険を加味して，MAP の差を広めにしておくために，式 (24) では $(1-Q)^{-1}$ を掛けている．これによって，有意水準 5%を確保するために必要な MAP の差はより大きくなる．たとえば分散が 0.03 で検索課題数が 50 の場合，未発見の適合文書の存在を仮定しない表 7 では，有意水準 5%を得るために必要な MAP の差は 4.93%であるが，上記 (a) の仮定の下に式 (24) で計算すると，5.5%に広がる．なお，式 (24) において標本分散の減少に対応する項 $1-H$ は逆に MAP の差を小さくする役割を果たしているが，結果的に $(1-Q)^{-1}$ によって MAP の差は広がることになる．

式 (24) を使った計算例を**表 8** に示す．表 8 の形式は表 7 とほとんど同じであるが，測定誤差の割合 K は 0.00 のみを使った．有意差について慎重な態度をとりたいならば，表 7 よりも表 8 を用いたほうが危険がより少ない．

6. おわりに

本稿では，最初に平均精度の数学的定義を導入し，最小値や無作為出力における期待値，変化量など，その定義式から導かれるいくつかの性質を議論した．次に，2 つの手法間の性能を比較する場合の統計的な検定方法に焦点を当て，まず，平均精度のデータを対標本ととらえた場合の差の検定を検討した．実際に，対標本を仮定しない検定に比べて，対標本の場合には両側確率がかなり小さくなることが 4 章の実験で明らか

になった．

続いて，適合判定の変動および pooling での未発見の適合文書の 2 つに着目し，それらが平均精度を用いた統計的検定の結果に与える影響を議論した．前者に関しては，ベルヌーイ試行を仮定した測定誤差モデルを導入し，実際のデータによるシミュレーションを用いて適合判定の変動の影響を調べた．その結果，適合判定の変動にともなう誤差は小さく，大局的な分析には大きな影響を与えないものの，手法間の差の検定における有意差の程度は判定の変動に比較的敏感であり，注意を要することが明らかになった．

後者の未発見の適合文書についても，シミュレーションで調べた結果，平均精度による性能比較の分析には大局的には大きな影響を与えないことが分かった．本稿の 2.3 節で議論したように，平均精度という指標は上位文書の適合/不適合の相違が相対的に大きな効果を持ち，下位文書は大きな影響力を持たない．100 位までを pooling すれば未発見の適合文書は必ず 101 位以下にしか出現しないから，平均精度という指標を採用する限り，未発見適合文書の影響は大きくならない．しかしその一方，判定変動の場合と同様に，差の検定の有意差のレベルでは，未発見の適合文書の影響を考慮する必要のあることも明らかとなった．

以上の結果をふまえて，5 章では，適合判定の変動や未発見の適合文書の程度に応じて，有意水準 5%での有意差を確保するには，どの程度の MAP の差が必要となるかについて論じ，それを計算するためのモデルを提示した．

本稿の最後として，本研究の限界と今後の研究課題をいくつかあげておく．まず，適合判定の変動の分析における，検索課題に対する個々の文書の適合確率の設定に関する問題がある．これに関しては，すでに述べたように，一定数の判定者を用意した，より精密な実験を実施することが望ましいが，本稿では既存のテストコレクションで利用可能なデータのみを用いざるをえなかった．このため，適合確率の推計値は近似的なものにとどまっている．本稿の 4.2 節や 4.3 節の結果はすべてこの近似的な適合確率に依拠している点に注意しなければならない．

未発見の適合文書の影響に関しても，本研究は，実際のテストコレクションでその探索を試みたわけではなく，あくまで仮想的な例を設定して，シミュレーションを行ったのみである．上記の適合判定の変動に関する計算例では，その適合確率の推計が，近似的ではあるものの，利用可能なデータを使って現実の状況を再現しようと努力しているのに対して，未発見の適合文書については，できる限り最悪の事例を設定して，その場合について分析したのにとどまっている．

以上のような限界のため，有意水準 5%を達成するための MAP の差を求めるモデル式 (23) あるいは式 (24) におけるパラメータ K，H，Q の設定がそれほど現実的ではない可能性がある．そこで，より正確なデータを使ってこれらのパラメータの大きさを調べてみる必要がある．これは今後の課題である．

また，t 検定には正規母集団の仮定が必要であることから，これが成立する保証はないとして，ノンパラメトリックな検定を考える研究者が多い[4),11)]．この場合の適合判定の変動の影響および未発見の適合文書の影響についても今後調べる必要がある．さらに，本稿の議論はあくまで 2 つの手法間の比較に限定されている．手法間の比較をよりマクロ的に行うための分散分析[3)]に対する探究もまた今後の課題といえる．

謝辞 本研究で使用した NTCIR-1 のデータは，国立情報学研究所情報学資源研究センター客員助教授として特別に使用を許可されたものです．関係各位に御礼申し上げます．

