社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ＦＩ－ 89 (1)
2008－ＮＬ－183 (1)
2008／1／21

チュートリアル: 情報検索テストコレクションと評価指標
Information Retrieval Test Collections and Evaluation Metrics: A Tutorial

酒井 哲也†
Tetsuya Sakai

1. はじめに

近年，テレビのコマーシャルや電車の中の広告にさえ検索窓と検索ボタンが出てくることからもわかるように，**情報検索 (information retrieval)** は日常生活に浸透している．そして，携帯電話からブログを書き込む情報発信の環境なども整い始め，Web 上の情報量は増加の一途を辿り，膨大な情報の中からユーザに有用なものだけを探し当てる情報検索技術の重要性は増すばかりである．

情報検索技術の進歩には，検索システムを適切に評価する方法が不可欠である．もっと言えば，単独の研究機関による自前の検索手法の評価にとどまらず，多くの研究機関が同一条件下で検索手法の比較を行うことにより切磋琢磨していくことが望ましい．そこで本稿では，情報検索研究者が共有し，検索システムの評価を行うためのデータセットである**テストコレクション (test collections)** と，システムの良し悪しを定量化するための**評価指標 (evaluation metrics, evaluation measures)** について説明する．次に，これらを用いて検索結果の質，すなわち**検索有効性 (retrieval effectiveness)** を評価する一般的な方法について述べる．なお，検索システムがいかに早く結果をユーザに提供できるかを意味する**検索効率 (retrieval efficiency)** もまた重要な研究課題であるが，本稿では扱わない．さらに，実際の検索システムでは入出力インタフェースの良し悪しが非常に重要であるが，検索有効性の評価では，検索対象のうち何がどのような順序で検索され何が検索されなかったかという，検索システムの基本機能の側面のみを扱う．

本稿の構成は以下の通りである．2 章で情報検索評価用テストコレクションの概要とその作成過程について，3 章で検索システムの評価指標について述べる．4 章でテストコレクションと評価指標を用いた情報検索評価の方法について述べる．最後に，5 章で本稿のまとめを述べる．

2. テストコレクションとその作成過程

情報検索システムの検索結果を評価するには，被験者を何人も使って任意の**検索要求 (search request)** に関する検索を実際に行ってもらい，主観的な評価を行ってもらうのが一番手っ取り早いと思うかも知れない．最終的にシステムを使うのはユーザなのだから．しかし，ユーザが介入する評価実験には以下のような弱点がある．

- 被験者を雇い，検索結果をその都度評価してもらうにはコスト (お金と時間) がかかる．従って，実験の規模が小さくならざるを得ない．小規模な実験から得られる結果は信頼性に乏しい．

†(株) ニューズウォッチ
sakai@newswatch.co.jp
http://voice.fresheye.com/sakai/

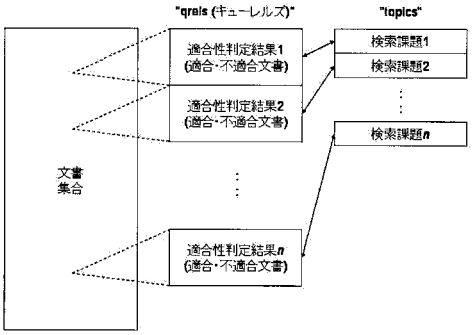


図 1: 情報検索テストコレクションの基本構成

- 被験者の主観的判断は一貫性に欠けることが多い．複数の被験者の判断が異なるばかりか，同一の被験者による同一の検索結果に対する評価も異なる場合がある．従って，同一条件下で実験を再現しようとしても，得られる結果が異なることが多い．

小規模で信頼性に乏しく，再現不能な実験から得られた結果の蓄積のみでは，情報検索技術の進歩は望めない．そこでテストコレクションは，実験環境からユーザという要素を思い切って排除することにより，客観的で再現可能，かつ大規模で効率的な評価実験を可能にしている．

2.1 テストコレクションとは

情報検索テストコレクションは，図 1 に示すように，検索対象である**文書コレクション (document collection)** と，検索要求を特定の形式で記述した**検索課題**の集合と，各検索課題に対応する**適合性判定結果 (relevance assessments, relevance judgments)** により構成される．米国において 1992 年より毎年開催されている情報検索の評価ワークショップ TREC (Text REtrieval Conference) [30] では，検索課題のことを **topics**，適合性判定結果のことを **qrels**‡と呼んでいる．

図 2 に，1999 年より日本で開催されている情報検索の評価ワークショップ NTCIR(エンティサイル) の検索課題の例を示す．これは，2007 年 5 月に成果報告会が開催された第 6 回 NTCIR (NTCIR-6) の言語横断情報検索

‡query-relevance set の略語だが [29]，通常，**クエリ (query)** とは検索課題をシステム可読な形式に変換したものを指す．適合性判定はシステム依存であるクエリに対してではなく，検索課題に対して行われるものなので，本来は topic-relevance set とでも呼ぶべきであろう．

```
<TOPIC>
<NUM>003</NUM>
<ONUM>NTCIR4-003</ONUM>
<SLANG>CH</SLANG>
<TLANG>JA</TLANG>
<TITLE>ES 細胞</TITLE>
<DESC>ヒト ES 細胞の紹介記事を探したい</DESC>
<NARR>
<BACK>2 つのアメリカの研究グループが，実験室でヒト ES 細胞の培養に成功した．かれらはこの研究が心筋から脳組織まで，いかなる組織の培養にも応用できると考えている．医学における ES 細胞の用途と特性について，また，倫理論争があるのかどうか，論争があるのであればどのようなものなのかについて知りたい．</BACK>
<REL>用途，医学的特性，倫理論争の紹介を含む文書を適合とする．科学者の研究過程は適合しない．</REL>
</NARR>
<CONC>ES 細胞，医学的特性，用途，倫理，論争</CONC>
</TOPIC>
```

