

検索実験における評価指標としての Mean Average Precision の性質

岸田和明

駿河台大学文化情報学部, 国立情報学研究所情報学資源研究センター
*〒357-8555 埼玉県飯能市阿須 698
kishida@surugadai.ac.jp

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、検索実験において利用される検索質問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し、評価指標としての平均精度の特徴を明らかにすることである。第一に、適合判定の変動が平均精度を用いた平均値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おこなう。第二に、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。これらの分析のための具体例として、NTCIR-1の結果の一部を利用する。

Property of Mean Average Precision as Performance Measure in Retrieval Experiment

Kazuaki KISHIDA
Faculty of Cultural Information Resources, Surugadai University
698 Azu, Hanno, Saitama 357-8555
kishida@surugadai.ac.jp

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 to discuss some issues on statistical test for determining a difference of retrieval performance between two systems by using mean average precision as a evaluation measure. 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 example 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その中でも特に検索性能 (retrieval performance) に焦点が当てられ、検索質問 (トピック) に対する正解文書 (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 ratio) と精度 (precision ratio) 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」として定義される。しかし、これらの指標は本来的にはブール代数に基づく伝統的な検索方法における評価のために開発されたものである。そのため、最近主流となりつつある、文書を適合度の順序で出力する手法 (またはシステム) を評価する場合には、これらの指標を直接的に適用することはできず、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。これは簡単に言えば「各適合文書が検索された時点での精度の平均」[1]を意味する (より詳しい定義は後述)。そして、平均精度自体は検索質問ごとに計算されるが、検索実験では複数の検索質問が用意されるため、検索質問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 MAP が中心的な評価指標として扱われる傾向にある。

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、平均精度の歴史は比較的浅く、評価指標としてのそ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もよく、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性についての研究が急務となっている。本稿の目的は、検索性能の比較を目的とした平均精度による統計的検定の問題を議論することによって、評価指標としての MAP の性質に対する理解を深めることにある。以下、第2節では平均精度の基本的性質を整理し、第3節では平均精度による統計的検定について、(1)対標本の場合の検定方法、(2)適合判定が確率的に変動する場合の影響、(3)pooling の方法論的問題の影響の3点を議論する。第4節では、NTCIR-1 の具体的なデータを用いて、第3節の議論に関する数値計算例を示す。

2 平均精度の基本的な性質

2.1 平均精度と MAP の数学的定義

平均精度の数学的定義は次のように与えられる[3]。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 ($i = 1, \dots, N$)。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。ある 1 つの検索実行における平均精度 v は

$$v = \frac{1}{\sum_{i=1}^N x_i} \sum_{i=1}^N \left[\frac{x_i}{i} \left(1 + \sum_{k=1}^{i-1} x_k \right) \right] = \frac{1}{\sum_{i=1}^N x_i} \sum_{i=1}^N \frac{1}{i} (x_1 x_i + x_2 x_i + \dots + x_{i-1} x_i + x_i) \quad (2.1)$$

で定義される。なお、(2.1)式中の $\sum_{i=1}^N x_i$ は N 件に含まれる適合文書数であり、本稿ではこれを R と表記することがある。なお平均精度の最大値は 1.0 であるが、(2.1)式から容易にわかるように、最小値は 0.0 ではなく N と R に依存し、 $R^{-1} \sum_{k=1}^R [k/(N-R+k)]$ である[3]。また N 件の文書を無作為に出力した場合

の平均精度の期待値は $E(v|N, R) = \frac{1}{N(N-1)} \sum_{i=1}^N \left(R + \frac{N-R}{i} - 1 \right)$ となる[3]。

さらに平均精度 v に添え字を付けて、 v_h を第 h 番目の検索質問に対する平均精度とする ($h = 1, \dots, L$)。

MAP は L 個の検索質問に対する v_h の平均であり $\bar{v} = L^{-1} \sum_{h=1}^L v_h$ で定義される。

2.2 平均精度の変化量

適合文書を 1、不適合文書を 0 で表わし、第 1 位から順に並べれば、ある検索実行の出力結果を「101110...」のようなビット列で表現できる。ここでは、適合文書の順位が移動したとき (例えば「11000...」から「10100...」へ変化) の平均精度の変化量について議論する。

