—解 説—

よりよい検索システム実現のために

正解の良し悪しを考慮した情報検索評価の動向

酒井 哲也
（株）東芝 研究開発センター
tetsuya.sakai@toshiba.co.jp

本稿のねらい

近年のWeb検索エンジンの普及により，「検索」という言葉はすっかり市民権を得たようである．しかし，検索機能を持つソフトウェアや検索エンジンを普段使っていて，一度も不満を感じたことがないという方は果たしていらっしゃるだろうか．

情報検索の研究者は日夜，**検索有効性**（retrieval effectiveness）の向上に励んでいる．すなわち，どんな入力に対してもユーザを満足させる検索結果を提示することを最終ゴールとし，このために検索結果の質をなんらかの指標により定量化してシステムを最適化しようとする．検索有効性を議論するための基本概念として，情報検索の分野では**適合性**（relevance）という言葉がよく出てくるが，本稿では検索された情報（たとえば文書，Webページ，画像）が「正解」か「不正解」かといったいい加減な表現を用いることにする．古典的な**文書検索**（document retrieval）以外のタスクについても触れるつもりなので，このほうが分かりやすいだろう．

一口に「正解」といってもいろいろある．たとえば文書検索の場合，入力された検索要求にぴったりマッチしユーザを大満足させるような「大正解」もあれば，検索要求の内容と多少ずれはあるがある程度役に立つ文書や，文書のごく一部が検索要求の内容とマッチするものもあるだろう．またWeb検索において，Webページの内容のみにとどまらずメタ情報や信頼性も考慮し，たとえば公式サイトは「大正解」とし，個人のホームページやブログは「ふつうの正解」として扱いたい場合もあるかもしれない．いずれにしても，実用的な検索システム構築のために「正解レベル」を考慮した評価を行うことは自然だろう．

ところが驚いたことに，従来の情報検索研究では正解レベルを扱わず「正解か否か」という値の判断基準に基づいてシステムの評価を行うものが主流であった．このような評価の枠組みでは当然，「大正解」を検索できるシステムと「おまけの正解」しか検索できないシステムの区別ができない．また，古典的な情報検索研究では，**再現率**（recall）すなわち「全正解のうちどれだけをシステムがちゃんと見つけたか」が検索有効性を表す重要な指標として扱われてきたが，最近ではWebのように検索対象の大規模化が進み「全正解」の定義があまり意味をなさない検索シーンも見られるようになってきた．また，文書のリストを提示する文書検索とは異なり，ユーザの入力した質問に対して回答文字列をずばり出力する**質問応答**（question answering）や，そもそも文書という単位が不明確な**XML検索**（XML retrieval）といった新しい検索タスクも注目を集めており，従来の検索システム評価方法では手に負えないものも出てきた．実際の利用シーンに即した適切な評価指標のもとで改良が行われないかぎり，システムはいつまでたってもユーザを満足させることができないだろう．

以上のような背景から，本稿では古典的な文書検索タスクにとどまらず，Web検索，質問応答，特許検索，XML検索といった「新しい」情報検索タスクを視野に入れ，正解レベルを考慮した情報検索指標を中心に紹介する．より具体的には，まず古典的な情報検索評価手法からはじめ，「正解か否か」に基づいた検索評価指標について説明する．次に，比較的最近提案されたものを中心に，正解レベルに基づく検索評価指標について説明する．次に，質問応答，特許検索，XML検索といった従来の文書検索の枠組みでは扱いにくい情報検索タスクにおける評価の難しさについて説明し，今回紹介する検索評価指標のこれらのタスクへの適用可能性について述べる．最後に本稿のまとめを述べる．

「正解か否か」に基づく検索評価指標

本章では，「正解か否か」に基づく検索評価指標をいくつか紹介する．なお「正解か否か」は情報検索の用語では**二値適合性**（binary relevance）という．

まず，「ユーザはなるべくたくさんの正解を検索して欲しい」という前提に基づく古典的な情報検索タスク向けの評価指標を紹介する．次に，「ユーザは正解が1件見つかれば満足する」という前提に基づく評価指標を紹介する．後者は，近年のWebに代表される大規模データに対する検索において，あらゆる正解を検索してユーザに提示することや，そもそも「あらゆる正解」を定義すること自体が現実的でない場合に有用だろう．また，対象が大規模であるか否かにかかわらず，検索したい内容によって正解が1件見つかれば十分という場合もあるだろう．

正解をたくさん検索するタスクのための指標

大昔の情報検索はそもそも検索対象が少なく，たとえば数百件の文書の中から数十件の文書を取り出し順位もつけずに提示すれば事足りたのかもしれない．このような場合，検索有効性は検索もれの少なさを表す**再現率**（recall）と，ごみの少なさを表す**精度**（precision）により評価できる．以後，検索される単位を便宜上「文書」と呼ぶことにすると

$$\text{再現率} = \frac{\text{検索された正解文書数}}{\text{全正解文書数}}. \quad (1)$$

$$\text{精　度} = \frac{\text{検索された正解文書数}}{\text{検索された文書数}}. \quad (2)$$

ところが，情報洪水時代になって検索対象は膨大になり，たとえば数百万件の文書から数万件の文書を取り出してユーザに丸投げするのは許されない状況になってきた．そういうわけで，今日では情報検索システムといえば検索された文書をランキングして提示するのが常識である．この場合，検索結果の上位のほうだけを切り取って評価を行えば比較的低い再現率と高い精度が得られる．一方，検索結果の下位のほうまで見渡して評価を行えば比較的高い再現率と低い精度が得られる．そこでランキング検索の評価においては，このような再現率と精度のトレードオフ関係をグラフ化してシステムの優劣を議論することが多い（実際には，再現率が0, 0.1, …, 1.0の11点における精度のきれいな単調減少曲線を得るために**補間**（interpolation）を行う）．

さて，システムの優劣を議論するには，グラフの形や上下を比較するよりも，なんらかの単一の数値で示される評価指標を比較するほうが分かりやすいし統計処理もやりやすい．**再現率–精度曲線**（recall-precision curves）から単一の数値を求める方法としては，たとえば**11点平均精度**（eleven-point average precision）が知られている．これは前述の11点の再現率における精度を平均したもので，（補間された）再現率–精度曲線を無理矢理水平にしたときの高さを意味する．しかし，現在最も広く用いられている検索評価指標は米国の評価型ワークショップTREC（Text REtrieval Conference）[☆1]で用いられてきた**（補間なし）平均精度**（（non-interpolated）average precision）である．これは（補間なし）再現率-精度曲線の下側の面積のようなもので，11点平均精度よりも検索結果の上位の変動に敏感な指標である．最も一般的な検索評価指標であるので，以下にきちんと定義しておこう．

情報検索の評価では通常，文書集合と，検索課題の集合と，各検索課題について上記文書集合からあらかじめ探し出した正解集合からなる**テストコレクション**（test collection）というデータを利用する[12]（ただし，テストコレクションによる評価は検索システムの実用性を示すための必要条件ではあるが，十分条件ではない．すなわち，いくら「実験室」でよい評価結果が出たからといって，それで実際にシステムを使うユーザが満足するとは限らない）．テストコレクション中のある特定の検索課題に対する全正解の数，すなわち再現率の分母をRで表すことにする．一方，システムが出力したランクつき検索結果のサイズをLで表すことにする．そして，検索結果の第r位における文書が正解であるとき1，正解でないとき0となるフラグを$I(r)$で表し，また第r位までに含まれる正解の個数を$count(r)$ $(\leq r)$で表すことにする．第r位における精度が