図 2: 検索課題の例 (NTCIR-6 CLIR)

(Cross-Lingual Information Retrieval[§], CLIR) タスクで用いられたものである．図 3 に示したように，NTCIR-6 CLIR タスクの検索課題は，過去の NTCIR で作成された検索課題を別の文書コレクションに適用して再利用したものである．図 2 の検索課題がもともと NTCIR-4 で作成されたものであることは，ONUM というフィールドを見ればわかる．また，一口に検索課題と言っても，TITLE, DESC (description), NARR (narrative [¶]), CONC (concepts) などの複数のフィールドから構成されていることがわかる．この形式は，基本的に米国の TREC を踏襲している．仕様の詳細については NTCIR のサイト [13] などを参照いただきたい．

検索システムを評価する研究者は，用途に応じてシステムに入力するフィールドを選択すればよい．例えば，現状の Web 検索のように語の羅列が入力されることを想定したシステムの評価では，TITLE フィールドのみをシステムに与えて評価を行うとよいだろう．一方，簡潔な自然言語一文による入力が期待される場合には，DESC フィールドのみをシステムに与えればよい．さらに，ユーザが図 2 の TITLE，DESC，NARR，CONC のような情報をフィールド毎に全て入力してくれると仮定した上でシステム評価を行うことも可能である．実際，TREC や NTCIR では複数のフィールドを組み合わせた検索実験が多く行われている[‖]．

適合性判定結果とは，簡単に言うと「正解データ」のことである．通常，適合性判定は人手により行われ，適合性判定結果ファイルには，適合と判定された文書，および不適合と判定された文書が含まれる．適合・不適合という**二値適合性 (binary relevance)** に基づく判定だけでなく，例えば高適合・適合・部分適合・不適合のような**多値適合性 (graded relevance)** に基づく判定が行われる場合もある．

[§]Cross-Language Information Retrieval という用語が一般的となっているが，NTCIR では文法的により好ましい cross-lingual という語を採用しているようだ．

[¶]ナラティブと発音することに注意．

[‖]このような複数フィールドからなる詳細情報をユーザが本当にシステムに与えてくれるとは考えにくい．しかし例えば，DESC フィールドのみを使った場合の検索結果と DESC と NARR フィールドを併用した場合の検索結果とを比較し，NARR に含まれる情報の有用性を検証した上で，このような情報をシステムに自動獲得させる方法を模索する研究のアプローチはありうるだろう．

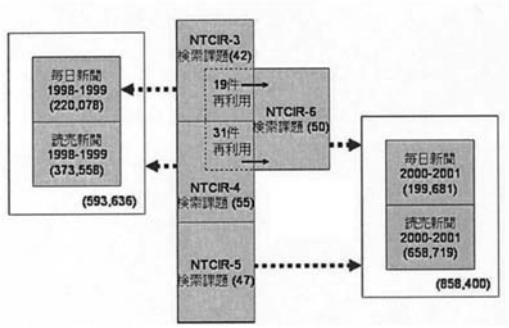
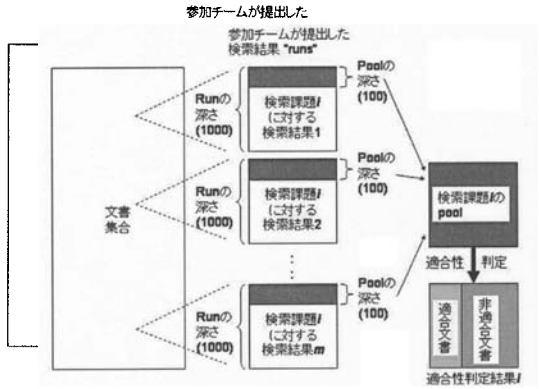


図 3: NTCIR-3〜6 言語横断タスク日本語テストコレクションの関係



文書コレクションが小さい場合には，各検索課題について全数調査により適合性判定結果を作成することが考えられる．しかし，近年の数十万〜数百万文書を扱うテストコレクションにおいては，全数調査は不可能に近い．このため，TREC や NTCIR などの評価ワークショップでは，次節で説明する**プーリング (pooling)** という手法を用いて適合性判定を行っている．

表 1 に，英語および日本語の**文書検索 (document retrieval)** のために作成された新旧テストコレクションの例を示す．詳細については表中に示した文献を参照されたい．なお，文書検索は歴史的には情報検索とほぼ同義であるが，文書の一部のみを検索するパッセージ検索などとともに情報検索の下位概念と位置づけることも可能である．

2.2 プーリング

図 4 に，大規模テストコレクションの適合性判定のためのプーリングの仕組みを示す．TREC や NTCIR のような評価型ワークショップでは，まず主催者が参加チー

