CD-002

ユーザの感覚に近い多様化検索評価指標

Diversified Search Evaluation Measures That Align with User Perception

酒井 哲也[1,a)]　Zhaohao Zeng[1,b)]

1. はじめに

我々の日常生活全般を支えているウェブ検索エンジンは日々改良されている．最近では一部の検索クエリに対して天気・株価・事典など様々なタイプの情報も提示されるようになっているが [31]，検索結果 (Search Engine Result Page: SERP) の基本要素はウェブページのランクつきリストであり，その質の向上は依然として重要課題である．ウェブ検索エンジンを保有する企業は，クエリログ中の頻出クエリを対象に人手による適合性判定データを収集し，[*1] nDCG (normalised Discounted Cumulative Gain) [30] などの評価指標に基づいて SERP の質を定量評価し，改良や調整を行っている．さらに最近では，検索エンジンに入力される短いクエリの背後にあるユーザの様々な意図を考慮し，SERP に含まれる情報を多様化する検索結果多様化 (search result diversification) [26], [30] の研究が進んでおり，この目的に特化した検索評価指標がいくつか提案されている [1], [3], [5], [8], [9], [18], [22]．これらの新しい評価指標は，適合性 (relevance) のみを考慮する nDCG のような従来型指標とは異なり，適合性と多様性 (diversity) のバランスを考慮する．検索エンジンの改良や調整の主目的はユーザの満足度を高めることであるから，これらの指標が下す SERP の優劣に関する判断は，ユーザの実際の判断と合致すべきである．本研究ではこの観点から，検索結果多様化のための評価指標を検証する．

多様化検索評価指標は，各検索クエリに付随する複数の検索意図が既知であるものとして，各検索意図の確率，および検索意図毎の適合性判定結果をもとに SERP を評価する．対象となるクエリには，例えば “apple” (果物か企業か) のように多義性に起因する曖昧性を含むものや，“harry potter” (本か映画か登場人物か) のように検索意図の詳細化が不十分なものがある [9], [30]．このような前提に基づく評価のアプローチは 2 つに大別される．第一は，Intent-Aware な評価指標 [1], [5], [18], [30] と呼ばれるものであり，一連の数学的公理に基づき最近提案された RBU (Rank-Biased Utility) [3] もこのグループに属する．第二は，D♯-measures [18], [22], [30] と呼ばれるものである．例えば ERR (Expected Reciprocal Rank) の Intent-Aware 版である ERR-IA [5] は TREC (Text Retrieval Conference) の多様化検索タスクで用いられ [7]，一方 nDCG の D♯-measure 版である D♯-nDCG は NTCIR (NII Testbeds and Community for Information access Research) の同様なタスクで用いられた [23]．しかし，一般に多様化検索評価指標は従来型評価指標よりも複雑であり，どの指標がどの程度実際のユーザの感覚と合致するか不明であった．本研究は，この問題に対する明確な答えを提供する．具体的には，与えられた SERP 対に対して各評価指標が下した優劣の判断が，15 名の判定者が同じ対に下した判断 (適合性・多様性それぞれの観点から下した判断) とどれほど一致するかを定量化し，統計的多重比較法 [19], [32] により D♯-measures のアプローチの有効性を示す．

2. 関連研究

ここでは，検索評価指標に対する評価に関連する既存研究について議論する．

2.1 ユーザに依存しない検索評価指標の評価

検索評価指標に対する評価の多くは，実際のユーザの判断に依存しない方法で行われてきた．最も広く行われているのは，各評価指標の平均値によりシステム群をそれぞれランクづけし，ランキング対の類似度を順位相関係数で定量化することである [18], [30]．しかし，この方法は指標の対がどの程度似ているか，似ていないかを示すことが出来るだけで，どの指標が正しいかを測ることが出来ない．また，この他に比較的広く用いられてきた手法として判別能力 (discriminative power) の測定がある [16], [18], [30]．これは，全システム対について統計的検定を行った場合に得られる一連の p-value を比較することにより，評価指標の安定性を定量化するものである．しかし，統計的安定性は評価指標が満たすべき必要条件であはあるが十分条件ではないため，この方法によってもどの指標が正しいかがわかるわけではない．この点は，判別能力に似た評価指標の比

[1] 早稲田大学 (Waseda University)
[a)] tetsuyasakai@acm.org
[b)] zhaohao@fuji.waseda.jp
[*1] クリックデータ等をもとに各ウェブページの適合性を推定することもある程度可能である [1].

較方法である逆転法 (swap method) [18], [28], [30] などについても同様である．

多様化検索評価指標は，ある検索クエリ q が与えられた下での各検索意図 i の確率 $Pr(i|q)$ と，検索意図 i 毎の文書 d の適合レベル $x_i(d)$ を考慮して適合性と多様性のバランスを評価する必要があるため，伝統的な検索評価指標に比べ複雑である．一致度テスト (intuitiveness test, concordance test) [17], [18], [30] は，このような複雑な指標群がより直観的にわかりやすい単純な指標とどれくらい一致するかを定量化する方法であり，多様化検索評価指標の他にも統合検索評価指標 [29] やクリックモデル [6] などの評価に用いられている．しかし，この方法はあくまで単純な評価指標を正解と見なすものであり，この正解自体がユーザの判断と一致する保証はない．