参 考 文 献

1) Buckley, C. and Voorhees, E.: Evaluating measure stability, *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.33–40 (2000).
2) Voorhees, E.: Variations in relevance judgements and the measurement of retrieval effectiveness, *Information Processing & Management*, Vol.36, pp.697–716 (2000).
3) Tague-Sutcliffe, J. and Blustein, J.: A statistical analysis of the TREC-3 data. *Overview of the 3rd Text REtrieval Conference* (*TREC-3*), Harman, D.K. (Ed.), National Institute of Standards and Technology, pp.385–398 (1995).
4) Robertson, S.E.: On sample sizes for non-matched-pair IR experiments, *Information Processing & Management*, Vol.26, No.6, pp.739–753 (1990).
5) Schamber, L.: Relevance and information behavior. *Annual Review of Information Science and Technology*, Vol.29, pp.3–48 (1994).
6) Cochran, W.G.: *Sampling Techniques*, 3rd ed. pp.377–399, John Wiley & Sons, New York (1977).
7) Särndal, C.-E., Swensson, B. and Wretman, J.: *Model Assisted Survey Sampling*, pp.601–636, Springer-Verlag, New York (1992).
8) Chen, A., Gey, F., Kishida, K., Jiang, H. and

Liang, Q.: Comparing multiple methods for Japanese and Japanese-Engish text retrieval, *Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, Kando, N. and Nozue, T. (Eds.), pp.49–58, National Center for Science Information Systems (1999).

9) 栗山和子，神門典子，野末俊比古，大山敬三：大規模テストコレクション構築のためのプーリングについて：NTCIR-1 の予備テストの分析，99-FI-54-4, pp.25–32 (1999).

10) Burgin, R.: Variations in relevance judgments and the evaluation of retrieval performance, *Information Processing & Management*, Vol.28, No.5, pp.619–627 (1992).

11) Wilbur, W.J.: Non-parametric significance tests of retrieval performance comparison, *Journal of Information Science*, Vol.20, No.4, pp.270–284 (1994).

付　　録

1. 無作為出力の場合の平均精度の期待値の証明

本文の式 (5) が成り立つことを以下に示す．

たとえば，$N = 4$，$R = 3$ で出力結果が「1110」の場合には，平均精度は式 (1)' より，

$$R^{-1}\left[\frac{1}{1}x_1+\frac{1}{2}(x_1x_2+x_2)+\frac{1}{3}(x_1x_3+x_2x_3+x_3)\right]$$

である．ただしこの式では，x_1x_4 などのその値が 0 となる項は省略し，1 となる項のみを表示してある．

$N = 4$，$R = 3$ の場合，そのほか「1101」「1011」「0111」の出力パターンがある．このうちの 1 つを無作為抽出する過程を無限回繰り返し，それぞれのパターンの出現回数を数えれば，結局，それらは等しくなる．したがって，この場合，無作為出力の平均精度の期待値は，これらの 4 つのパターンの平均精度を平均すれば求められることになる．

「1110」の平均精度は上に示したが，残りの 3 パターンに対して同様の式を求め，すべてを足し合わせた量を T と表記する．この T を 4 で割れば，$N = 4$ かつ $R = 3$ の場合の無作為出力の期待値が得られる．なお，パターンの総数は ${}_NC_R = {}_4C_3 = 4$ で計算できる．

x_1 や x_2 などの単項が T の算出の際に何回出てくるかを考えると，これは ${}_{N-1}C_{R-1} = {}_3C_2 = 3$ 回である（ここでの式の表現法では x_1 や x_2 が 0 になる場合には項としては出てこないことに注意．以下同様）．たとえば x_1 が出現するパターンは「1110」「1101」「1011」の 3 パターンであり，それぞれ 1 回ずつ x_1 を含んでいる．このようなパターンの総数は，先頭の「1」を固定したときの残りのパターンの組合せの数（「110」「101」「011」）として求められるから，${}_{N-1}C_{R-1}$ 回である．

次に，x_1x_2 や x_2x_3 などの 2 つの変数が掛け合わされた項がそれぞれ何回出現するかを考えてみると，これは ${}_{N-2}C_{R-2} = {}_2C_1 = 2$ 回である．たとえば x_1x_2 が出現するパターンは「1110」「1101」の 2 つであり，これは先頭から「11」を固定したときの残りのパターンの組合せの数（「10」,「01」）である．ゆえに ${}_{N-2}C_{R-2}$ 回である．

したがって，結局，一般的に，

$$T = R^{-1}\left[\sum_{i=1}^{N}\left({}_{N-1}C_{R-1}\times\frac{1}{i}x_i\right) + \sum_{i=2}^{N}\left({}_{N-2}C_{R-2}\times\frac{1}{i}\sum_{k=1}^{i-1}x_ix_k\right)\right]$$