☆1 http://trec.nist.gov/

$$P(r) = \frac{count(r)}{r} \quad (3)$$

と書けることがお分かりだろう．このとき，平均精度は以下のように定義される．

$$\text{平均精度} = \frac{\sum_{1 \le r \le L} I(r)P(r)}{R}. \quad (4)$$

これを言葉で表現すると，「各正解が検索された時点での精度を，全正解について平均したもの」となる．分子はシステムが検索できた正解に関する精度の和だが，分母は全正解数Rであることに注意して欲しい．これは，システムが検索できなかった正解の精度を0と見なして平均をとっていることを意味する．

TRECをはじめとする評価型ワークショップで平均精度と共に用いられている評価指標としては，**R-精度**（R-precision）がある．これは，$P(R) = count(R)/R$，すなわち第R位における精度である．テストコレクションの検索課題セットは一般に正解数Rのばらつきが大きいことが知られているが，平均精度もR-精度も共に分母をRとすることにより，再現率を基準にした検索課題間の比較を可能にしている．

このほかによく用いられる評価指標に，**第l位における精度**（precision at document cut-off l）がある．これは$P(l) = count(l)/l$で定義され，lとしては検索課題によらず10, 50, 100などの固定値が用いられる．平均精度やR-精度が再現率を基準にしているのに対し，第l位における精度はユーザが何件文書を調べたか，すなわちユーザの労力を基準にしている．このため，平均精度やR-精度を**システム指向**（system-oriented），第l位における精度を**ユーザ指向**（user-oriented）の指標と呼ぶ人もいる．しかし，平均精度やR-精度と異なり，第l位における精度は検索課題ごとに上限値が異なってしまうため，この指標をテストコレクションの検索課題セットについて平均するのは好ましくない．たとえば正解数$R = 100$である検索課題Aと$R = 10$である検索課題Bがある場合に，50位における精度で評価をする場合を考える．このとき，検索課題Aの場合の上限値は1であるが，検索課題Bの場合の上限値は10/50 = 0.2である．システムがどんなに頑張っても正解は10件しかないからである．さらに，50位における精度によれば，検索結果の1～10位が正解であるシステムも41～50位が正解であるシステムも同等と見なされてしまう．実際，第l位における精度は平均精度よりも著しく**安定性**（stability）が低いことが実証されている（ここで，安定性とは，たとえば別の検索課題セットを用いた場合でも同じような実験結果が得られるかどうかを示す指標である[7]）．

このほかにも，たとえば**平均探索長**（expected search length）や**正規化再現率**（normalized recall）などの評価指標が知られているが，今日まであまり用いられていない．

正解を1つだけ検索するタスクのための指標

前述のように，実際の検索システムの利用シーンでは，「正解が1件見つかればよい」という場合も多いだろう．このような前提に基づく評価指標としては，たとえばTRECのWebタスクや質問応答タスクなどにおいて利用されてきた**逆数順位**（reciprocal rank）がある．逆数順位は，検索結果が1つも正解を含まない場合0と定義される．そうでない場合，検索結果中の最も上位にある正解の順位をr'としたとき，

$$\text{逆数順位} = \frac{1}{r'} \quad (5)$$

と定義される．

逆数順位の特徴は，検索結果中にいくつ正解が含まれていてもとにかく最上位の正解の順位r'しか考慮せず，システムの有用性はr'に反比例すると仮定している点である．たとえば最初の正解が1位にあれば逆数順位は1すなわち満点だが，最初の正解が2位にあると逆数順位は一気に0.5になる．

r'は最初に見つかった正解の順位であるから，$count(r') = 1$である．したがって，実は逆数順位は$P(r') = count(r')/r'$と一致する．言葉でいうと，逆数順位とは検索結果中で最初の正解が見つかった時点での精度にほかならない．すなわち，平均精度が全正解についての精度を見渡しているのに対し，逆数順位は最初に検索された正解についての精度のみを考慮している．このため，逆数順位による評価結果は平均精度による評価結果に比べるとかなり不安定である[9]．これは，逆数順位のほうが「たまたま上位に検索できた正解」や「たまたま上位に検索できなかった正解」に左右されやすいためである．したがって，逆数順位によりある程度信頼性の高い評価を行うには，できるだけ多くの検索課題を用いて評価を行う必要がある．

「正解レベル」に基づく検索評価指標

さて，ここからが本題である．本章では，正解レベルに基づく検索評価指標をいくつか紹介する．「正解レベル」を情報検索の用語でいうと**多値適合性**（graded relevance）となる．

情報検索評価の研究というと，古くは英米，90年代以降は米国のTREC主導という感が否めないが，正解

レベルに基づく検索評価指標の研究に関してはフィンランド人や日本人が頑張っている．たとえばフィンランド・タンペレ大のSormunenはACM SIGIR☆2 2002において，もともと正解レベルの情報を持っていないTRECのテストコレクションを手作業で再検査して新たに正解レベルを付与し，この結果，TRECの正解データの半分程度は実は「ぎりぎり正解」（marginally relevant documents）でしかないと報告している．同様に，SakaiとSparck JonesもACM SIGIR 2001において，TRECの一部の文書セットに対する「大正解」（highly relevant documents）は正解データの半分程度であったと報告している．繰り返しになるが，「ぎりぎり正解」と「大正解」を同一視したデータおよび評価指標により評価を行っている限り，「大正解」を最優先して検索してくれるシステムの実現はいつまでたってもできないだろう．

上記の観点からは，日本をはじめ東アジアの情報検索研究者は恵まれた環境にいる．なぜなら，TRECのアジア版と呼ばれる国立情報学研究所主催の国際ワークショップNTCIR（NII-NACSIS Test Collections for Information Retrieval systems）☆3では，文書検索の正解データにはじめから正解レベルの情報がつけられているためである（大量の文書データから効率的に正解を選出する方法については，本誌Vol.41, No.8の特集「情報検索システムの力くらべ―テストコレクションによる評価―」（神門典子編）などを参照いただきたい）．また，NTCIR発足以前に用いられていたBMIR-J1およびJ2という小規模な日本語テストコレクションも正解レベルの情報を持っていた[12]．

本章ではまず，前章の前半で紹介した指標と同様に「正解をたくさん検索する」タスク向けであり，かつ正解レベルを考慮できるものを紹介する．すなわち，これらは「大正解をたくさん検索する」タスクのための指標である．次に，前章の後半で紹介した逆数順位と同様に「正解を1つだけ検索する」タスク向けであり，かつ正解レベルを考慮できるものを紹介する．すなわち，これらは「大正解を1つだけ検索する」タスクのための指標である．

大正解をたくさん検索するタスクのための指標

正解レベルをうまく扱うために，まず，ACM SIGIR 2000においてJärvelinとKekäläinenが提案した**ご褒美の貯金**という考え方をご紹介する（彼らはこの年のベストペーパー賞を受賞した）．以下，正解レベルとして「大正解」「ふつうの正解」「おまけの正解」の3段階が与えられているとする．そして，システムがこれらの正解を