評価指標が満たすべき性質をいくつかの公理 (axioms) により表現し，これらを満たすように評価指標を設計するアプローチがある [3], [13]．逆に，既存の評価指標がいくつの公理を満たすかは，評価指標の比較評価方法となりうる．評価指標の数学的性質を明らかにすることには一定の意義があるが，予め用意した公理群で十分であるか，また各公理が実ユーザにとって本当に重要であるかには議論の余地が残る．例えば，Amigó ら [3] が多様化検索指標の設計のために設定した公理には，「各検索意図に対する適合性は多値でななく二値 (すなわち適合か非適合か) である」「ひとつの文書が複数の検索意図を同時に満たすことはない」などのあまり現実的ではない大前提に依存するものがある．実際，本研究は，彼らが ACM SIGIR 2018 にて公理的アプローチにより提案した多様化検索評価指標 RBU [3] が，ユーザによる SERP の優劣判断との一致度という観点から必ずしも最適ではないことを示している．

以上の議論より，検索評価指標が本当にユーザの体感を近似するものになっているかどうか検証するためには，人間の判断を正解と見なした評価を行うことが不可避であると考えられる．

2.2 ユーザに依存する検索評価指標の評価

検索評価指標とユーザの関係を扱う研究のひとつの流れとして，特定の評価指標の値の高低をコントロールし，ユーザによる情報検索の効率やユーザの満足度への影響を分析するものがある．例えば Turpin ら [27] は，二値適合性に基づく評価指標である平均精度 (Average Precision) [18], [30] の高低とユーザによる検索効率の相関は低いと報告している．一方，Al-Maskari ら [2] は，評価指標値の高低がユーザによる検索効率およびユーザ満足度にある程度影響を与えると報告している．

前述の通り，検索評価指標は日頃よりユーザ満足度を近似するものとして検索エンジンの調整や改良に用いられている．このことから，本研究の主眼は，ある評価指標が 2 つの SERP に対して下した優劣の判断がユーザの判断と一致することを保証することである．このアプローチに最も近い先行研究として，Sanderson らの研究がある [25]．しかし，2010 年当時の彼らの研究は，現在の我々の観点からすると以下の点で不十分である．第一に，彼らは 4 つの多様化検索評価指標しか扱っておらず，最新の指標を検討していない．第二に，彼らが実験に用いた SERP 対はわずか 79 件であり，実験結果から明確な結論が得られていない．第三に，彼らの実験では，判定者に検索クエリ自体ではなくそれに不随する検索意図のひとつのみを提示しており，いずれの SERP がより多様性が高いかという観点からの人手による判断を収集できていない [24]．これに対し本研究では，最近提案されたものを含む 21 種類の多様化検索評価指標を対象としており，1,127 件もの SERP 対を用意し，各 SERP 対に対する適合性・多様性それぞれの観点に基づく判定ラベルを 15 名の判定者から収集している．

我々が ACM SIGIR 2019 において行った中間報告 [24] では，上記 15 名の判定者による判定結果の多数決により各 SERP 対に対する正解をひとつ定めた上で，各評価指標の判断が正解と一致するか否かを議論した．しかし，この方法では，例えば 15 名全員が「第一の SERP のほうが良い」と判定したケースと，15 名中 8 名が同様に判定したケースを区別できない．さらに，評価値の上限 (どこまでユーザの判断に近づくことが可能か) の議論が欠けていた．そこで本研究では，SERP 対に対する評価指標の判断が何%の判定者の判断と一致するかを表す一致率により評価を行う．さらに，個々の判定者もまた評価指標の一種と見なすことにより，各評価指標の判断と判定者の判断との差異を統計的多重比較法 [19], [32] により明らかにする．

3. 検索評価指標

TREC では主として Intent-Aware な評価指標が [7], NTCIR では D♯-measures [23] が多様化検索の評価に用いられてきた．前述の公理的アプローチに基づく RBU [3] も前者に属する．以下，3.1 節で多様化検索評価指標の基礎をなす nDCG などの従来型の評価指標について概説する．次に，3.2 節で Intent-Aware な評価指標および D♯-measures の概要を述べる．以下，例えば nDCG の Intent-Aware 版は nDCG-IA，D♯ 版は D♯-nDCG のように表記する．なお，本研究では 4 章で説明するデータを公開しているため (7 章参照)，本研究でカバーしていない従来型評価指標および多様化検索評価指標について検討したい研究者は，容易に独自の評価を行うことができる．

3.1 適合性のみを考慮した従来型評価指標

適合性レベル x の各文書に利得 (gain) $gv(x)$ を与えるものとする [30]．本研究では $x = 0, 1, 2, \ldots, 4$ をもつ適合性判定データを扱うため，$gv(x) = 2^x - 1 = 0, 1, 3, \ldots, 15$ と

する [5]．これに基づき評価対象の SERP および理想的な SERP の第 r 位の文書に与える利得をそれぞれ $g(r), g^*(r)$ とするとき，従来型評価指標の中で 21 世紀になり最も広く用いられてきた nDCG は以下のように定義できる [18], [30]．

$$nDCG = \frac{\sum_{r=1}^{l} g(r)/\log(r+1)}{\sum_{r=1}^{l} g^*(r)/\log(r+1)} \quad (1)$$

本研究では，ウェブ検索においては検索結果の 1 ページ目の質が最重要であることから $l = 10$ とする．

本研究で扱う nDCG 以外の従来型評価指標は，全て Normalised Cumulative Utility (NCU) [18], [21], [30] の一種として説明できる．

$$NCU = \sum_{r=1}^{l} P_S(r) NU(r) \quad (2)$$

前提として，与えられた SERP の最上位からスタートし，得られた情報に満足した時点で SERP の走査を停止するユーザの母集団を考える．$P_S(r)$ は，ユーザが SERP の第 r 位で満足して停止する確率を表し，$NU(r)$ は，このユーザ群にとっての SERP の有用性 (utility) を表す．