表 1: 情報検索テストコレクションの例

名前	言語	文書数	文書の種類	検索課題数	適合性判定
Cranfield [5]	英語	1,400	航空学文献抄録	225	多値
TREC-8 ad hoc [27]	英語	528,155	主にニュース ("TREC Disks 4&5")	50	二値
TREC 2005 HARD/Robust [1, 31]	英語	1,033,461	ニュース ("AQUAINT コーパス")	50	多値
TREC Terabyte 2004-2006 [4]	英語	25,000,000	.gov ドメインの Web ページ	149	多値
BMIR-J2 [14]	日本語	5,080	新聞	50	多値
NTCIR-2 CLIR (日本語部分のみ) [8]	日本語	400,248	科学技術文献抄録	49	多値
NTCIR-6 CLIR (日本語部分のみ) [8]	日本語	858,400	新聞	50	多値
NTCIR-6 PATENT (日本語部分のみ) [6]	日本語	3,496,252	特許全文	1,685	多値

```
003 0 JA-000724057 1 10000.000000 TSB-J-J-D-01
003 0 JY-20010614J10YMAA1400020 2 9514.784350 TSB-J-J-D-01
003 0 JA-000905037 3 9286.559150 TSB-J-J-D-01
003 0 JA-011105003 4 9249.715280 TSB-J-J-D-01
003 0 JA-000221055 5 9205.224570 TSB-J-J-D-01
```

図 5: trec_eval 形式の検索結果の例 (NTCIR-6 CLIR における TSB-J-J-D-01 の一部)

ムに検索対象となる文書コレクションと検索課題セットを配布する．この時点では，これらの検索課題に対する適合性判定結果は存在しない．次に，参加チームは，各検索課題に対する検索結果を作成し，これらをひとつのファイルにまとめて提出期限までに主催者に提出する．これらの検索結果のことを通常 **runs** と呼ぶ．図 5 に，NTCIR-6 CLIR タスクに実際に提出されたある run の一部を示す．これは TREC で配布されている情報検索評価用プログラム trec_eval[30] の入力形式に従っている．フィールドはそれぞれ，検索課題番号，ダミー，文書 ID，順位，システム独自の文書スコア，run の名前を表している．

TREC や NTCIR では各検索課題につき高々rd 件の文書 ID を順位つきで出力することになっている．この上限値 rd を **run depth** と呼ぶ．通常は 1,000 件である．例えば検索課題が 50 件ある場合は，提出するファイルの行数は高々50,000 行となる．一般には，ひとつの参加チームが複数の runs を提出する．一方，主催者は，図 4 に示したように，提出された各 run から上位 pd 件の文書 ID のみを取り出して，重複を除く．この文書 ID の集合を**プール (pool, document pool)** と呼び，pd のことを **pool depth** と呼ぶ．通常，pd は 100 件程度である．次に主催者は，適合性判定者にプールに含まれる全文書に対して適合性判定を行わせる．適合性判定は人間の主観に基づく作業であるので，複数の適合性判定者が同一の検索課題に対する適合性判定を行うと，当然食い違いが生じる．しかし，この食い違いはシステム間の優劣の比較にはほぼ影響を与えないという報告がある [28]．

プーリングの結果，検索対象の文書コレクションに含まれる文書は，検索課題毎に以下の 3 種類に分類される：

(a) **判定済適合文書 (judged relevant documents)**. 多値適合性による判定の場合，高適合・部分適合文書なども含む．

(b) **判定済不適合文書 (judged nonrelevant documents)**.

(c) **未判定文書 (unjudged documents)**.

図 4 から，文書コレクションが大きい場合には (a) および (b) よりも (c) の割合のほうが格段に大きくなることが想像できる．このように，適合性判定が網羅的に行われていないテストコレクションもしくは適合性判定結果のことを，**不完全 (incomplete)** であるという．通常，情報検索評価を行う場合は，(b) のみならず (c) も不適合文書と見なして評価を行うので，テストコレクションの不完全性がシステム評価に与える影響の検証は重要である [2, 12, 21, 24]．少なくとも，プーリングにより作成されたテストコレクションを利用する際には，(c) の未判定文書を全て不適合と仮定した上で評価を行っているという事実に留意する必要がある．

また，不完全さと密接に関わる概念として，適合性判定結果の**偏り (bias)** がある．これは，真の適合文書集合のうち，特定の性質を満たすもののみが適合性判定結果に含まれている状況を表す．例えば，テストコレクション作成のためのプーリングに参加しなかったシステムを，このテストコレクションにより評価することを考える．この新しいシステムがプーリング参加システムとは全く異なる検索手法を用いているとすると，このシステムは過小評価される可能性がある．なぜなら，このシステムは，プーリング参加システムのいずれもが検索できなかった多くの適合文書を検索できているかも知れず，そのような文書は全て不適合扱いになってしまうからである．新しいシステムの評価を公正に行うことができるテストコレクションを，**再利用可能 (reusable)** であると言う．なお，表 1 の TREC 2005 HARD/Robust コレクションの適合性判定結果は，プールの大きさが充分でなかったがために，検索課題の TITLE に出現する語を含む文書に偏ってしまっているという報告がある [3]．従って，テストコレクション利用の際にはそのテストコレクション固有の性質にも注意を払う必要がある．

3. 情報検索の評価指標

3.1 再現率・精度曲線と平均精度

検索対象である文書集合が極めて小さい場合，システムはこの文書集合の部分集合を出力すれば充分であろう．検索結果を集合として出力することを **set retrieval** と