順位 j の適合文書と順位 l の不適合文書が入れ替わったときの平均精度の変化量を考える。この場合には、適合文書数 $R = \sum_{i=1}^N x_i$ 自体は変化しないから、この項を除いた(2.1)式を x_j で微分すると、

$$\frac{d}{dx_j} \left\{ \sum_{i=1}^N \left[i^{-1} x_i \left(1 + \sum_{k=1}^{i-1} x_k \right) \right] \right\} = j^{-1} \left(1 + \sum_{k=1}^{j-1} x_k \right) + \sum_{i=j+1}^N i^{-1} x_i \quad (2.2)$$

を得る[3]。(2.2)式右辺第2項から明らかなように、平均精度の場合、第 j 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 $j+1$ 位から第 N 位までの間に存在する文書からも影響を受ける。このこと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位までですべて見出せる場合(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(2.2)式に示されている平均精度の特徴をよく表わしている。

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

検索実験では通常いくつかの検索質問(トピック)が用意されるが、当然これらは想定されるすべての検索質問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やシステム間の性能比較をおこなう場合にはこれらの検索質問の集合を母集団から抽出された標本として見なす必要がある。実際に Tague-Sutchiff and Blustein[4]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。

検索質問の集合を単純無作為抽出による標本として考えれば、 $\text{MAP}(\bar{v} = L^{-1} \sum_{h=1}^L v_h)$ は平均精度 v_h の母平均の推定量を意味することになる。そして、ある2つの手法(あるいはシステム)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推定量 $\bar{v}$ によって検定することにほかならない。

このための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 α と β 、それぞれの平均精度の平均を $\bar{v}_\alpha$ と $\bar{v}_\beta$ 、平均精度の標本分散を s_α^2 と s_β^2 とかく。両者の標本の大きさが等しければ、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

$$t_1 = (\bar{v}_\alpha - \bar{v}_\beta) / \sqrt{s_\alpha^2/L + s_\beta^2/L} \quad (3.1)$$

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(正規母集団の仮定が必要)。あるいは s_α^2 と s_β^2 がそれぞれの母分散に等しいと仮定できれば統計量 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同一の検索質問に対する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)と捉える場合がある[5]。この場合には、検索質問ごとの手法間の差 $v_{ha} - v_{h\beta}$ 自体を標本と考える(v_h に添え字 α と β をつけた)。この差の平均は

$L^{-1} \sum_{h=1}^L (v_{ha} - v_{h\beta}) = \bar{v}_\alpha - \bar{v}_\beta$ であるから、標本分散は簡単な計算から $s_\alpha^2 + s_\beta^2 - 2\text{Cov}_{\alpha\beta}$ となる。ここで $\text{Cov}_{\alpha\beta} = (L-1)^{-1} \sum_{h=1}^L (v_{ha} - \bar{v}_\alpha)(v_{h\beta} - \bar{v}_\beta)$ である(v_{ha} と $v_{h\beta}$ 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、

$$t_2 = (\bar{v}_\alpha - \bar{v}_\beta) / \sqrt{s_\alpha^2/L + s_\beta^2/L - 2\text{Cov}_{\alpha\beta}/L} \quad (3.2)$$

が自由度 $L-1$ の t 分布に従うことになる。 $\text{Cov}_{\alpha\beta} > 0$ 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(3.2)式を使ったほうが帰無仮説はより棄却されやすくなる(有意差が出やすい)。

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[6]。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種類の誤差要因が含まれる可能性がある。