表 1　NCU ファミリーに属する評価指標のまとめ

評価指標	$P_S(r)$	$NU(r)$
Q	$I(r)/\min(l, R)$	$BR(r)$
RBP	$(1-p)p^{r-1}$	$g(r)/gv_{max}$
ERR	$P_{ERR}(r)$	$1/r$
EBR	$P_{ERR}(r)$	$BR(r)$
iRBP	$P_{ERR}(r)$	p^r

表 1 に，本研究で扱う NCU ファミリーに属する評価指標をまとめた．Q [15], [18] は，停止確率 $P_S(r)$ として，平均精度と同様に全適合文書上の一様分布を，有用性関数として以下のブレンド比 (Blended Ratio) を採用している．

$$BR(r) = \frac{\sum_{k=1}^{r} I(k) + \sum_{k=1}^{r} g(k)}{r + \sum_{k=1}^{r} g^*(k)} \quad (3)$$

$I(k)$ は第 k 位の文書が非適合文書の場合に 0，それ以外の場合に 1 となるフラグである．ブレンド比は，精度 ($Precision(r) = \sum_{k=1}^{r} I(k)/r$) と，nDCG に似た nCG (normalised Cumulative Gain) [11] を統合したものである．RBP (Rank-Biased Precision) [14] は，ユーザが SERP の第 r 位から第 $(r+1)$ 位に遷移する確率が常に一定であるという仮定の下に，この確率 p をパラメタとしてもつ．有用性関数には利得を最大利得 gv_{max} (本研究の場合 $gv_{max} = 2^4 - 1$) により正規化したものを用いている．ERR [5] の停止確率 $P_{ERR}(r)$ は，第 $(r-1)$ までの文書に対してユーザが満足しない確率と，第 r 位の文書に満足する確率に基づく直感的なユーザモデルに基づく．

$$P_{ERR}(r) = P_{sat}(r) \prod_{k=1}^{r-1} (1 - P_{sat}(k)) \quad (4)$$

ここで $P_{sat}(r) = g(r)/(gv_{max} + 1) = g(r)/2^4$ とする．また，ERR では第 r 位の文書のみがこのユーザ群にとって適合すると考えるため，有用性関数は $1/r$ となっている．ERR は，他の指標とは異なり，informational ではなく navigational な検索意図 [4], [17], [30] (すなわち多くの適合文書ではなく特定の適合文書を探したいという意図) に適している．EBR (Expected Blended Ratio) [24] は $P_{ERR}(r)$ と $BR(r)$ を組み合わせたものである．iRBU (intentwise RBU) [24] は多様化検索評価指標 RBU [3] の構成要素であり，停止確率に $P_{ERR}(r)$ を，有用性関数にはユーザが第 r 位までの文書を調べる労力の関数である p^r を用いている．ここで，確率 p は RBP と同様なパラメタである．本研究では，RBP [13] および RBU [3] の既存研究より，$p = 0.85, 0.99$ について検討する．iRBU は，有用性関数に各文書の適合性の情報を用いていない点に特色がある．

3.2　多様化検索評価指標

多様化検索テストコレクションでは，各検索クエリ q に対する検索意図の集合 $\{i\}$ を既知とし，各検索意図 i の確率 $Pr(i|q)$ と，各文書の検索意図 i 毎の適合レベル x_i が与えられている．ある検索意図 i に着目し x_i をもとに計算した従来型評価指標を M_i とするとき，対応する Intent-Aware な指標は以下のように計算できる [1], [5]．

$$M\text{-}IA = \sum_i Pr(i|q) M_i \quad (5)$$

また，公理的アプローチにより設計された RBU は，ユーザの労力の影響を制御するパラメタ e (Amigo ら [3] に倣い $e = 0.01$ とする) を用い，以下のように記述できる [24]．

$$RBU = iRBU\text{-}IA - e \sum_{r=1}^{l} p^r \quad (6)$$

すなわち，iRBU の Intent-Aware 版から労力分を減点したものである．なお，本研究の結果は e の値に依存しない．[*2]

D♯-measures は，まず SERP の第 r 位の文書の検索意図 i に関する適合レベル x_i に応じた利得を $gv(x_i)(= 2^{x_i} - 1)$ とするとき，この文書の総利得を $G(r) = \sum_i Pr(i|q) gv(x_i)$ により求める．また，全文書を総利得によりソートして理想的な SERP を定義し，これをもとに理想的な SERP 中の各文書の総利得 $G^*(r)$ を得る．これらを $g(r), g^*(r)$ と同様に扱って従来型評価指標 M を計算したものを D-measures と呼び，$D\text{-}M$ のように表記する．一方，純粋に SERP の多様性のみを測る評価指標として，全検索意図のうち SERP によりカバーされたものの割合を表す I-rec (intent recall) を考える．NTCIR INTENT タスク [23] では，D-nDCG を縦軸に，I-rec を横軸にとって多様化検索システムを評価した．D♯-measures は，両者を線形結合したものである [22]．