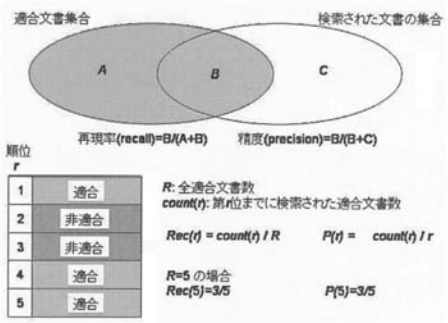


図 6: 再現率と精度

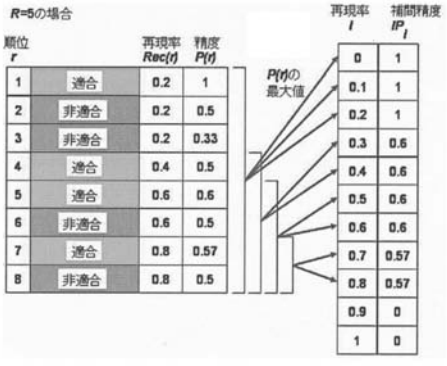


図 7: 補間精度の算出方法

いう．また，検索対象が大規模であっても，例えば特許検索における**ブール検索 (Boolean retrieval)** のようにset retrievalを行うケースはある．このような場合のシステムの検索有効性は，図6の上に示したようなベン図を考えた上で，**再現率 (recall)** および**精度 (precision)** により表すことができる**.「もれの少なさ」を示す再現率と「ごみの少なさ」を示す精度がトレードオフの関係にあることはお分かりだろう．

さて，近年のWeb検索のように検索対象の量が膨大になってくると，システムにより検索結果の各文書に優先順位をつけて出力する**ランクつき検索 (ranked retrieval)** が必須となる．この場合，検索結果の上位 r 件に着目すれば，再現率 $Rec(r)$ および精度 $P(r)$ を図6のように計算できる．すなわち，特定の検索課題の全適合文書数を R，第 r 位までに含まれる適合文書数を $count(r)$ とすると，$Rec(r) = count(r)/R$，$P(r) = count(r)/r$ である．当然，r を大きくすると，新しい適合文書が出てくる可能性があるので再現率の値は一般に大きくなる．一方，こうすると不適合文書の比率が高まり，精度の値は小さくなる可能性がある．このトレードオフを視覚化し，ランキング検索における検索結果全体の質を表現したものが**再現率・精度曲線 (recall-precision curve)** である[32].

実際に再現率・精度曲線を描く手順を図7に示す．この例では $R = 5$ である．まず，同図の左にあるように，検索結果の各順位 r における $Rec(r)$ と $P(r)$ を計算する．次に，11点の再現率 $i \in \{0.0, 0.1, 0.2, \ldots, 1.0\}$ に対応する**補間精度 (interpolated precision)**

$$IP_i = \max_{r,\ Rec(r) \geq i} P(r)$$

を算出する．ただし，$Rec(r) \geq i$ を満たす再現率が達成されていない場合は，$IP_i = 0$ とする．図7の右側はこの様子を示している．

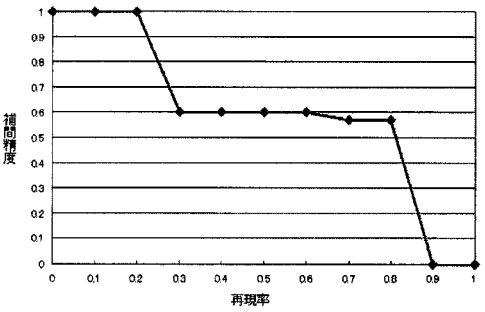


図 8: 再現率・精度曲線

図8に，図7に対応する再現率・精度曲線を示す．図7のもともとの精度 $P(r)$ が順位 r に対して単調減少になっていないのに対し，補間精度は再現率に対して必ず単調減少になることに注意して欲しい．また，この例は単一の検索課題に関する曲線であるが，検索課題セット全体に関する再現率・精度曲線を描くには，各再現率ポイント i 毎に，IP_i を全検索課題について平均すればよい．

実際に複数のシステムについて再現率・精度曲線を描くと，曲線が交差し，システムの優劣の議論がしにくいことが多い．そこで，システムの検索有効性を単一の数値で表す評価指標の出番となる．その一つは上記の11点の再現率における補間精度を平均した**11点平均精度 (11-point average precision)** であるが，TREC やNTCIR においては，これとは異なる**(補間なし) 平均精度 ((noninterpoloated) Average Precision)** が広く用いられている．平均精度は，11点平均精度よりも検索結果の上位の変動に敏感であるという好ましい性質を有している．

検索結果の第 r 位における文書が適合しているか否かを表すフラグを $I(r)$ で表すと，平均精度は以下で定義

精度のことを適合率**と言う人もいる．たしかにこのほうが再現率との語呂がよいが，適合率とは本来，precision よりも前に使われていた **relevance ratio** という用語の訳語である．

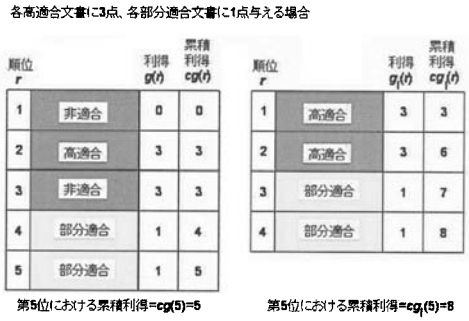


図 9: 累積利得 (cumulative gain)

される：

$$AP = \frac{1}{R}\sum_{r} I(r)P(r)$$

これを言葉で表すと，「全ての適合文書について，その適合文書が検索された時点での精度を計算し，これらの値を平均する．ただし，検索されなかった適合文書についての精度は 0 と見なす」となる．さらに，平均精度を検索課題セットに関して平均したものを **MAP (Mean Average Precision)** と呼ぶ．

以上で説明した再現率・精度曲線や平均精度は，あまり深く考えなくても TREC により配布されている C プログラム trec_eval により計算できる．しかし筆者は，一度は自分で評価プログラムを実装してみることを強くお勧めする．既存の評価方法を鵜呑みにせずに，再現率・精度曲線や平均精度がどのような意味合いをもっているか，本当に妥当かといった問題意識をもつことは，情報検索研究者であれば必要なプロセスであろう††．