前者の適合判定の変動を x_i の測定の際の誤差と捉えれば、統計学分野で開発された測定誤差の数学的モデル[7][8]を適用することが可能になる。ある1つの検索手法(またはシステム)を固定して、上記①の変動に対する期待値を $E_m(\cdot)$ 、分散を $V_m(\cdot)$ 、上記②の変動に対する期待値を $E_p(\cdot)$ 、分散を $V_p(\cdot)$ とし、 $E_p[E_m(\cdot)] = E_{pm}(\cdot)$ 、 $V_p[V_m(\cdot)] = V_{pm}(\cdot)$ と表記する。標本調査の教科書[7][8]に従って計算すれば、結果的に $E_{pm}(\bar{v}) = \mu$ 、 $V_{pm}(\bar{v}) = L^{-1}(\sigma_d^2 + \sigma_\mu^2)$ を得る。ここで μ は h 番目の検索質問に対する μ_h の母平均であり、 $\mu_h = E_m(v_h|h)$ である(μ_h は、 h を固定して適合判定を無限回繰り返し、その結果得られる複数個の v_h を平均することにより変動要因を除去したものに相当する)。また σ_d^2 は $\sigma_h^2 = V_m(v_h|h)$ の母集団での平均値である。最後に σ_μ^2 は μ_h の母分散である。

適合判定に測定誤差が含まれている場合には、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でなく測定誤差 σ_d^2 による散らばりも含まれることになる。すなわち σ_μ^2 に対する「純粋な標本抽出誤差」よりも $L^{-1}s^2$ が過大に評価されていることになり、この結果、(3.1)式や(3.2)式の分母が大きくなって、帰無仮説が棄却されにくくなっている(手法間の差が有意になりにくい)可能性がある。

3.3 poolingの方法論的課題に起因する平均精度の変動

大規模なデータベースの場合には各検索質問に対するすべての適合文書を発見するのは難しく、そのために多くの場合poolingの方法が採用されている。これは、数多くの検索手法(あるいはシステム)による同一検索質問に対するそれぞれの検索結果の上位 x 件を合併し、それらに対してのみ適合判定をおこなう方法である。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に過ぎず、すべての適合文書が発見されていない可能性がある。

例えばある検索実験プロジェクトにおいて、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いければ、この適合文書は未発見のまま埋もれることになる(不適合文書と見なされる)。もし仮にこの文書 d が適合していることが後から発見されたとすれば、その平均精度の変化量 Δ は、

$$\Delta = (R+1)^{-1} \left[j^{-1} \left(1 + \sum_{k=1}^{j-1} x_k \right) + \sum_{t=j+1}^N t^{-1} x_t \right] - (R+1)^{-1} \bar{v} \quad (3.3)$$

で与えられる。ここで R と $\bar{v}$ はともに文書 d が発見される以前の適合文書数と平均精度である。 j は適合文書が発見された順位であり、ここでの例では $j=101$ となる。(3.3)式は次のように求められる。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、(2.2)式を直接応用している。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 に関連した部分以外で平均精度が減少する分に相当し、 $[R^{-1} - (R+1)^{-1}]R\bar{v} = (R^2 + R)^{-1}R\bar{v} = (R+1)^{-1}\bar{v}$ として導かれる。

もし第 $(j+1)$ 位(=102位)以下に適合文書がなければ、 $1 + \sum_{k=1}^{j-1} x_k = 1 + R$ 、 $\sum_{t=j+1}^N t^{-1} x_t = 0$ であるから、(3.3)式は $\Delta = j^{-1} - (R+1)^{-1}\bar{v}$ となる(ここで $j=101$)。この式を使って、第101位の文献が実は適合文書であるということが発見されたときの平均精度の変化を計算すると表3.1のようになる(R と $\bar{v}$ は適当に

表3.1 101位に適合文書が発見された影響
(表例が適合文書数、表頭が平均精度)

	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

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 の発見によって平均精度が下がることもある。そして、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と考えられる。ただし、比較対象の他の手法の平均精度もまた(3.3)式にしたがって変化する点には注意しなければならない(なお上記のpoolingのメカニズムにより文書 d が100位以内に出現することはない)。