[*2] e および p を定めた下でのサイズ l の SERP に対する労力項は定数であり，RBU の大小に影響しないためである．

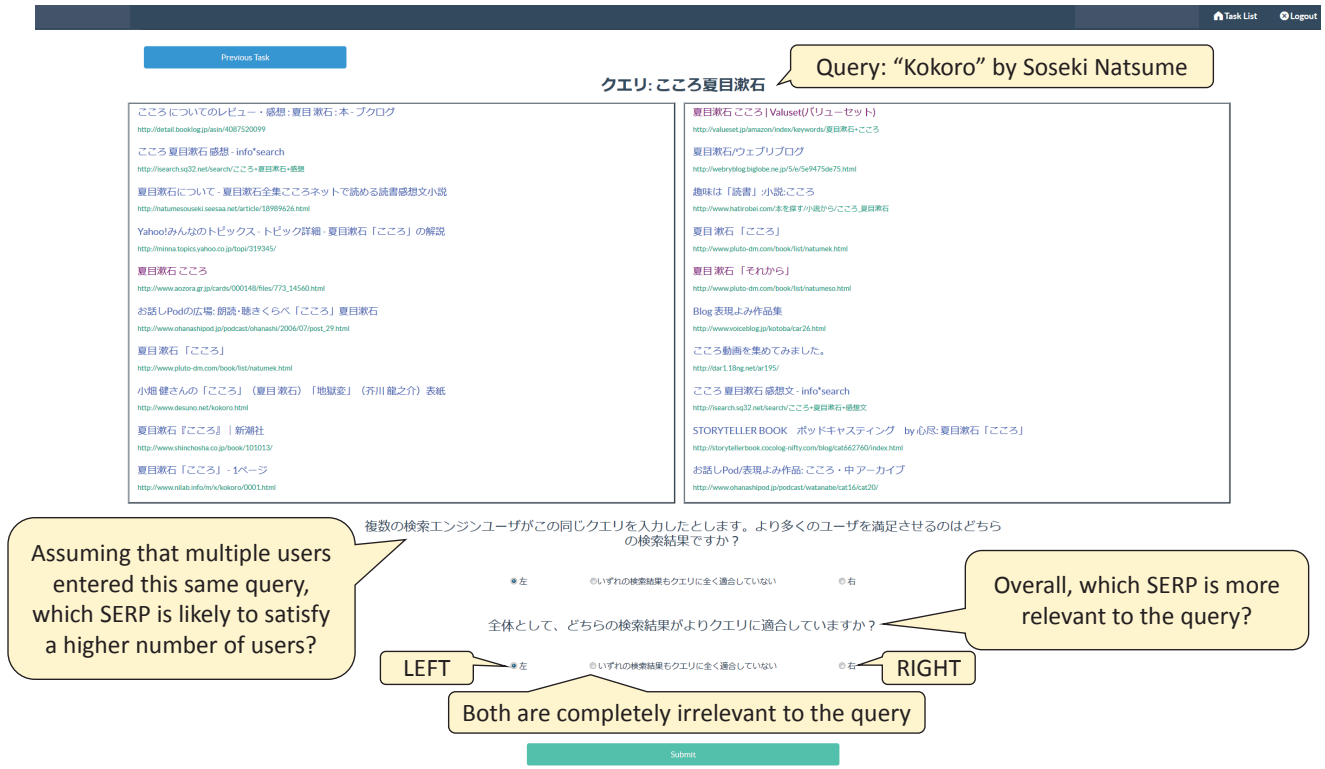

図 1　判定者が用いた SERP 選考判定インタフェース [24].

$$D\sharp\text{-}M = \gamma I\text{-}rec + (1-\gamma) D\text{-}M \tag{7}$$

λ はパラメタであり，本研究では INTENT タスクに倣い $\lambda = 0.5$ とする．

4. SERP 選好判定データ

本研究の目的は，2 つの SERP のいずれが優れているかという多様化検索評価指標の判断が何%のユーザの判断と一致するかを定量化し，ユーザの体感に近い評価指標はどれか，およびユーザの判断にどこまで近づけるのかを明らかにすることである．このため，ウェブ検索多様化タスクである NTCIR-9 INTENT の日本語サブタスク [23] に提出された実際の検索結果と，同タスクで構築されたテストコレクションを利用する．このテストコレクションは各検索意図の確率 $Pr(i|q)$ および検索意図毎の多値適合性判定結果を有しており，我々の研究目的に適している．[*3]

NTCIR-9 INTENT 日本語テストコレクションは，検索対象である約 6730 万件のウェブページからなる日本語コーパス clueweb09JA[*4]と，100 件の検索クエリを含む．各検索クエリには平均 10.91 件の検索意図および確率が付随している．また，各適合文書には検索意図毎に適合性レベル $x = 1, \ldots, 4$ が付与されている．このタスクには 15 のシステムが参加したため，$15 * 14/2 = 105$ 件のシステム対の比較が可能である．

上記のデータをもとに，我々は独自に以下のように SERP 選好判定 (preference assessment) データを作成した．まず，上記 INTENT データより，100 クエリ $\times 105$ SERP 対 $= 10,500$ 件の < クエリ, SERP1, SERP2> の 3 つ組を作成した．ここで，各 SERP は前述の通り上位 $l = 10$ 件のウェブページを含む．次に，ほぼ質が同じ SERP 対を除去するため，精度 (10 件中の適合文書数の割合) および intent recall (SERP でカバーできている検索意図の割合) の差がいずれも 0.1 以上である 3 つ組のみを抽出した．これにより 3 つ組は 1,247 件に絞り込まれた．さらに，clueweb09JA のウェブページデータに文字コードの問題があるものなどを除去した結果，最終的に 43 件のクエリをカバーする 1,127 件の 3 つ組が得られた [24].