3.2 nDCG

平均精度は二値適合性向けの評価指標であり，このままでは多値適合性を扱うことができない．従って，平均精度により検索システムを評価している限り，高適合文書を部分適合文書よりも優先して検索できるシステムの実現は難しいだろう．ところが，表 1 に示したように，多値適合性に対応したテストコレクションがたくさん作成されているにもかかわらず，現実には未だに平均精度を用いるのが主流である．これは，TREC が多値適合性に対応したテストコレクションを作成し始めたのが比較的最近であることに加え，NTCIR などの他の評価ワークショップが基本的に TREC のやり方を踏襲する傾向があることに起因していると考えられる．

多値適合性に対応した評価指標として比較的広く用いられているのは，**nDCG (normalized Discounted Cumulative Gain)** であろう [7]．例えば，情報検索の分野で研究が活発なマイクロソフト研究所は，自社の Web 検索エンジンの評価に nDCG の一種を採用しているという．これについては後述する．

さて，例えば高適合・部分適合・不適合の 3 段階の多値適合性判定情報をもつ検索課題を考えよう．そして，高適合文書を 1 件検索する毎に 3 点，部分適合文書を 1 件検索する毎に 1 点，システムに「ご褒美」を与えることにしよう．例えば図 9 の左側の検索結果において，第 2 位における**利得 (gain)** を $g(2) = 3$ とする．さらに，第 r 位における**累積利得 (cumulative gain)** $cg(r) = \sum_{i \le r} g(i)$ を計算する．次に，この検索課題に対する理想的な検索結果というものを考える．仮にこの検索課題については $R = 4$ であり，高適合文書が 2 件，部分適合文書が 2 件用意されているとすると，理想的な検索結果は図 9 の右側の検索結果のように，高適合・部分適合の順に全て列挙したリストになる．理想的な検索結果の第 r 位における累積利得を $cg_I(r)$ で表すことにすると，この例では $r \ge 4$ について $cg_I(r) = 8$ となることがわかるだろう．すなわち，もうこれ以上「ご褒美」は出ない．

nDCG は基本的に，特定の順位におけるシステムの累積利得と理想的な累積利得の比を求めるものである．ただし，単純に利得を足し合わせた累積利得同士の比ではシステムを適切に評価することができない．このことは，例えば図 9 の左側の検索結果を並べ替えて，高適合文書を第 1 位に，部分適合文書を第 2 位および第 3 位にもってきたとしても，第 5 位における累積利得が依然として $cg(5) = 5$ となることから明らかであろう．すなわち，単純な累積利得では，適合文書の順位が下がってもペナルティを課すことができない．そこで nDCG では，第 r 位における利得を $\log_b(r)$ で割ることにより目減りさせた**減損利得 (discounted gain)** を累積する．より正確には，$r > b$ なる r については $dg(r) = g(r)/\log_b(r)$，$r \le b$ なる r については $dg(r) = g(r)$ を用いる．同様に，理想的な検索結果の減損利得を $dg_I(r)$ と表記することにすると，第 l 位における nDCG は以下のように計算される：

$$nDCG_l = \frac{\sum_{r \le l} dg(r)}{\sum_{r \le l} dg_I(r)}$$

上記における log の底 b は，ユーザの忍耐力を表すパラメタとして考案された．小さな b の値は，検索結果を下位のほうまで調べる忍耐力のないユーザに相当する．例えば，$g(100) = 3$ であるとき，$b = 10$ とすると $dg(100) = 3/\log_{10}(100) = 1.5$ であるが，$b = 2$ とすると $dg(100) = 3/\log_2(100) = 0.45$ 点しか与えられない．すなわち後者の場合，第 100 位に高適合文書があってもシステムに与えられるご褒美は小さい‡‡．

しかし，nDCG にはまだ不具合がある．特に，その不具合は b の値を大きくすると顕著になる [18, 20, 24]．これは，上記の $dg(r)$ の定義を見るとわかるように，log による除算は $r \le b$ の場合に適用できず，前述の累積利得の問題が解消されないためである．例えば，$b = 10$ とした場合，図 9 の左側の検索結果と，前述のこれを

††trec_eval にはいくつか罠があるので，自前のプログラムによる評価結果と是非比較してみて欲しい．例えば，trec_eval は図 5 のような検索結果中の順位は無視し，文書スコアにより内部でソートを行ってから平均精度などを計算する．また，-c というオプションをつけ忘れると，MAP などを算出する際，全検索課題数で割るかわりに run に含まれる検索課題の数で割ってしまう．

‡‡文献 [15] p.152 には「忍耐力のないユーザを想定するなら…大きな底を用いればよい」と，逆のことを書いてしまった．

並べ替えて改善した検索結果には，nDCG をもってしても同じ評価が下されてしまう．マイクロソフト研究所ではこの不具合を解消するため，r の値によらず一律に $dg(r) = g(r)/\log_b(r+1)$ とした nDCG の変形版を用いている．この変形版では，log の底 b はシステムと理想的検索結果との比をとる際にキャンセルし，nDCG の値に影響を与えない．一方，これは本来の nDCG が意図していたユーザの忍耐力を表すパラメタが失われてしまったことを意味する．