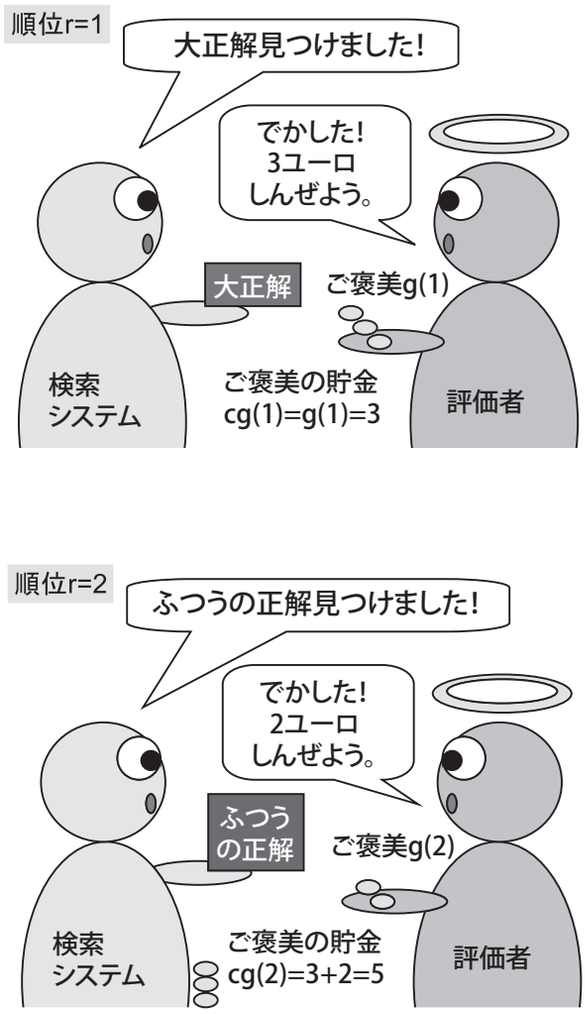

図-1　累積利得cg（ご褒美の貯金）のイメージ

1件検索するごとに，それぞれたとえば3, 2, 1ユーロの**ご褒美**を与えることにしよう☆4．これはたとえば，「大正解」は「おまけの正解」3つぶんの価値があると見なすことに相当する．

図-1をご覧いただきたい．ここでは，検索システムが1位に「大正解」を，2位に「ふつうの正解」を検索している．そこで，システムは1位においてご褒美$g(1) = 3$ユーロをもらう．次に，2位においてご褒美$g(2) = 2$ユーロをもらう．したがって，2位におけるご褒美の貯金は$cg(2) = 3 + 2 = 5$となる．もちろん，不正解を検索した場合はご褒美をもらえない．なおcgは「ご褒美の貯金」の略ではなく**累積利得**（cumulative gain）の略である（実はこの考え方は，60年代に考案された**スライド比**（sliding ratio）という評価指標においてすでに使われていた）．さて，ご褒美を一律1ユーロにして古典的な情報検索のように正解レベルをあえて無視すると，$cg(r) = count(r)$が

☆2　情報検索の国際会議．
☆3　http://research.nii.ac.jp/ntcir/index-ja.html
☆4　フィンランド人のJärvelinとKekäläinenに敬意を表して通貨単位をユーロとした．

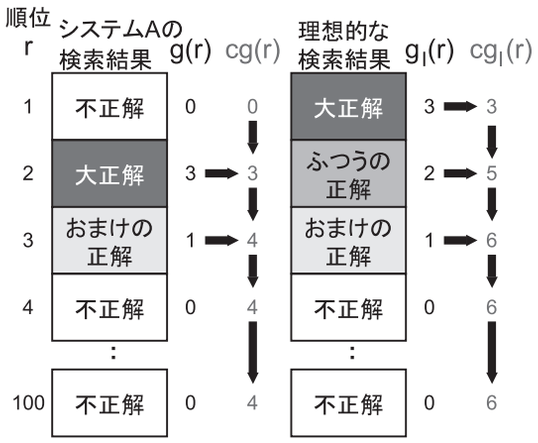

図-2　大正解を2位に，おまけの正解を3位に持つシステムAの累積利得

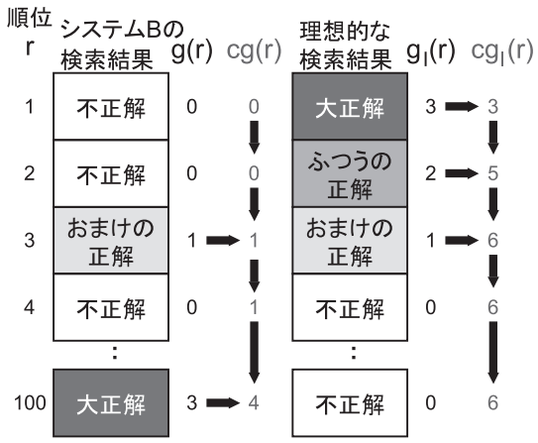

図-3　大正解を100位に，おまけの正解を3位に持つシステムBの累積利得

成り立つことがお分かりだろう．

JärvelinとKekäläinenは当初，**第l位における累積利得**$cg(l)$（および後述する**減価累積利得**$dcg(l)$）をそのまま全検索課題について平均し，評価指標として用いていた．しかし，これらの値は正解数（特に大正解数）の大きい検索課題についてはそれだけ大きくなってしまうので，きちんとした評価を行うには平均をとる前に**正規化**（normalization）を行うべきである．実際，JärvelinとKekäläinenは2002年に（第l位における）**正規化累積利得**（normalized cumulative gain）および**正規化減価累積利得**（normalized discounted cumulative gain）を提案している[4]．以下，まず正規化累積利得について説明する．

まず，与えられた検索課題に対する**理想的な**（ideal）検索結果というものを考えよう．これは，正解レベルの高い順に正解を列挙することにより得られる．たとえば簡単のために「大正解」「ふつうの正解」「おまけの正解」をそれぞれ1件ずつ持つ検索課題を考えると，これに対する理想的検索結果は**図-2**の右側のようになる．ここでは，理想的検索結果の第r位におけるご褒美およびご褒美の貯金をそれぞれ$g_I(r)$, $cg_I(r)$で表している．一方，図-2の左側はこの検索課題に対するシステムの検索結果の一例であり，この場合は「大正解」を2位に，「おまけの正解」を3位に検索できているが「ふつうの正解」はどこにも検索できていない（検索結果のサイズ$L = 100$としている）．

そこで，正解レベルに基づく情報検索指標として，「ご褒美の貯金に関する理想と現実のずれ」を測定するために以下を用いるとどうだろうか．

$$\text{第}l\text{位における正規化累積利得} = \frac{cg(l)}{cg_I(l)}. \qquad (6)$$

こうすれば，評価値はシステムの検索結果が理想的なものであるとき，かつそのときに限って1となり，検索課題セットに関して平均をとるのにも都合がよいのではないか．

ところが実際は，正規化累積利得では正当な評価が行えない場合がある．まず，図-2において100位における正規化累積利得を計算してみると，$cg(100)/cg_I(100) = 4/6$となる．一方，**図-3**のように，「大正解」が2位ではなく100位に検索されてしまったシステムを考えてみると，$g(2)$および$cg(2)$は0となるが，かわりに$g(100)$が3になるので結局$cg(100)$は4のままになる．すなわち，システムBの100位における正規化累積利得も4/6になってしまう．