上記の 3 つ組データに対して「ユーザの判断」を下すため，早稲田大学の情報理工学科，情報通信学科 (学部) および情報理工・情報通信専攻 (大学院) に属する日本語ネイティブの学生を 15 名雇用し，各自に全 3 つ組データに対して適合性・多様性の 2 つの観点から SERP 選好判定を行ってもらった．図 1 はこの作業のために判定者に提供したウェブブラウザ上のインタフェースである．この例のように，判定者は画面上部のクエリ (個々の検索意図ではない) を見た上で，左右いずれの SERP が好ましいか選択する．判定者に対する質問は以下の 2 つである．

(適合性)　全体として，どちらの検索結果がよりクエリに適合していますか?

(多様性)　複数の検索エンジンユーザがこの同じクエリを

[*3] TREC の多様化検索タスクで構築されたテストコレクション [7] は $Pr(i|q)$ の情報をもっていない．

[*4] http://www.lemurproject.org/clueweb09.php/

入力したとします．より多くのユーザを満足させるのはどちらの検索結果ですか?

なお，上記質問の順序，左右の SERP の位置，クエリの順序は各判定者に対しランダムに提示した．以上により，適合性選好ラベルを含む $1,127 \times 15$ の行列と，多様性選好ラベルを含む $1,127 \times 15$ の行列が構築できた．各選好ラベルは LEFT・RIGHT・EQUAL のいずれかである．

上記選好判定データの信頼性を Krippendorff の α [12], [20] により定量評価したところ，適合性選好について 0.406，多様性選好について 0.356 と良好な結果が得られた．さらに，15 人の判定者のうち 1 人のデータを除去して再度 α を計算することにより各判定者の質のチェックを行ったが，いずれの判定者についても問題は見られなかった [24]．また，適合性選好と多様性選好の関係を調べるため，それぞれの行列について，各行の 15 個のラベルから「LEFT の個数と RIGHT の個数の差」を計算し，その Pearson 相関係数 (サンプルサイズ $n = 1,127$) を計算した結果は 0.898 (95%CI[0.886, 0.909]) となった．すなわち，適合性の観点から多くの判定者に好まれる SERP はまた，多様性の観点からも多くの判定者に好まれる傾向が顕著であった．

5. 評価指標の評価方法

与えられた < クエリ, SERP1, SERP2> の 3 つ組に対し，各評価指標は LEFT・RIGHT・EQUAL のいずれかの判断を下す．これが何%の判定者の選好ラベルと一致しているかを一致率 (Agreement Rate) と定義する．各判定者は適合性・多様性に基づき個別に判断を下すので，一致率も 2 種類定義できる．また，判定者は 15 名いるので，これが一致率の分母となる．一致率を 1,127 件の 3 つ組について平均したものを MAR (Mean Agreement Rate) と呼ぶ．さらに，多様化検索評価指標は適合性と多様性の両方を考慮することが要求されているため，本研究では，下記のように両方を同時に考慮した一致率を主要な評価尺度とする．まず，前述の 2 つの選好ラベル行列を比較し，各判定者が適合性と多様性について同じ判断を下した場合のみを残した第 3 のラベル群を作成する．例えば，ある 3 つ組について，ある判定者の適合性選好ラベルが LEFT，多様性選好ラベルが RIGHT もしくは EQUAL である場合にはこの判定者のラベルは除外される．この結果，各 3 つ組に付随するラベル数は平均 13.10 (最小 5，最大 15) となった．従って，この場合の一致率の分母は 3 つ組によって変化する．

本研究では，各評価指標の判断がユーザの判断と比べてどれほど信頼できるかを定量化するため，個々の判定者の選好ラベル群もまた評価指標の一種であると見なし，これらについても一致率を計算する．(分母には判定者自身の選好ラベルが含まれる．) この際，例えば判定者 01 の適合性選好ラベル群を 01rel，多様性選好ラベル群を 01div のように表記する．

表 2 各判定者による適合性と多様性の選考判定が一致した場合に着目した各評価指標の MAR ($n = 1,127$).

(a) 評価指標	MAR	(b) 選好ラベル群	MAR
D♯-nDCG	0.7484	12rel (best)	0.7724
D♯-RBP ($p = 0.85$)	0.7439	12div	0.7708
D♯-Q	0.7434	02rel	0.7620
D♯-EBR	0.7374	10div	0.7590
D♯-RBP ($p = 0.99$)	0.7330	04div	0.7545
I-rec	0.7330	15div	0.7475
RBU ($p = 0.99$)	0.7308	03rel	0.7411
D♯-ERR	0.7250	02div	0.7406
RBU ($p = 0.85$)	0.7228	04rel	0.7383
D-Q	0.7199	15rel	0.7362
D-RBP ($p = 0.99$)	0.7159	10rel	0.7362
nDCG-IA	0.7159	07rel	0.7329
Q-IA	0.7119	09div	0.7154
D-RBP ($p = 0.85$)	0.7102	07div	0.7121
D-nDCG	0.7101	11div (median)	0.7105
D-EBR	0.7100	13rel	0.7089
RBP-IA ($p = 0.99$)	0.7097	09rel	0.7087
RBP-IA ($p = 0.85$)	0.7095	13div	0.7026
EBR-IA	0.7052	01rel	0.7026
ERR-IA	0.6973	05div	0.6995
D-ERR	0.6894	05rel	0.6933
		01div	0.6899
		03div	0.6880
		11rel	0.6870
		14rel	0.6866
		14div	0.6748
		06rel	0.6437
		08rel	0.6380
		06div	0.6352
		08div (worst)	0.6198