3.3 Q-measure

nDCG の他にも多値適合性を扱うことができる評価指標はある．筆者が提案した Q-measure[15, 16, 18] は平均精度の拡張であり，nDCG のような不具合もない．さらに，マイクロソフト版の nDCG や**一般化平均精度 (generalized average precision)**[11] とは違い，ユーザの忍耐力に相当するパラメタを有している [20]．Q-measure は通常の文書検索のみならず，XML 文書検索 [9] や質問応答 [19] の評価にも用いられた実績があり，また NTCIR-6 言語横断タスクにおいては，nDCG とともに補助的な指標として採用された．

まず，与えられた正の値 β に対して，**ブレンド比 (blended ratio)** を以下のように定義する：

$$BR(r) = \frac{\beta cg(r) + count(r)}{\beta cg_I(r) + r}$$

パラメタ β を 0 にすると $BR(r) = P(r)$ となることは明らかであろう．さらに，二値適合性の評価環境下では，$r \leq R$ のとき，かつそのときに限り $BR(r) = P(r)$ が成り立ち，$r > R$ のとき，かつそのときに限り $BR(r) > P(r)$ が成り立つ．

Q-measure は平均精度の $P(r)$ を $BR(r)$ に置き換えたものである：

$$Q\text{-}measure = \frac{1}{R}\sum_r I(r)BR(r)$$

パラメタ β を 0 にすると $Q\text{-}measure = AP$ となることは明らかであろう．また，例えば β を大きな値 (例えば 100) にすると，ブレンド比の式の分母の r の影響が小さくなり，適合文書が下位に検索された場合のペナルティが小さくなる．さらに，Q-measure は以下の性質を満たす：

(1) 検索結果が理想的なものであるとき，かつそのときに限り $Q\text{-}measure = 1$ となる．

(2) 二値適合性の評価環境下では，検索結果が第 R 位より下位に適合文書を含まないとき，かつそのときに限り $Q\text{-}measure = AP$ となる．また，第 R 位より下位に適合文書を含むとき，かつそのときに限り $Q\text{-}measure > AP$ となる．

なお，平均精度は上記の (1) と同様な条件を満たさないことに注意して欲しい．平均精度は高適合文書と部分適合文書の区別ができないため，検索結果の上位 R 件が全て適合してさえいれば，例え部分適合文書が高適合文書より上位にあっても $AP = 1$ となってしまう．

3.4 その他の評価指標

ここでは，平均精度，nDCG，Q-measure の他に，21 世紀になってから提案されたいくつかの評価指標について簡単に紹介する．

Moffat ら [12] が ACM SIGIR 2007 において用いた **Rank-Biased Precision (RBP)** は，その名のとおり精度を拡張した評価指標であり，再現率の要素を全くもたないという特徴をもつ．ユーザが第 r 位の文書から第 $(r+1)$ 位の文書に遷移する確率 p が一定値であるというモデルに基づく多値適合性に対応した指標である．RBP の場合，このパラメタ p が前述のユーザの忍耐力に相当する．しかし，この指標については最大値が必ずしも 1 にならない，**判別能力 (discriminative power)** が Q-measure や nDCG に比べると格段に低いなどの欠点が指摘されている [24]．ここでの判別能力とは，統計的検定における第一種の誤りの確率 α が与えられた場合に，全てのシステム対のうち，統計的有意差が得られた割合を意味する [17, 22]．

また，最近の TREC では，テストコレクションの不完全性に対応する試みとして **bpref (binary preference)** という評価指標が trec_eval に組み込まれ，平均精度と共に用いられている．しかし，この指標は検索結果から全ての未判定文書を取り除いた**短縮リスト (condensed list)** に対して平均精度を適用することと本質的に変わらず，かつ，検索結果上位の変動に鈍感であるという欠点をもつことが示されている [21]．この他に不完全性を直接扱う評価指標としては，**inferred average precision** [33] が知られている．これは，不完全な適合性判定結果をもとに，真の平均精度の値を推定しようとする試みである．いずれの指標も，多値適合性を扱うことができない．

TREC や NTCIR における主要な情報検索タスクのゴールは，適合文書をできるだけたくさん検索結果の上位に検索することである．しかし，例えば実際の Web 検索では，適合文書を 1 件見つけさえすればそれで充分というケースも多い．適合文書をひとつだけ検索するタスクのための評価指標として最も広く用いられているのは**逆数順位 (Reciprocal Rank)** である．検索結果が適合文書を含まない場合，逆数順位は 0 と定義される．さもなくば，検索結果中の最上位の適合文書の順位を r_1 としたとき，逆数順位は $1/r_1$ と定義される．

逆数順位を多値適合性に対応させた評価指標としては，第 r_1 位におけるブレンド比 $BR(r_1)$ として定義される **O-measure** [16] がある．同様の試みとしては NWRR(Normalized Weighted Reciprocal Rank)[15] があるが，システム順位の安定性や判別能力の観点から O-measure のほうが NWRR や逆数順位よりも優れていることが示されている．さらに，O-measure と Q-measure の中間に位置する指標として P^+ という指標もある [23]．逆数順位，NWRR，O-measure が「ユーザは，例え部分適合文書であっても何かしら 1 件見つければその時点で検索結果を調べるのをやめる」という仮定に基づくのに対し，P^+ は「ユーザは，なるべく適合レベルの高い文書を 1 件見つけるまで検索結果を調べ続ける」という仮定に基づく．詳しくは各文献を参照されたい．