図-2および図-3の一番右側の値を眺めてみると分かるが，問題は$cg_I(r)$の値が第R位以降は定数となることである．つまり，正解が全部でR件しかないのだから，R回ご褒美をもらったあとは貯金がいっこうに増えない．したがって，$cg_I(r)$を評価指標の分母としているかぎり，検索結果の下のほうに検索された正解に対して減点を行うことができない．全検索課題について定数lを用いる正規化累積利得に限らず，より一般に，以下で定義される第r位における**重みつき精度**（weighted precision）を基礎とした評価指標では，第R位より下に検索された正解を扱う場合に不具合が生じる．

$$WP(r) = \frac{cg(r)}{cg_I(r)}. \qquad (7)$$

上記の問題を解決するには少なくとも2つのアプローチがある．第1は，検索結果の下位のほうに進むにつれてシステムへのご褒美（式(7)の分子に相当）を小さくするものである．第2は，検索結果の下位のほうに進むにつれて理想的検索結果のほうの評価値（式(7)の分母に相当）を大きくするものである．

第1のアプローチでは，累積利得の代わりに**減損累**

積利得（discounted cumulative gain）を用いる．これは，個々のご褒美の値を順位のlogで割ってから貯金するものである．以下，第r位におけるシステムの検索結果の減損累積利得および理想的検索結果の減損累積利得をそれぞれ$dcg(r)$および$dcg_l(r)$と表記することにする．たとえばlogの底$b = 2$として，図-2および図-3をもう一度見てみよう．システムAが2位に持っている「大正解」に対するご褒美は，そのまま3ユーロとする（順位$r \leq b$の場合，**減損**（discounting）は行わない）．次に，システムAが3位に持っている「おまけの正解」に対するご褒美は，$1/\log_2 3 = 0.63$とする．したがって，$r \geq 3$について$dcg(r) = 3 + 0.63 = 3.63$となる．また理想的検索結果のほうも，3位の「おまけの正解」に対するご褒美が0.63になるので，$r \geq 3$について$dcg_l(r) = 3 + 5 + 0.63 = 8.63$となる．一方，システムBのほうは，3位に持っている「おまけの正解」に対するご褒美が$1/\log_2 3 = 0.63$，100位に持っている「大正解」に対するご褒美が$3/\log_2 100 = 0.45$と非常に小さな値になる．したがって，$dcg(100) = 0.63 + 0.45 = 1.08$となり，システムAの値よりも低い値となる．

以上より，第l位における正規化累積利得の代わりに，第l位における正規化減損累積利得を用いれば正解レベルを考慮した正当な評価が行えることが分かる．

$$\text{第}l\text{位における正規化減損累積利得} = \frac{dcg(l)}{dcg_l(l)}. \quad (8)$$

本誌Vol.46, No.9掲載の記事「マイクロソフト社独自開発のMSN Search Engine」（浅川，Selberg）によれば，マイクロソフト社はMSNサーチエンジンの精度向上のために実際に正規化減損累積利得（の一種）を利用しているそうである．

lというパラメタを持っていることから分かるように，正規化減損累積利得は「ユーザ指向」の評価指標である．ただ，実際にはlをいくつにするか（たとえば10, 100, 1000）により，システムの比較結果が大きく変わることが報告されている[7]．信頼性の高い評価結果を得るにはlを大きくとること，すなわちなるべく検索結果全体を見渡して評価を行うことが望ましい．また，（正規化）減損累積利得を用いる場合は，ご褒美の値に加えてlogの底bもパラメタとして決定しなければならない．たとえば忍耐力のないユーザを想定するなら，下位に検索された正解に対するご褒美をうんと小さくするように大きな底を用いればよい．

第2のアプローチでは，精度（式（3））と重みつき精度（式（7））を統合した第r位における**ブレンド比**（blended ratio）[7), 9)]を用いる．

$$BR(r) = \frac{cg(r) + count(r)}{cg_l(r) + r}. \quad (9)$$

ブレンド比は，重みつき精度と同様「理想と現実のギャップ」を測るが，分母にrを含んでいるため検索結果の下位に進むにつれて（すなわちrが大きくなるにつれて）低い評価値を与えることができ，前述の問題を解消している．また，その式の形から，ご褒美の値を大きく設定すると重みつき精度の性質が強くなり，小さく設定すると従来の精度の性質が強くなることが分かる（正規化（減損）累積利得では，「大正解」「ふつうの正解」「まあまあ正解」にそれぞれ3, 2, 1ユーロ与えることと，それぞれ30, 20, 10ユーロ与えることは等価であるが，ブレンド比の場合はご褒美の大きさ自体も1つのパラメタである）．また，ブレンド比は以下の性質を満たし，従来の精度の比較的自然な拡張になっている[9]．

- ご褒美の値を一律1ユーロにした場合，$r \leq R$のとき，かつそのときに限って$BR(r) = P(r)$となり，$r > R$のとき，かつそのときに限って$BR(r) > P(r)$となる．

Sakai[6]は平均精度（式（4））の精度の項をブレンド比に置き換えた評価指標**Q-measure**を提案している．

$$Q\text{-}measure = \frac{\sum_{1 \leq r \leq L} I(r) BR(r)}{R}. \quad (10)$$

上記を言葉で表現すると，「各正解が検索された時点でのブレンド比を，全正解について平均したもの」となる．

図-2についてQ-measureを計算してみよう．まず$BR(2) = (3 + 1)/(5 + 2) = 4/7$，そして$BR(3) = (4 + 2)/(6 + 3) = 6/9$であるからこれらの和を$R = 3$で割って$Q\text{-}measure = 0.62$となる．一方，図-3の場合は$BR(3) = (1 + 1)/(6 + 3) = 2/9$，そして$BR(100) = (4 + 2)/(6 + 100) = 6/106$で$Q\text{-}measure = 0.09$となる．

Q-measureは平均精度同様「システム指向」の指標であり，Q-measureによるシステムの比較結果は平均精度による比較結果と非常に相関が高い．さらに，正規化減損利得と平均精度の相関よりも，Q-measureと平均精度の相関のほうが高い[6), 7)]．以下にQ-measureに関する定理をまとめておく[9]．

- 検索結果が理想的なものであるとき，かつそのときに限って$Q\text{-}measure = 1$となる．
- ご褒美が一律1ユーロの環境下では，検索結果が第R位より下位に正解を含まないとき，かつそのときに限って$Q\text{-}measure$ = 平均精度となり，また第R位より下位に正解を含むとき，かつそのときに限って$Q\text{-}measure$ > 平均精度となる．

Q-measureは，もともと質問応答の評価のために考案された指標であるが（“Q”はQuestion Answeringの頭文

字），最近ではXML検索の評価型ワークショップINEX（INitiative for the Evaluation of XML retrieval）☆5のタスクに適用した研究事例も報告されている[10]．

第R位における精度として定義されるR-精度にならって，第R位におけるブレンド比である**R-measure** = $BR(R)$を考えることもできる．ただ，前述の重みつき精度の不具合はあくまで第R位より下位に検索された正解に対して起こるものなので，わざわざブレンド比を持ち出さず**第R位における重みつき精度**$WP(R)$により評価を行っても差し支えはない．ただし，前述のようにQ-measureが理想的検索結果のみに対して1となるのに対し，R-measureや第R位における重みつき精度は上位R件がすべてなんらかの正解でありさえすれば1になってしまう．たとえば図-2の例において，1～3位に「おまけの正解」「ふつうの正解」「大正解」をこの順番で検索したシステムに対しても満点を与えてしまう．この意味で，これら2つの指標は正解レベルを十分に活用しているとは言えない．