以上のように計算した各評価指標および各判定者の MAR の差を厳密に議論するため，統計的多重比較法の一種である Tukey HSD 検定を行う [19], [32]．また，本研究では，p-value にとどまらず，効果量 (effect size) [10], [19] を議論するに十分な実験データを提供する．

6. 結果と考察

表 2 に，**各判定者の適合性・多様性選好ラベルが一致した場合に着目した MAR の結果**を示す．(a) は 21 種類の多様化検索評価指標の結果，(b) は各判定者による適合性および多様性選好ラベル群の結果である．全評価指標と，判定者による選好ラベル群のうち MAR が最高なもの (「best 判定者」)，最低なもの (「worst 判定者」)，および中間値 (「median 判定者」) [*5]からなる 24 件について対応のある Tukey HSD 検定 (有意水準 5%) を行った結果を表 3 に示す．これらから以下のことがわかる．

(A) 表 3(I) より，best 判定者は D♯-nDCG を除く全ての指標，および median・worst 判定者を統計的に有意に上回っている．

(B) 表 3(II) より，表 2 の上位 4 指標 (D♯-nDCG〜D♯-EBR)

[*5] 評価指標に対して辛い評価を行うため，30 件中 15 位の選好ラベル群を中間値と見なす．

表 3 表 2 に対応する Tukey HSD 検定の結果の抜粋．対象としたのは多様化検索評価指標 21 件および best・median・worst 判定者である．有意水準 $\alpha = 0.05$ にて統計的に有意な差のみを示している．(紙面の制約上，(IV) および (VI)-(IX) の詳細は省略している．) 二元配置分散分析 (繰り返しなし) の誤差分散は $V_{E2} = .0225$ であり，各効果量は $\Delta MAR/\sqrt{V_{E2}}$ により算出できる [19].

Comparison	ΔMAR	p-value
(I) best assessor > X (X: median, worst, all measures except D♯-nDCG)		
best assessor > worst assessor	.1526	.0000
best assessor > D-ERR	.0829	.0000
best assessor > ERR-IA	.0750	.0000
best assessor > EBR-IA	.0672	.0000
best assessor > RBP-IA ($p = 0.85$)	.0629	.0000
best assessor > RBP-IA ($p = 0.99$)	.0627	.0000
best assessor > D-EBR	.0624	.0000
best assessor > D-nDCG	.0623	.0000
best assessor > D-RBP ($p = 0.85$)	.0622	.0000
best assessor > median assessor	.0619	.0000
best assessor > Q-IA	.0605	.0000
best assessor > nDCG-IA	.0565	.0000
best assessor > D-RBP ($p = 0.99$)	.0565	.0000
best assessor > D-Q	.0525	.0000
best assessor > RBU ($p = 0.85$)	.0496	.0000
best assessor > D♯-ERR	.0474	.0000
best assessor > RBU ($p = 0.99$)	.0412	.0000
best assessor > I-rec	.0394	.0000
best assessor > D♯-RBP ($p = 0.99$)	.0394	.0000
best assessor > D♯-EBR	.0349	.0003
best assessor > D♯-Q	.0290	.0124
best assessor > D♯-RBP ($p = 0.85$)	.0285	.0161
(II) X > median assessor (X: top 4 measures)		
D♯-nDCG > median assessor	.0379	.0000
D♯-RBP ($p = 0.85$) > median assessor	.0334	.0009
D♯-Q > median assessor	.0329	.0012
D♯-EBR > median assessor	.0270	.0348
(III) median assessor > X		
median assessor > worst assessor	.0907	.0000
(IV) (X: all 21 measures)		
X > worst assessor		
(V) D♯-nDCG > X (X: worst 12 measures)		
D♯-nDCG > D-ERR	.0590	.0000
D♯-nDCG > ERR-IA	.0510	.0000
D♯-nDCG > EBR-IA	.0432	.0000
D♯-nDCG > RBP-IA ($p = 0.85$)	.0389	.0000
D♯-nDCG > RBP-IA ($p = 0.99$)	.0387	.0000
D♯-nDCG > D-EBR	.0384	.0000
D♯-nDCG > D-nDCG	.0383	.0000
D♯-nDCG > D-RBP ($p = 0.85$)	.0382	.0000
D♯-nDCG > Q-IA	.0365	.0001
D♯-nDCG > nDCG-IA	.0325	.0015
D♯-nDCG > D-RBP ($p = 0.99$)	.0325	.0015
D♯-nDCG > D-Q	.0285	.0162
(VI) (X: worst 11 measures)		
D♯-RBP ($p = 0.85$) > X		
D♯-Q > X		
(VII) (X: worst 8 measures)		
D♯-EBR > X		
(VIII) (X: worst 3 measures)		
D♯-RBP ($p = 0.99$) > X		
I-rec > X		
(IX) (X: worst 2 measures)		
RBU ($p = 0.99$) > X		
D♯-ERR > X		
(X) other		
RBU ($p = 0.85$) > D-ERR	.0333	.0009
D-Q > D-ERR	.0305	.0053
D-RBP ($p = 0.99$) > D-ERR	.0265	.0434
nDCG-IA > D-ERR	.0264	.0446

は，median 判定者を統計的に有意に上回っている．

(C) 表 3(III)-(IV) より，median 判定者および全 21 指標は，worst 判定者を統計的に有意に上回っている．

(D) 表 3(V) より，表 2 においてトップの D♯-nDCG は，D-Q 以下 12 指標を統計的に有意に上回っている．同様に，表 3(VI)-(IX) に示した各指標は，それぞれ下位 $11, 8, 3, 2$ 件の指標を統計的に有意に上回っている．

(E) 表 3(V)-(X) より，D-ERR は，上位 12 指標 (D♯-nDCG〜nDCG-IA) を統計的に有意に**下回っている**．

なお，表 3 より，例えば D♯-nDCG と D-nDCG の差の効果量は $\Delta MAR/\sqrt{V_{E2}} = .0383/\sqrt{.0225} = 0.2553$ (標準偏差の 1/4 程度) のように算出できる [19].