4 実際のデータを用いた分析

4.1 使用データ

この節では、前節で議論したMAPによる統計的検定の問題を実際のデータを使って議論する。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BKJJD CFUとを選び、それらの比較評価を試みる。前者は索引作成方法としてbigramを応用した「短い検索質問」に対する実行結果であり、後者は辞書との最長一致法を用いた「長い検索質問」に対する実行結果である[9]。ともにロジスティック回帰型検索モデルを使っている。なお以下の計算に使う検索質問(トピック)はNo.31~No.83までの53件である(すなわち標本の大きさ $L = 53$)。

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

3.1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、検索質問を固定して、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 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り返しによって得られることになる。この繰り返しは相互に独立であるならば、これをベルヌーイ試行として捉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($i = 1, \dots, N$)。なお、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、(2.1)式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 $\mu_h = E(v_h|h)$ は一種の比推定のかたちになってしまう。このためこの値を解析的に正確に計算することは難しい。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 件の検索質問ごとにそれぞれ全文書に対して何らかの方法で p_i ($0.0 \sim 1.0$)を設定する($i = 1, \dots, N$)。
- ② ある1つの方法(例えばBKJJBIDS)を選び、その L 件の検索質問(トピック)に対する実行結果(文書のランキング)を用意する。
- ③ 1つの検索質問(トピック)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おこなう:(1-1)文書ごとに $0.0 \sim 0.1$ の一樣乱数を発生させ、(1-2)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つの検索質問に対して 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(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 件の検索質問に対しておこなう。
- ⑥ 最終的に、 μ_h ($h = 1, \dots, L$)から σ_μ^2 を推計し(通常の標本分散を計算する方法を使う)、 σ_h^2

($h = 1, \dots, L$) を平均して σ_d^2 の推計値とする。

4.2.2 実際の計算例

ここでは NTCIR-1 のデータを用いた計算の一例を示す。日本語のテストコレクションである 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[10]。したがって、2 人の判定者の判定結果のパターンは組み合わせで 6 通り存在するが、ここでは仮に表 4.1 のようにパラメータ p_i を設定してみる。

表 4.1 各文書のパラメータ p の設定

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

4.2.1 で説明した手順に従って計算した結果を表 4.2 に示す。なお文書数については $N = 1,000$ ，計算の反復回数については

$M = 100,000$ とした。また表 4.2 には 2 人の判定者それぞれについて、①「正解」と「部分的正解」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平均精度の平均 (MAP) とその分散を示してある。

表 4.2 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

手法	統計量	判定者 1		判定者 2		測定誤差モデルでのシミュレーション
		正解＋部分的正解	正解のみ	正解＋部分的正解	正解のみ	
BKJJBIDS	標本平均 (MAP)	.29042	.27685	.28930	.27984	.27973
	標本分散	.05593	.05231	.05288	.06297	.05171($\hat{\sigma}_\mu^2$)
	測定誤差 $\hat{\sigma}_d^2$	—	—	—	—	.00188
BKJJDCFU	標本平均 (MAP)	.34901	.33558	.33726	.31579	.32588
	標本分散	.05401	.05545	.04893	.05346	.04558($\hat{\sigma}_\mu^2$)
	測定誤差 $\hat{\sigma}_d^2$	—	—	—	—	.00299

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。一方、 μ_h の標本分散 $\hat{\sigma}_\mu^2$ 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。そして測定誤差 σ_h^2 の標本平均 $\hat{\sigma}_d^2$ はそれぞれ 0.00188 と 0.00299 であった。この結果は、モデルから予想されたように、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。しかし一方、表 4.1 のパラメータ設定の下では σ_d^2 の影響はそれほど大きくはない。すなわち μ_h の標本分散と σ_h^2 の標本平均とを足し合わせると（すなわち $\hat{\sigma}_d^2 + \hat{\sigma}_\mu^2$ ），それぞれ 0.05359 と 0.04857 であり、 $\hat{\sigma}_d^2 + \hat{\sigma}_\mu^2$ に対する $\hat{\sigma}_d^2$ の構成比はそれぞれ 3.5% と 6.2% にすぎない。