となる．この式中に出現する x_i の値は必ず 1 であることに注意すると，上の式の [ ] の中の各項はそれぞれ次のように書ける（2 番目の項については添え字が動く範囲を広げたことに注意）．

$${}_{N-1}C_{R-1}\sum_{i=1}^{N}\frac{1}{i},\ \text{および}$$

$${}_{N-2}C_{R-2}\sum_{i=1}^{N}\frac{i-1}{i} = {}_{N-2}C_{R-2}\sum_{i=1}^{N}\left(1-\frac{1}{i}\right)$$

これらを再び足し合わせて整理すると，

$${}_{N-1}C_{R-1}\left[\sum_{i=1}^{N}\frac{1}{i}+\frac{{}_{N-2}C_{R-2}}{{}_{N-1}C_{R-1}}\sum_{i=1}^{N}\left(1-\frac{1}{i}\right)\right] = {}_{N-1}C_{R-1}\sum_{i=1}^{N}\left[\frac{1}{i}+\frac{R-1}{N-1}\left(1-\frac{1}{i}\right)\right]$$

となる．さらに，

$$\frac{1}{i}+\frac{R-1}{N-1}\left(1-\frac{1}{i}\right) = \frac{(N-1)i^{-1}+(R-1)(1-i^{-1})}{N-1} = \frac{(R-1)+i^{-1}(N-R)}{N-1}$$

であることに注意すれば，結局，

$$T = \frac{{}_{N-1}C_{R-1}}{R}\sum_{i=1}^{N}\frac{R+i^{-1}(N-R)-1}{N-1}$$

となる．この T をパターン総数 ${}_NC_R$ で割ったものがここで求める期待値であるが，

$$\frac{{}_{N-1}C_{R-1}}{{}_NC_R}\times\frac{1}{R} = \frac{R}{N}\times\frac{1}{R} = \frac{1}{N}$$

なので，最終的に

$$\frac{1}{N(N-1)}\sum_{i=1}^{N}[R+i^{-1}(N-R)-1]$$
$$=\frac{1}{N(N-1)}\left[N(R-1)+\sum_{i=1}^{N}\frac{N-R}{i}\right]$$

を得る．これは本文式 (5) に等しい．（証明終）

2. 測定誤差を含んだ場合の分散 $V_{pm}(\bar{\nu})$ の導出

ここでは，標本抽出理論の教科書[6]に従って，本文式 (17) を導出する過程を示す．

まず，統計学でよく知られた結果により，

$$V_{pm}(\bar{\nu})=V_p[V_m(\bar{\nu}|S)]$$
$$=E_p[V_m(\bar{\nu}|S)]+V_p[E_m(\bar{\nu}|S)]$$

のように分解できる（証明は Cochran[6]の p.275 を参照）．

本文式 (14) から $E_m(\bar{\nu}|S)$ は統計量 μ に対しての標本平均である．したがって，上の式の最右辺の第 2 項はいわば標本平均の分散であるから，統計学の初等的な結果より，μ の母分散 σ_μ^2 を標本の大きさで割ったものになる．すなわち，

$$V_p[E_m(\bar{\nu}|S)]=\sigma_\mu^2/L$$

である．

残りの最右辺の第 1 項については以下のようになる．まず，その [ ] の中身は，

$$V_m(\bar{\nu}|S)=V_m\left[L^{-1}\sum_{h=1}^{L}\nu_h\middle|S\right]=L^{-2}\sum_{h=1}^{L}V_m(\nu_h|S)$$

となるが，本文中で仮定したように検索課題間の適合判定は独立なので，$V_m(\nu_h|S)=V_m(\nu_h|h)$ となるから，上の式の最右辺は，$V_m(\nu_h|h)=\sigma_h^2$ を使って，

$$L^{-2}\sum_{h=1}^{L}V_m(\nu_h|h)=L^{-2}\sum_{h=1}^{L}\sigma_h^2$$

である．これを代入すれば，

$$E_p\left[L^{-2}\sum_{h=1}^{L}\sigma_h^2\right]=L^{-1}E_p\left[L^{-1}\sum_{h=1}^{L}\sigma_h^2\right]=L^{-1}\sigma_d^2$$

となる．最右辺の導出は本文式 (18) の定義式から明らかである．

以上の結果を合わせれば，本文式 (17) が得られる．

(平成 13 年 9 月 25 日受付)
(平成 13 年 11 月 6 日採録)

（担当編集委員　大山 敬三）

岸田　和明（正会員）

1991 年慶應義塾大学大学院文学研究科図書館・情報学専攻博士課程中退．同年図書館情報大学助手．1994 年駿河台大学文化情報学部助教授，現在に至る．2000 年より国立情報学研究所情報学資源研究センター客員助教授．ACM，日本統計学会，日本図書館情報学会等会員．