一方，表 4 および表 5 に，判定者の適合性選好判定データのみ，および多様性選好判定データのみを用いて計算した MAR の結果を示す．(紙面の制約上，詳細な統計的検定結果の表は割愛する．) まず，**適合性選好判定データのみを用いた MAR の結果**は以下の通りである．

- best 判定者は，全ての指標および median・worst 判定者を統計的に有意に上回っている．
- median 判定者は，worst 判定者および表 4 の Q-IA 以下 6 指標を統計的に有意に上回っている．
- 全 21 指標とも，worst 判定者を統計的に有意に上回っている．
- 表 4 においてトップの D-Q は，D-RBP ($p = 0.99$) 以下 16 指標を統計的に有意に上回っている．同様に，D♯-nDCG，D♯-Q，D♯-RBP ($p = 0.85$)，D♯-EBR，D♯-RBP ($p = 0.99$) は，それぞれ下位 $6, 6, 3, 3, 2$ 件の指標を統計的に有意に上回っている．
- D-ERR は，上位 14 指標 (D-Q〜D-RBP ($p = 0.85$)) を統計的に有意に**下回っている**．

次に，**多様性選好判定データのみを用いた MAR の結果**は以下の通りである．

- best 判定者は，D♯-nDCG を除く全ての指標，および median・worst 判定者を統計的に有意に上回っている．

表 4 判定者による適合性の選考判定結果のみを考慮した各評価指標の MAR ($n = 1,127$).

(a) Measure	MAR	(b) Assessor	MAR
D-Q	0.7345	03rel (best)	0.7600
D♯-nDCG	0.7202	10rel	0.7565
D♯-Q	0.7201	02rel	0.7433
D♯-RBP ($p = 0.85$)	0.7146	12rel	0.7423
D♯-EBR	0.7117	04rel	0.7421
D-RBP ($p = 0.99$)	0.7055	07rel	0.7315
RBU ($p = 0.99$)	0.7023	01rel (median)	0.7196
D♯-RBP ($p = 0.99$)	0.7015	15rel	0.7393
I-rec	0.7015	11rel	0.7117
RBU ($p = 0.85$)	0.7007	05rel	0.7073
RBP-IA ($p = 0.99$)	0.6990	13rel	0.7054
D-nDCG	0.6984	09rel	0.6746
D♯-ERR	0.6981	14rel	0.6661
D-RBP ($p = 0.85$)	0.6974	08rel	0.6554
RBP-IA ($p = 0.85$)	0.6968	06rel (worst)	0.6384
Q-IA	0.6941		
D-EBR	0.6932		
nDCG-IA	0.6929		
EBR-IA	0.6849		
ERR-IA	0.6788		
D-ERR	0.6723		

表 5 判定者による多様性の選考判定結果のみを考慮した各評価指標の MAR ($n = 1,127$).

(a) Measure	MAR	(b) Assessor	MAR
D♯-nDCG	0.7347	12div (best)	0.7542
D♯-RBP ($p = 0.85$)	0.7331	10div	0.7484
D♯-RBP ($p = 0.99$)	0.7276	04div	0.7444
I-rec	0.7276	15div	0.7429
D♯-Q	0.7273	02div	0.7208
D♯-EBR	0.7262	07div	0.7171
RBU ($p = 0.99$)	0.7208	03div	0.7054
D♯-ERR	0.7159	09div (median)	0.7042
RBU ($p = 0.85$)	0.7093	01div	0.6990
nDCG-IA	0.7015	11div	0.6946
D-RBP ($p = 0.99$)	0.6971	13div	0.6832
EBR	0.6941	14div	0.6771
Q-IA	0.6938	05div	0.6668
EBR-IA	0.6921	06div	0.6363
D-RBP ($p = 0.85$)	0.6920	08div (worst)	0.5993
D-nDCG	0.6914		
RBP-IA ($p = 0.85$)	0.6913		
RBP-IA ($p = 0.99$)	0.6903		
ERR-IA	0.6831		
D-Q	0.6811		
D-ERR	0.6753		

- 表 5 のトップ 2 である D♯-nDCG，D♯-RBP ($p = 0.85$) は，median 判定者を統計的に有意に上回っている．
- median 判定者および全 21 指標は，worst 判定者を統計的に有意に上回っている．
- 表 5 においてトップの D♯-nDCG は，RBU ($p = 0.85$) 以下 13 指標を統計的に有意に上回っている．同様に，D♯-RBP ($p = 0.85, 0.99$)，I-rec，D♯-Q，D♯-EBR，RBU ($p = 0.99$)，D♯-ERR，RBU ($p = 0.85$) は，それぞれ下位 $12, 12, 12, 12, 12, 10, 6, 2$ 件の指標を統計的に有意に上回っている．
- D-ERR は，上位 10 指標 (D♯-nDCG〜nDCG-IA) を統計的に有意に下回っている．