以上で紹介した「大正解をたくさん検索する」タスクのための評価指標の中では，「ユーザ指向」の正規化減損累積利得と，「システム指向」のQ-measureが最も安定性が高く，また，これらの指標の安定性は平均精度のそれと少なくとも同程度であることが実証されている[7]（ただし正規化減損累積利得のパラメタlは大きくとる必要がある）．さらに，この2つの評価指標の考え方を組み合わせて，たとえば「ユーザ指向」のブレンド比$BR(l)$や，「システム指向」の正規化減損累積利得

$$\frac{\sum_{1 \le r \le L} I(r)(dcg(r)/dcg_l(r))}{R} \qquad (11)$$

を考えることもできる．

Q-measure以外にも平均精度を一般化して正解レベルを扱えるようにした提案はある．Kishida[5]は**一般化平均精度**（generalized average precision）を提案し，簡単な模擬データを用いた考察をもとにその性質がQ-measureと正規化減損累積利得との中間に位置するのではないかと述べている．一方，VuとGallinari[10]も平均精度を一般化しQ-measureと比較しているが，こちらは正規化の仕方がまずいため理想的検索結果に対して満点を与えることができない．

なお残念なことに，アジア言語情報検索に関する評価型ワークショップNTCIRは，正解レベルを持つテストコレクションを作成しているにもかかわらず，システム評価には今のところTRECにならって平均精度を使い続けている（ただしNTCIRのWebタスクでは正規化なしの減損累積利得$dcg(l)$が用いられている）．これではせっかくつけた正解レベルの情報が無駄になってしまうので，平均精度に加えて正規化減損累積利得やQ-measureなどの信頼性の高い評価指標を公式に採用するよう勧めたい．

大正解を1つだけ検索するタスクのための指標

本節では，大正解が1つ見つかれば十分な検索シーン向けの情報検索指標を紹介する．具体的には，正解レベルを扱えるように逆数順位を拡張する2つのアプローチを紹介する．

NTCIRのWebタスクの主催者であるEguchiら[2]は，逆数順位の拡張として**重みつき逆数順位**（weighted reciprocal rank）を提案している．以下，「大正解」「ふつうの正解」「おまけの正解」のような正解レベルをまとめて「X正解」と表記することにする（すなわち$X \in$ {大, ふつうの，おまけの}）．重みつき逆数順位は，各Xについてパラメタβ_X (>1)を必要とする．これは前述のご褒美とは逆に，正解レベルが高いほど低い値に設定しなければならない．そこで以降，$\beta_{大} = 2$, $\beta_{ふつうの} = 3$, $\beta_{おまけの} = 4$としよう（ブレンド比におけるご褒美と同様，β_Xの大きさ自体も1つのパラメタである）．そして，検索結果が正解を含まないとき，重みつき逆数順位を0と定義する．そうでない場合，逆数順位の場合と同様に検索結果中の最も上位にある正解の順位をr'とし，さらに第r'位における正解の正解レベルをX'で表すことにする．このとき，

$$\text{重みつき逆数順位} = \frac{1}{r' - 1/\beta_{X'}} \qquad (12)$$

である☆6．

図-4に重みつき逆数順位の計算例を示す．この例から重みつき逆数順位の以下の特徴が分かる．

(a) 値が1を超える場合がある．
(b) 1位に「おまけの正解」を検索したシステムCのほうが，2位に「大正解」を検索したシステムDよりも優れていると見なされる．
(c) 1位に「おまけの正解」を検索したシステムCと，1位に「おまけの正解」を，2位に「大正解」を検索したシステムEは同等であると見なされる．

まず(a)についてであるが，重みつき逆数順位の上限値は，ある検索課題の正解データの中で最も高い正解レベルをYとしたとき$1/(1 - 1/\beta_Y)$で与えられる．ここ

☆5 http://inex.is.informatik.uni-duisburg.de/

☆6 原典[2]では，この式の分子も{0, 1}の値をとるパラメタである．しかしこれは「そもそもどこまでを正解と見なすか」を決めるものであり重みつき逆数順位の本質を表すものではないため，ここでは1としている．また，本稿における重みつき逆数順位の定義は原典の定義と見かけ上異なるが，両者が等価であることは容易に証明できる[8]．

で，一般にはYがテストコレクションの全検索課題で共通であるとは限らない．たとえば，「大正解」を持つ検索課題と持たない検索課題があるかもしれない．前述のβ_Xの値を用いた場合，「大正解」を持つ検索課題についての重みつき逆数順位の上限値は$1 - /(1 - (1/2)) = 2$であるが，「大正解」は持たずに「ふつうの正解」と「まあまあ正解」のみを持つ検索課題についてのそれは$1 - /(1 - (1/3)) = 1.5$である．したがって，このようなテストコレクションにおいて重みつき逆数順位の平均をとるのは問題がある．また，たとえYが全検索課題について共通であるとしても上限値がβ_Yに伴って変化するのは不便なので，一般には重みつき逆数順位を各検索課題の上限値で割った**正規化重みつき逆数順位**（normalized weighted reciprocal rank）を用いるべきだろう．

$$\text{正規化重みつき逆数順位} = \frac{1 - 1/\beta_Y}{r' - 1/\beta_{X'}}. \qquad (13)$$

次に(b)についてであるが，これが重みつき逆数順位の本質である．つまり，重みつき逆数順位は「高い正解レベルの正解を検索することよりも，まずはなんらかの正解を上位に検索することが先決」という仮定に基づいている．この性質は，パラメタβ_Xをどのように調整しても変わらない．したがって，図-4のシステムDをシステムCよりも高く評価したいような利用シーンには適さない．

そして（c）についてであるが，これは重みつき逆数順位があくまで最上位に検索されたなんらかの正解（すなわち第r'位の正解）をもとに評価を行っていることに起因する．これは，「ユーザはなんらかの正解（おまけの正解でもよい）を発見した時点で満足する」という仮定に相当する．この仮定が正しいかどうかは，少なくとも検索インタフェースに依存するだろう．たとえば，ユーザに提示されるランクつきリストが各文書のタイトルを含んでいるが，タイトルを読んだだけでは内容が推測しにくい場合，ユーザは図-4のシステムEが2位に持っている「大正解」にはまったく気づかずに，1位の「おまけの正解」を見た時点でリストを破棄するかもしれない．あるいは，検索結果のリストをユーザにまったく提示せずに「next」ボタンにより文書を1件ずつ提示するインタフェースにおいても，ユーザは次に「大正解」が来るかどうかまったく分からないので最初に見つけた「おまけの正解」で満足するかもしれない．上記の仮定はこのような場合には当てはまるだろう．一方，良質な検索エンジンの検索結果リストのように，ある程度どの文書が正解でありそうか想像がつくような検索シーンにおいては，ユーザは図-4のシステムEが2位に持っている「大正解」に直接アクセスするかもしれず，上記の仮定は妥当ではないだろう．

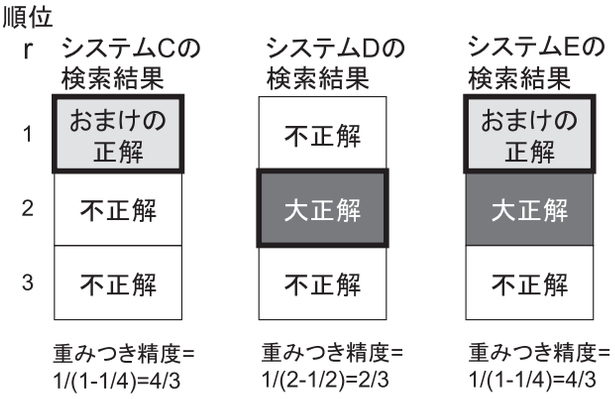

図-4　重みつき平均の計算例 ($\beta_{大}=2, \beta_{ふつうの}=3, \beta_{おまけの}=4$)