4. 検索システムの評価の仕方

テストコレクションと評価指標を計算するプログラムさえ揃えば，検索システムの評価を行うことは容易である．一般的な手順は以下のとおりである：

1. 検索システムに文書コレクションを登録する．すなわち，**索引 (index)** を作成する．
2. 検索システムに検索課題をひとつずつ与え，各検索課題に対する検索結果を作成する．
3. 各検索課題について，適合性判定ファイルおよび検索結果を評価プログラムに入力し，評価指標の値を算出する．また，評価指標の検索課題セットに関する平均値などを算出する．

システムAとシステムBの総合力の優劣を議論したければ，それぞれの評価指標の平均値の差の統計的検定などを行うべきである．同一テストコレクションによりシステム対を評価する場合，各検索課題に対応するそれぞれの評価値を対になっているデータ (paired data) と見なし，t 検定 [10] やブートストラップ検定 [17, 22] を実施すればよい．t 検定は正規母集団を仮定している．ブートストラップ検定はこのような仮定を必要としないが，検索課題セットからの復元抽出を繰り返し行うため，少々計算機パワーを必要とする．なお，両者の統計的検定結果はよく似ていることが知られている．

ただし，統計的検定結果は「参考程度」と考えよう．仮に統計的有意差が得られなかったとしても，必ずしも絶望する必要はない．提案手法が有効でないことが示されたわけではなく，今回の実験では有効であるという充分な証拠が得られなかっただけなのだから．逆に，統計的有意差があったからといって安心してはいけない．その有意差は実際のユーザにとってどれほどの意味があるものだろうか? 小さな有意差を，一体どれくらいコツコツと積み上げていけばユーザがひと目見てわかるくらいのシステム改善になるのだろうか?

この他，情報検索の評価におけるチェック項目を列挙しておく．

- **複数のテストコレクションを用いること．** 単一のテストコレクションを用いた評価では，実験から得られた知見がそのテストコレクションにのみ当てはまるのか，他のデータに対してもある程度一般化できそうなのか，見当がつかない．もちろん，個々のテストコレクションの規模，特に検索課題数は多いほうがよい．
- **目的にあった評価指標を，できれば複数種類用いること．** 例えば，適合文書が1件見つかれば充分だと思われる利用シーンのために，平均精度で評価を行うのは適切ではない．また例えば，平均精度で見るとわからないが，上位10件の精度 ($P(10)$) で見ると傾向がわかる場合もある．何事も，いろいろな角度から見てみよう．
- **平均値の大小の議論に終始せず，統計的検定を行い，かつ，検索課題毎の分析を行うこと．** 特に，検索有効性が低かった検索課題について，何故うまくいかなかったかを詳細に分析することは非常に重要である．何故失敗したかを分析しないと，システムの改善は難しい．
- **適切なベースラインを用いること．** テストコレクションを用いて自分の提案手法をひとつだけ評価してもあまり意味がない．比較対象となる手法 (ベースライン) が必要である．ただし，ベースラインとして，原始的な tf-idf に基づく検索手法を用いるのは反則に近い．例えば Okapi BM25[25] のように，世の中には単純な tf-idf よりも格段に有効な手法が出回っている．このようなハイレベルなベースラインに対する提案手法の優位性を限定的にでも示すことができれば，それは情報検索技術のブレイクスルーにつながるかも知れない．
- **システムのチューニングは公正に行うこと．** 単一のテストコレクションに対して，検索システムのパラメタを徹底的にチューニングして，一番良かった結果を論文に載せるのは反則である．正しいやり方は，パラメタ調整のためのデータと，評価用データをきちんとわけることである．例えば，単一のテストコレクションを用いるにせよ，検索課題セットを訓練用と評価用に分割して用いる．検索課題数が少なければ，**交差検定法 (cross-validation)** という手もある．なお，評価対象のテストコレクションに対して最適化したシステムの結果を，理論的上限値という位置づけで論文中に示すことが有用である場合もあるだろう．

5. まとめ

本稿では，情報検索評価用テストコレクションの概要とその作成過程，検索有効性の評価指標，およびこれらを用いた情報検索評価の方法について述べた．特に，評価指標の説明に関しては，筆者の独断と偏見に基づく部分もあるかも知れないので，興味や疑問をもたれた方には是非ご自分で文献調査および実験を行っていただきたい．さらに，例えば「被験者が適合文書を1件見つけるまでの時間と平均精度の間には相関がない」[26] といった，テストコレクションを用いた「実験室型」評価について否定的な報告もあるので，このような研究についてもウォッチすることは必要であろう．

繰り返しになるが，検索評価用のプログラムは一度は自分で実装してみることをお勧めする．また，テストコレクションを用いた評価がユーザという要素を排除しているからこそ，自分はどのようなユーザのための検索システム実現を目指しているのかを意識した上で，評価指標や評価方法を選定すべきである．ブラックボックスの評価プログラムが出力する数値を論文の原稿に copy and paste しているばかりでは，日本における情報検索技術研究の大きな発展は望めないのではないだろうか．

参考文献

[1] Allan. J.: HARD Track Overview in TREC 2005: High Accuracy Retrieval from Documents, *TREC 2005 Proceedings* (2006)

[2] Buckley, C. and Voorhees E. M.: Retrieval Evaluation with Incomplete Information, *ACM SIGIR 2004 Proceedings*, pp.25-32 (2004)

[3] Buckley, C., Dimmick, D., Soboroff, I. and Voorhees, E.: Bias and the Limits of Pooling for Large Collections, *Information Retrieval*, Vol.10, Number 6, pp.491-508 (2007)

[4] Buttcher, S., Clarke, C. L. A., and Soboroff, I.: The TREC 2006 Terabyte Track, *TREC 2006 Proceedings* (2007)