次に、BKJJBIDS と BKJJDCFU との間に検索性能の有意差があるかどうかを 3.1 での議論に基づいて検定してみる。結果を表 4.3 に示す。(3.1)式による統計量 t_1 と(3.2)式による統計量 t_2 とを比較すると、対標本として捉えた場合の t_2 のほうが大きく、上側確率が小さい（すなわち棄却されやすい）。測定誤差モデルによって判定変動を除去した場合には、これらの値は、表 4.2 と同様に、確率設定の基となった 4 つの判定結果による検定結果をちょうど平均したような中間的な値になっている。上で述べたように、表 4.1 の確率設定の下では判定変動は MAP の計算にそれほど大きな影響を与えないので、このような結果になったと考えられる。

表 4.3 手法間の性能比較のための平均値の差の検定結果 (53 件の検索質問による結果)

手法 α : BKJJD CFU 手法 β : BKJJBIDS	判定者1		判定者2		測定誤差 モデル
	正解・部分	正解	正解・部分	正解	
$\bar{v}_\alpha - \bar{v}_\beta$ (MAPの差)	0.0587	0.0586	0.0360	0.0480	0.0462
統計量 t_1 : (3.1) 式	1.3025	1.2866	0.7671	1.0583	1.0773
上側確率 $P(t_1 < x)$: 標準正規分布	0.0964	0.0991	0.2215	0.1450	0.1407
上側確率 $P(t_1 < x)$: t分布 ¹⁾	0.1956	0.2011	0.4448	0.2924	0.2838
$v_{ha} - v_{hb}$ の標本分散	0.0337	0.0422	0.0349	0.0446	0.0330 ³⁾
統計量 t_2 : (3.2) 式	2.3304	2.0763	1.4006	1.6531	1.8503
上側確率 $P(t_2 < x)$: 標準正規分布	0.0099	0.0189	0.0807	0.0492	0.0321
上側確率 $P(t_2 < x)$: t分布 ²⁾	0.0237	0.0428	0.1673	0.1043	0.0700

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ものである。なおこの場合には $\hat{\sigma}_d^2 = 0.00487$ であった。

ところで、上で述べたように、 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 4.1)。この図は 53 件の検索質問ごとに得られた μ_h の値 (すなわち $\mu_h = E_m(v_{ha} - v_{hb} | h)$) を昇順に並べて番号 $n = 1, \dots, 53$ を付与し、第 n 位の μ_h の場合には、 x 座標を μ_h 、 y 座標を $n/53$ としてプロットした図である。図中の正規分布の曲線は、表 4.3 に示されている $\bar{v}_\alpha - \bar{v}_\beta$ と $v_{ha} - v_{hb}$

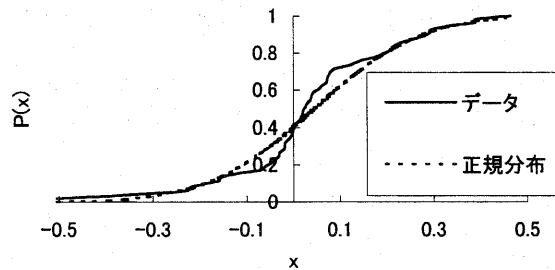


図 4.1 標本データの経験分布: 手法間の差

の標本分散を用いて、 μ_h を変数とする累積正規分布を計算して描いた。図からは標本中の μ_h の分布が正規分布に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