なお，NTCIRのWebタスクのために提案された重みつき逆数順位は，実際には上記タスクで活用されてはいない．より具体的には，パラメタβ_Xをすべての正解レベルXについて無限大（∞）に設定することにより，式(12)を式(5)に帰着させた上で評価が行われている．すなわち，実際に使われているのは従来の逆数順位にほかならない．

さて，「大正解を1つだけ検索する」タスクのための評価指標としてもう1つ，**O-measure**というものを紹介しよう[9]（"O"は「1つ（one）」の頭文字である）．O-measureは「大正解をたくさん検索する」タスクのための指標であるQ-measureやR-measureの親戚で，（重みつき）逆数順位と同様にr'を用いて以下のように定義される．

$$O\text{-}measure = BR(r') = \frac{g(r') + 1}{cg_1(r') + r'}. \qquad (14)$$

すなわち，O-measureとは最初に見つかった正解の順位におけるブレンド比である（r'は最初の正解なので，$cg(r') = g(r')$および$count(r') = 1$が成り立つ）．Q-measureが全正解のブレンド比を調べるのに対し，O-measureは特定の正解のブレンド比のみを調べるため，O-measureによる評価結果の安定性はQ-measureの場合に比べるとだいぶん低いが，逆数順位に比べると高い．以下に，O-measureに関する定理をまとめておく．

- 検索結果が最も正解レベルの高い文書の1つを最上位に持つとき，かつそのときに限って$O\text{-}measure = 1$となる．
- ご褒美が一律1ユーロの環境下では，$r' > R$のとき，かつそのときに限って$O\text{-}measure =$ 逆数順位となり，$r' > R$のとき，かつそのときに限って$O\text{-}measure >$ 逆数順位となる．

上記の1つ目の定理から，O-measureは平均を取るのに適した指標であることが分かる．また，O-measureは重みつき逆数順位の（b）の仮定（「高い正解レベルの正解

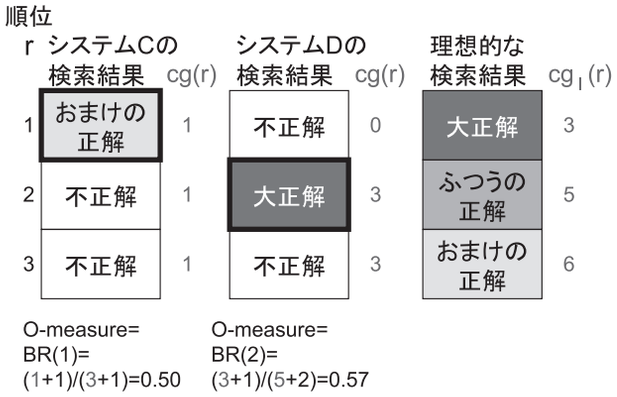

図-5　O-measureの計算例（「大正解」「ふつうの正解」「おまけの正解」にそれぞれ3, 2, 1ユーロ与える場合）

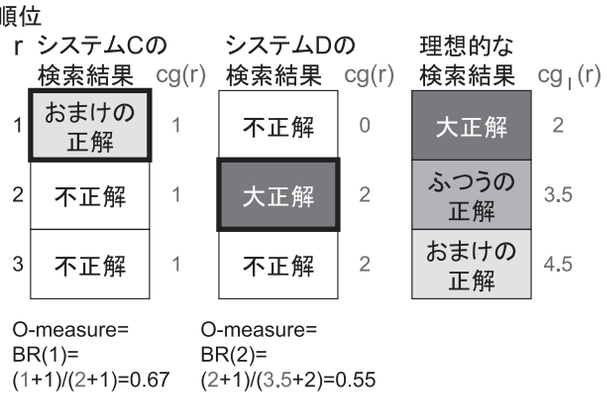

図-6　O-measureの計算例（「大正解」「ふつうの正解」「おまけの正解」にそれぞれ2, 1.5, 1ユーロ与える場合）

を検索することよりも，まずはなんらかの正解を上位に検索することが先決」）とは無縁である．このことを**図-5**および**図-6**を用いて説明しよう．図-5は「大正解」「ふつうの正解」「おまけの正解」を各1件ずつ持つ検索課題について，検索された「大正解」「ふつうの正解」「おまけの正解」にこれまでどおりそれぞれ3, 2, 1ユーロのご褒美をあげる場合のO-measureの計算例である．図の右端にあるのが理想的検索結果であり，図-4からそのまま拝借したシステムCおよびDの評価値がそれぞれ計算されている．このご褒美体系のもとでは，2位に「大正解」を検索したシステムDが，1位に「おまけの正解」を検索したシステムCに勝っている．一方，図-6は正解レベルの影響を減らすために「大正解」「ふつうの正解」「おまけの正解」にそれぞれ2, 1.5, 1ユーロ与えることにした場合の計算例である．この場合は重みつき逆数順位の場合と同様にシステムCがシステムDに勝っている．以上のように，O-measureは正解レベルを重視するか順位を重視するかをご褒美の調整により制御できる．

最後に，重みつき逆数順位のところで述べた（c）の仮定（「ユーザはなんらかの正解を発見した時点で満足する」）であるが，これはO-measureにも当てはまる．すなわち，O-measureもあくまで第r'位の正解のみに基づいて評価を行うため，図-4のシステムEに対して常にシステムCと同様の評価を下す．しかし前述の通り，システムEの検索結果を評価する際，1位の「おまけの正解」よりもむしろ2位の「大正解」に基づいて評価を行う指標が適切な検索シーンも実際にはあるかもしれない．Sakai[8]はこのような用途のために**P-measure**という指標を考案し，これがO-measureと同等以上の，かつ（正規化重みつき）逆数順位より高い安定性を示すことを確認している．

その他の新しい指標

本章ではこれまで，「大正解」「ふつうの正解」「おまけの正解」といった離散値の正解レベルに基づく評価指標を紹介してきたが，より極端に，正解レベルを連続値として扱おうとする研究もある．Della MeaとMizzaro[1]は，正解レベルを連続値として扱うだけでなく，システムに全検索対象文書の正解レベルの推定値を直接出力させるようにして，人手でつけた正解レベルとシステムが推定した正解レベルの絶対差を測ることによりシステムを評価することを提案している．しかし彼らの提案は，文書をランクづけしてユーザに提示する通常の情報検索とは異なる新しいタスクと見なすべきだろう[7]．また現実問題として，人手でどうやって連続値の正解レベルを付与するかといった課題も残る．

新しいタイプの検索への応用

本稿ではこれまで，いわゆる文書検索，すなわち文書の識別番号のランクつきリストを出力するタスクを前提に，個々の文書が正解か否か，あるいは個々の文書の正解レベルはどうかについて議論してきた．しかし最近では，このような枠組みでは手に負えない検索シーンが出現している．本章では質問応答，特許検索，XML検索を取り上げ，これらの新しいタイプの検索タスクにおける評価の難しさと，これまでに示した評価指標の適用可能性について議論する．

質問応答

質問応答システムとは，たとえば「ザ・ビートルズでベースを弾いていたのは？」という質問に対し，「ポール・マッカートニー」のように回答文字列をずばり出力

するシステムである．ユーザの目的が文書を読むことではなく回答を得ることである場合には，文書検索システムよりも質問応答システムを利用するほうが効率的だろう．特に本稿では，1999年のTREC-8において定義されたタイプの質問応答システムについて議論する．これは，ユーザが単一の質問を入力すると，人名，地名，組織名，数値などの回答候補をL件（たとえば5件）ランクつきで出力するというものである（本稿では，定義や原因を答えさせる質問や，一連の文脈をなす複数質問も扱わない．また，複数の回答候補に順位をつけずに出力する質問応答タスクも扱わない）．

さて，上記のように回答文字列に順位をつけて提示する質問応答システムを評価したい場合，これまでに説明したような検索評価指標が適用できるだろうか．

情報検索における「正解を1つだけ検索したい」タスクと同様に，質問に対する回答が1件見つかりさえすればよい場合には，逆数順位をそのまま適用できる．実際，TRECやNTCIRの質問応答タスクではシステム評価に逆数順位が用いられてきた．では，情報検索における「正解をたくさん検索したい」タスクと同様に，与えられた質問に対する多様な回答をなるべくたくさん見つけて欲しい場合にはどうだろうか．たとえば，「大リーグで活躍している日本人野球選手は?」という質問に対して，システムが1位に「イチロー」を，2位に「松井秀喜」を返してきたとする．このとき，「イチロー」だけでなく「松井秀喜」も考慮して評価するにはどうしたらよいか?