[5] Cleverdon, C. W.: The Cranfield Tests on Index Language Devices, Aslib Proceedings, Vol.19, pp.173-192 (1967)

[6] Fujii, A., Iwayama, M. and Kando, N.: Overview of the Patent Retrieval Task at the NTCIR-6 Workshop, *NTCIR-6 Proceedings* (2007)

[7] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM Transactions on Information Systems*, Vol.20, Issue 4, pp.422-446

[8] Kando, N.: Overview of the Sixth NTCIR Workshop, *NTCIR-6 Proceedings*, pp.i-ix (2007)

[9] Kazai, G., and Lalmas, M.: eXtended Cumulated Gain Measures for the Evaluation of Content-oriented XML Retrieval, *ACM Transactions on Information Systems*, Vol. 24, Issue 4, pp.503-542 (2006)

[10] 岸田, 岩山, 江口: 検索実験の方法と実際: NTCIR ワークショップでの試み, *Pre-Meeting Lecture at the NTCIR-3 Workshop* (2002)

[11] Property of Average Precision and its Generalization: An Examination of Evaluation Indicator for Information Retrieval Experiments, National Institute of Informatics Technical Report NII-2005-014E (2005)

[12] Moffat, A., Webber, W. and Zobel, J.: Strategic System Comparisons via Targeted Relevance Judgments, *ACM SIGIR 2007 Proceedings*, pp.375-382 (2007)

[13] NTCIR: `http://research.nii.ac.jp/ntcir/index-ja.html`

[14] 酒井, 小川, 木谷, 石川, 木本, 中渡瀬, 芥子, 豊浦, 福島, 松井, 上田, 徳永, 鶴岡, 安形, 神門: 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, サイエンス社 (1998)

[15] 酒井: 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. 情報処理, Vol.47, No.2, pp.147-158 (2006)

[16] 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, Vol.47, No.SIG 4 (TOD29), pp.13-27 (2006) Also available in IPSJ Digital Courier, Vol.2, pp.174-188 (2006) `http://www.jstage.jst.go.jp/article/ipsjdc/2/0/174/_pdf`

[17] Sakai, T.: Evaluating Evaluation Metrics based on the Bootstrap, *ACM SIGIR Proceedings*, pp.695-696 (2006)

[18] Sakai, T.: On the Reliability of Information Retrieval Metrics based on Graded Relevance, *Information Processing and Management*, Volume 43, Issue 2, pp.531-548 (2007)

[19] Sakai, T.: On the Reliability of Factoid Question Answering Evaluation, *ACM Transactions on Asian Language Information Processing (TALIP)*, Volume 6, Issue. 1, Article 3 (2007)

[20] Sakai, T.: On Penalising Late Arrival of Relevant Documents in Information Retrieval Evaluation with Graded Relevance, Proceedings of the First Workshop on Evaluating Information Access (EVIA 2007), pp.32-43, 2007. `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/EVIA/1.pdf`

[21] Sakai, T.: Alternatives to Bpref, *ACM SIGIR 2007 Proceedings*, pp.71-78 (2007)

[22] Sakai, T.: Evaluating Information Retrieval Metrics based on Bootstrap Hypothesis Tests, 情報処理学会論文誌データベース, Vol.48, No.SIG 9 (TOD35), pp.11-28, 2007. Also available in IPSJ Digital Courier, Vol.3, pp.625-642 (2007) `http://www.jstage.jst.go.jp/article/ipsjdc/3/0/625/_pdf`

[23] Sakai, T.: On the Properties of Evaluation Metrics for Finding One Highly Relevant Document, 情報処理学会論文誌データベース (IPSJ Transactions on Databases), Vol.48, No.SIG 9 (TOD35), pp.29-46, 2007. Also available in IPSJ Digital Courier, Vol.3, pp.643-660 (2007) `http://www.jstage.jst.go.jp/article/ipsjdc/3/0/643/_pdf`

[24] Sakai, T. and Kando, N.: A Further Note on Alternatives to Bpref, 情報処理学会研究報告 2007-FI-88, pp.7-14 (2007)

[25] Sparck Jones, K., Walker, S. and Robertson, S. E.: A Probabilistic Model of Information Retrieval: Development and Comparative Experiments, Information Processing and Management 36, Part I (pp. 779-808) and Part II (pp. 809-840) (2000)

[26] Turpin, A. and Scholer, F.: User Performance versus Precision Measures for Simple Search Tasks, *ACM SIGIR 2006 Proceedings*, pp. 11-18 (2006)

[27] Voorhees, E. M. and Harman, D.: Overview of the Eighth Text REtrieval Conference (TREC-8), *Proceedings of TREC-8* (2000)

[28] Voorhees, E. M.: Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness, *ACM SIGIR '98 Proceedings*, pp. 315-323 (1998)

[29] Voorhees, E. M.: The Philosophy of Information Retrieval Evaluation, *Proceedings of CLEF 2001*, Lecture Notes in Computer Science 2406, pp.355-370 (2002)

[30] Voorhees, E. M. and Harman, D. K. (eds.): *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press (2005)

[31] Voorhees, E. M.: Overview of the TREC 2005 Robust Retrieval Track, *TREC 2005 Proceedings* (2006)

[32] Witten, I. H., Moffat, A. and Bell, T. C.: Managing Gigabytes: Compressing and Indexing Documents and Images (Second Edition), Morgain Kaufmann Publishers, pp. 189-190 (1999)

[33] Yilmaz, E. and Aslam, J. A.: Estimating Average Precision with Incomplete and Imperfect Judgments, *ACM CIKM 2006 Proceedings*, pp.102-111 (2006)