最後に、3.3 で議論した pooling による適合判定を採用したために「埋もれてしまった」適合文書の影響について、実際のデータによる例を示す。ここでは非常に仮想的な計算例を示す。すなわち、BKJJBIDS による第 101 位の文書の適合確率を 1.0 に強制的に変えた時の平均精度を、上記の測定誤差モデルに基づいて計算してみる。53 検索質問に対して BKJJBIDS による第 101 位の文献を調べてみたところ、51 件の検索質問に対して適合確率が 0.0 であった。これらの適合確率を 1.0 に置き換えて μ_h の平均 (MAP) を計算すると (反復回数は上と同様に 10 万回)、0.26803 となり、表 4.2 の結果と比較して 0.0117 の減少 (約 4%) となった。比較対象となる BKJJD CFU の場合には R を $R+1$ に強制的に置き換えて計算を試みたところ MAP は 0.30619 であった (差は 0.0197, 約 6% の減少)。結果として両方の MAP の差は 0.03816 であり、また $v_{ha} - v_{hb}$ に対する $\hat{\sigma}_\mu^2$ はシミュレーションによれば 0.002937 と推計された。したがって統計量 t_2 は 1.6209 となる。これを表 4.3 の結果 1.8503 と比べれば、約 12% の変化である。この変化量は、51 件の検索質問で未発見の適合文書が存在するという極端な設定のわりには、比較的小さいように思われる。以上はほんの一

例に過ぎないが、この例から、一般的な状況においても pooling における未発見の適合文書の影響はそれほど大きくないことが予想される。

5 まとめ

本稿では、適合判定の変動および pooling での未発見の適合文書の 2 つに焦点を当て、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調べた。前者に関しては、本研究で導入したベルヌーイ試行を仮定する測定誤差モデルを利用した範囲内では、適合判定の変動による測定誤差は小さく、検定結果に大きな影響を及ぼさないことがわかった。この結果は、栗山ら[10]の実証分析と矛盾しない。

後者の未発見の適合文書についても、平均精度による性能比較の分析には大きな影響を与えないことが明らかとなった。本稿の第 2 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。

本稿では、また、2 つの手法間の性能を比較する場合に、平均精度を対標本と見なす方法の検討もおこなった。これは古典的問題であり、理論的には特に新しい知見はないが、今回のデータでは、対標本を仮定しない場合の平均値の差の検定に比べて、対標本の場合には上側確率が約半分になることがわかった(表 4.3 参照)。この上側確率はもちろん母分散の大きさに影響を受けるが、今回のデータでは、MAP でおおよそ 0.05~0.06 くらいの差があれば、有意水準 5% 程度の t 検定で帰無仮説は棄却されるようである(表 4.3 参照)。この点に関しては、もちろん、他のデータを使った確認が今後必要である。

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. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p.33-40 (2000)
- [2] Voorhees, E.: Variations in relevance judgements and the measurement of retrieval effectiveness. *Information Processing & Management*, Vol.36, p.697-716 (2000)
- [3] 岸田和明: Mean Average Precision の再考: ランキング出力の評価尺度についての考察 INFOSTA シンポジウム 2001 予稿集, p.1-6 (2001)
- [4] Tague-Sutcliffe and Blustein, James: A statistical analysis of the TREC-3 data. *Overview of the Third Text REtrieval Conference (TREC-3)*. D.K.Harman ed. National Institute of Standards and Technology, 1995, p.385-398. <http://trec.nist.gov/>
- [5] Robertson, S.E.: On sample sizes for non-matched-pair IR experiments. *Information Processing & Management*, Vol.26, No.6, p.739-753 (1990)
- [6] Schamer, L.: Relevance and information behavior. *Annual Review of Information Science and Technology*, Vol.29, p.3-48 (1994)
- [7] Cochran, William G.: *Sampling Techniques* Third ed. New York, John Wiley & Sons, 1977, p.377-399.
- [8] Samdal, Carl-Erik, Swensson, Bengt and Wretman, J.: *Model Assisted Survey Sampling*. New York, Springer-Verlag, 1992. p.601-636.
- [9] Chen, A., Gey, F., Kishida, K., Jiang, H. and Liang, Q.: Comparing multiple methods for Japanese and Japanese-English text retrieval, *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*. N. Kando and T. Nozue eds, National Center for Science Information Systems, 1999, p.49-58.
- [10] 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)