本稿を見返してみると，平均精度が自然な候補となる．ところが，実は平均精度のような「正解をたくさん検索する」ための指標を質問応答にそのまま適用することは難しい．なぜなら，文書検索がユニークな識別番号のリストを出力するのに対し，質問応答は任意の文字列のリストを出力するためである．たとえば先ほどの大リーグの例で，3位に「鈴木一朗」，4位に「松井」が出力されていた場合，どのように評価すべきだろうか？　さらに言えば，たとえば「松井」という回答よりも「松井秀喜」という回答のほうがより好ましいと感じるユーザもいるだろう．同様に，円周率をシステムに問うた場合の回答として，「3.14」のほうが「およそ3」よりも望ましいと感じるユーザもいるだろう．また，実際の質問応答システムは回答文字列を文書データから機械的に切り出すため，たとえば「好調イチロー」のような範囲的に不適切なものを出力する可能性がある．このような回答を「おまけの正解」としたい場合もあるかもしれない．質問応答の評価で正解レベルを扱うにはどうすればよいだろうか?

上記問題の解決策として，ここではNTCIR-4にて提案されたQ-measureを利用する評価方法を紹介する[6]．

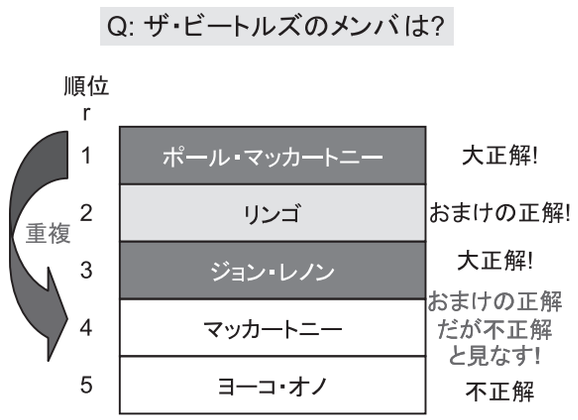

図-7　情報検索評価指標の質問応答への応用例

ここでは例として，「ザ・ビートルズのメンバは?」という質問を考えよう．このとき，正解データをたとえば以下のような形で用意しておく．

$A(1)$ = {ポール・マッカートニー（大正解），ポール（おまけの正解），マッカートニー（おまけの正解）}
$A(2)$ = {ジョン・レノン（大正解），ジョン（おまけの正解），レノン（おまけの正解）}
$A(3)$ = {ジョージ・ハリソン（大正解），ジョージ（おまけの正解），ハリソン（おまけの正解）}
$A(4)$ = {リンゴ・スター（大正解），リンゴ（おまけの正解），スター（おまけの正解）}

$A(i)$を回答の**同値クラス**（equivalence class）と呼ぶ．同値クラスは，「この中のどれか1つが出ていればOK」な回答文字列をまとめたものである．また，各回答文字列に対して正解レベルを付与していることに注意して欲しい．正解レベル付与の方針は用途に応じて決める必要があるが，ここでは単純にフルネームは「大正解」，姓あるいは名のみは「おまけの正解」としている．たとえば「故ジョン・レノン」を$A(2)$に含めるか，含める場合に正解レベルはどうするかといった判断は正解作成者に委ねられている．このような正解データの作成は大変だが，文書検索の場合に各検索課題につき数百の正解を用意していることを考慮すれば，不可能ではないだろう．

さて，上記のように回答同値クラスがうまく作成できた場合，質問応答システムについても「大正解がたくさん欲しい」場合の評価が行える．これには，**図-7**に示したように回答候補リスト中に含まれる**重複**（duplicate）を考慮してQ-measureなどの情報検索指標を計算すればよい．回答同値クラスの数は4であるから，$R = 4$である（なお，NTCIRのWebタスクにおいても重複を考慮した評価の試みがある[2]）．この例では，1位に「ポール・

マッカートニー」が，4位に「マッカートニー」が出力されているが，より下位にある「マッカートニー」のほうを「不正解」と見なすことにより回答リストの冗長さに対しペナルティを課している．従来のように逆数順位を使った評価では，この図における2位以下の回答がすべて無視され，また，1位の回答候補が「ポール・マッカートニー」だろうが「ポール」だろうが評価値は1となってしまうことに注意して欲しい．

なお，質問応答の評価において同値クラスを設けるアプローチは完璧な解決方法とは言えない．同値クラスの構成が破綻する場合があるからである．たとえば，「叶恭子」，「叶美香」，「叶姉妹」という回答文字列をすべて正解として扱いたい場合どうすればよいだろうか．

$A(1)$ = {叶恭子（大正解），恭子（おまけの正解）}
$A(2)$ = {叶美香（大正解），美香（おまけの正解）}
$A(3)$ = {叶姉妹（ふつうの正解）}

として，「叶」という回答は不可とするか？　あるいは，

$A(1)$ = {叶恭子（大正解），叶美香（大正解），叶姉妹（ふつうの正解），恭子（おまけの正解），美香（おまけの正解），叶（おまけの正解）}とするか？　それぞれの場合に，どのようなシステムが高く評価されるか考えてみていただきたい☆7．

さらに質問応答の場合，各回答候補と一緒に**根拠文書**（supporting document）と呼ばれるテキストデータが提示される場合がある．これらも含めて評価を行おうとすると話はさらに複雑になる．たとえば図-7において，2位の「リンゴ」という文字列をクリックしてみると以下のような根拠文書が提示されたとする．

「…ザ・ビートルズは青**リンゴ**のマークで有名なアップル・レコード社を設立した．」

すなわち，ユーザに提示された「リンゴ」が実は「リンゴ・スター」のことではなく果物のことであったとする．この場合は，「リンゴ」という回答を「不正解」と見なしたくなるだろう．

特許検索

本節では，特許検索，特に**無効資料調査**（invalidity search）[3] と呼ばれる文書検索タスクの一種について簡単に述べる．無効資料調査とは，たとえば新しい特許出願を検索要求としたとき，そこで請求されている権利を無効化できる既存の特許公報などを検索するものである．もしも既存の特許（すなわち正解）が単独で新しい特許を無効化できるのなら，問題は従来型の情報検索とまったく一緒であり，実際の利用シーンに即して「（大）正解をたくさん検索する」ための指標か，「（大）正解を1つだけ検索する」ための指標を用いて評価を行えばよい．

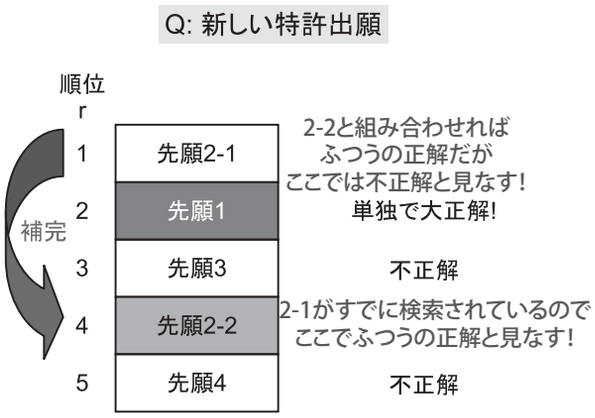

図-8　情報検索評価指標の特許検索への応用例

問題は，実際には複数の既存特許の組合せにより新しい特許を無効化できるケースが存在することである．すなわち，単独では半人前だが，他の特許を組み合わせると一人前に正解として機能する特許が存在する．このような正解のあり方を情報検索らしい用語でいうと**組み合わせ適合性**（combinatorial relevance）となる．

たとえば，単独で権利を無効化できる特許「先願1」と，半人前の特許「先願2-1」と「先願2-2」があり，この2つの半人前は合わせると一人前になるとしよう．このとき，従来の情報検索指標がうまく使えるように，以下のような構成の正解データを用意してみよう．

$A(1)$ = {先願1}（大正解）
$A(2)$ = {先願2-1, 先願2-2}（ふつうの正解）

ここでの$A(i)$は質問応答のものとは異なり，**補完クラス**（complement class）とでも呼ぶべきものである．質問応答の同値クラスの各要素がORで結ばれているイメージであるのに対し，補完クラスの各要素はANDで結ばれているイメージであるのはお分かりだろう．この場合，**図-8**のように評価を行えば，既存の情報検索評価指標がそのまま適用できると考えられる．この例では，まず1位に先願2-1が検索されているが，この時点では$A(2)$の要素はまだ他にも残っているので不正解と見なす．次に，4位に先願2-2が検索された時点で$A(2)$は完全にカバーされるので，こちらは「ふつうの正解」と見なす．質問応答と逆の処理を行っていることがお分かりだろう．ただし，質問応答に対しては前述の同値クラスを用いた評価方法を適用した実績があるのに対し[6]，特許検索について述べた上記の方法は現時点ではアイディアレベルである．

☆7　回答に順位をつけない質問応答タスクにおいても同値類の破綻問題は生じる．

XML検索

情報検索タスク「応用編」の最後に，XML検索について簡単に紹介しておこう．XML検索において通常の文書検索と同様にランクつき検索結果を出力するための研究は，2002年以来欧州で開催されている前述の評価型ワークショップINEXなどにおいて進められている[11]．XML検索の評価が難しいのは，そもそも「文書」という明確な検索単位が存在しないためである．つまり，ある「正解部分」を包含する1段落を検索して出力するシステムもあれば，さらにこの段落を包含する1ページ分のテキストを出力するシステムもある．また困ったことに，上記の両方を検索してしまう冗長なシステムもある．

「文書」という単位が存在しないため，INEXではなんと2次元の正解レベルを持つ正解データを作成している．1つ目の次元は，検索されたエレメント（段落，ページなど）が所望の情報をどれだけカバーしているかを表す**網羅性**（exhaustivity）であり，もう1つの次元は，検索されたエレメントがどれだけ所望の情報に絞った内容を含むかを表す**特定性**（specificity）である．システム評価の際にはこれらの正解レベルの組合せが見かけ上1次元の正解レベルとして扱われている．Q-measureなどをINEXのシステム評価に適用した試みもあるが[10]，このようなXML検索タスクの評価は質問応答，特許の無効資料調査にも増して複雑であり，まだ試行錯誤の段階にあると言えるだろう．

まとめ

本稿では，正解レベルを扱うことのできるもの，かつ比較的最近提案されたものを中心に，情報検索評価指標について簡単に解説した．また，古典的な情報検索の枠組みでは手に負えない質問応答，特許の無効資料調査およびXML検索という新しいタイプのタスクの評価の難しさについて紹介し，これらのタスクにおける情報検索評価指標の適用可能性について述べた．

評価指標は，別のデータを使って評価しても結果が変わらないかどうかを示す安定性や，システム間の差をいかに感度よく検出できるかを表す**判別能力**（discrimination power）などの観点からあらかじめ十分な検証を行った上で用いるべきだろう[7]．そして，信頼性の高い評価指標の中から実際の利用シーンに合った性質を持つものを選んで用いるべきだろう．

本稿で重点的に紹介した正解レベルに基づく評価指標の課題としては，「ご褒美」などのパラメタの設定方法があげられる．理想的には，ユーザの実際の満足度と評価指標との相関が高くなるようにご褒美体系を設定することが望ましいが，今のところ「何パターンかのご褒美体系を試してみる」という場当たり的なアプローチしかとられていない．最後に，繰り返しになるが，テストコレクションによる「実験室型」評価は必要条件であって十分条件ではないので，情報検索研究者はユーザの利用環境に直結したシステムの評価方法を追究し続ける必要がある．

謝辞　重みつき逆数順位の発案者である江口浩二先生（国立情報学研究所）には，その趣旨について個人的にご説明いただきました．ここに感謝いたします．

参考文献

1）Della Mea, V. and Mizzaro, S. : Measuring Retrieval Effectiveness : A New Proposal and a First Experimental Validation, Journal of the American Society for Information Science and Technology, Vol.55, No.6, pp.530-543 (2004).
2）Eguchi, K. et al. : Overview of the Web Retrieval Task at the Third NTCIR Workshop, Technical Report NII-2003-002E (2003).
3）Fujii, A., Iwayama, M. and Kando, N. : Overview of Patent Retrieval Task at NTCIR-4, NTCIR-4 Proceedings (2004).
4）Järvelin, K. and Kekäläinen, J. : Cumulated Gain-Based Evaluation of IR Techniques, ACM Transactions on Information Systems, Vol.20, No.4, pp.422-446 (2002).
5）Kishida, K. : Property of Average Precision and its Generalization : An Examination of Evaluation Indicator for Information Retrieval Experiments, Technical Report NII-2005-014E (2005).
6）Sakai, T. : New Performance Metrics based on Multigrade Relevance : Their Application to Question Answering, NTCIR-4 Proceedings (2004).
7）Sakai, T. : The Reliability of Metrics based on Graded Relevance, AIRS 2005 Proceedings, Lecture Notes in Computer Science 3689, pp.1-16 (2005).
8）Sakai, T. : A Further Note on Evaluation Metrics for the Task of Finding One Highly Relevant Document, 情報処理学会研究報告 2006-FI-82 (2006).
9）Sakai, T. : On the Task of Finding One Highly Relevant Document with High Precision, 情報処理学会論文誌：データベース TOD-29 (2006).
10）Vu, H.-T. and Gallinari, P. : On Effectiveness Measures and Relevance Functions in Ranking INEX Systems, AIRS 2005 Proceedings, Lecture Notes in Computer Science 3689, pp.312-327 (2005).
11）絹谷弘子 他：キーワードを利用したXML文書検索，情報処理学会論文誌：データベース TOD-22, pp.255-273 (2004).
12）酒井哲也 他 ： 情報検索システム評価のためのテストコレクション，Computer Today, Vol.9, No.87, pp.31-35 (1998).

（平成17年12月14日受付）