以上を総合すると，多様化検索評価指標について，人間による SERP 選好判定との一致率という観点から以下の結論が得られる．

- **最もユーザの感覚に近いのは D♯-nDCG である．**適合性・多様性双方を考慮した MAR および多様性のみを考慮した MAR において，best 判定者と統計的に有意差がないのは D♯-nDCG のみである．また，両 MAR において，D♯-nDCG は median 判定者を統計的に有意に上回っている．すなわち，**D♯-nDCG の判断の妥当性は人間と少なくとも同程度**である．
- 同様に，D♯-RBP ($p = 0.85$)，D♯-Q，D♯-EBR も人間と比べてほぼ遜色がない．適合性・多様性双方を考慮した MAR において median 判定者を統計的に有意に上回っており，特に D♯-RBP ($p = 0.85$) については多様性のみを考慮した MAR においても同様である．
- 総じて，**Intent-Aware な指標よりも D♯-measures のほうがユーザの感覚に近い．**例えば，D♯-nDCG および D♯-RBP ($p = 0.85$) は，適合性・多様性双方を考慮した MAR および多様性のみを考慮した MAR において，全ての Intent-Aware な指標を統計的に有意に上回っている．
- 式 7 により D-measures を intent recall と結合して D♯-measures を構成することは，多様性を考慮した場合のユーザの体感に近づけるのに有効である．例えば，表 2・表 5 いずれにおいても D♯-nDCG は D-nDCG および I-rec を平均的に上回っており，特に D-nDCG との差は統計的に有意である．

なお，ERR に基づく多様化検索評価指標の MAR が総じて低いのは，前述のとおり ERR が navigational な検索意図に適した評価指標であるためである [4], [17], [30]．すなわち，ERR が SERP 中の適合情報の「量」を考慮しないためである．また，D-Q は適合性のみを考慮した場合にトップであり (表 4)，多様性のみを考慮した場合に D-ERR に次いで成績が悪いことから (表 5)，適合性のほうに大きく振れた指標であることがわかる．

2 章で述べたように，RBU は公理的アプローチにより設計された指標である．しかし，我々の実験結果は，RBU よりも D♯-measures のほうが平均的に優れていることを示している．(前述の通り，表 5 における D♯-nDCG と RBU ($p = 0.85$) の差は統計的に有意である．) このことから，評価指標が一連の公理を満たすことは，ユーザの体感と合致させることと必ずしも一致しないことがわかる．

7. 結論

本研究では，NTCIR-9 INTENT タスクに提出されたラ

ンから抽出した 1,127 件の SERP 対に対し 15 人の判定者が独立に行った SERP 選好判定の結果を利用して，21 種類の多様化検索評価指標を一致率により定量評価した．Tukey HSD 検定に基づく我々の実験結果より，多様化検索の評価には Intent-Aware な評価指標ではなく D♯-measures，特に D♯-nDCG などを用いるべきであることがわかる．さらに，これらの指標は人間による SERP 選好判定と比しても遜色がないことが確認できた．我々の SERP 選好判定データ・各評価指標の値・一致率のデータは一般公開しているため，[*6] 本研究は他の研究者により再現・拡張可能である．

研究者や開発者が本来測定したいことを評価指標が本当に測定できているか否かは，多様化検索タスクや情報検索分野に限定されない極めて重要な問題である．冒頭で論じたように，システムの調整や改良の方向性は評価指標に大きく依存しているからである．我々は今後，人間による選好判定をもとに評価指標を検証するアプローチを多様化検索以外のタスクや研究分野に拡張する予定である．

参考文献

[1] Agrawal, R., Sreenivas, G., Halverson, A. and Leong, S.: Diversifying Search Results, *Proceedings of ACM WSDM 2009*, pp. 5–14 (2009).

[2] Al-Maskari, A., Sanderson, M., Clough, P. and Airio, E.: The Good and the Bad System: Does the Test Collection Predict Users' Effectiveness, *Proceedings of ACM SIGIR 2018*, pp. 59–66 (2008).

[3] Amigó, E., Spina, D. and de Albornoz, J. C.: An Axiomatic Analysis of Diversity Evaluation Metrics: Introducting the Rank-Biased Utility Metric, *Proceedings of ACM SIGIR 2018*, pp. 625–634 (2018).

[4] Broder, A.: A Taxonomy of Web Search, *SIGIR Forum*, Vol. 36, No. 2, pp. 3–10 (2002).

[5] Chapelle, O., Ji, S., Liao, C., Velipasaoglu, E., Lai, L. and Wu, S.-L.: Intent-based Diversification of Web Search Results: Metrics and Algorithms, *Information Retrieval*, Vol. 14, No. 6, pp. 572–592 (2011).

[6] Chuklin, A., Zhou, K., Schuth, A., Sietsma, F. and de Rijke, M.: Evaluating Intuitiveness of Vertica-Aware Click Models, *Proceedings of ACM SIGIR 2013*, pp. 1075–1078 (2013).

[7] Clarke, C. L., Craswell, N. and Voorhees, E. M.: Overview of the TREC 2012 Web Track, *Proceedings of TREC 2012* (2013).

[8] Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S. and MacKinnon, I.: Novelty and Diversity in Information Retrieval Evaluation, *Proceedings of ACM SIGIR 2008*, pp. 659–666 (2009).

[9] Clarke, C. L., Kolla, M. and Vechtomova, O.: An Effectiveness Measure for Ambiguous and Underspecified Queries, *Proceedings of ICTIR 2009 (LNCS 5766)*, pp. 188–199 (2009).

[10] Cohen, J.: *Statistical Power Analysis for the Bahavioral Sciences (Second Edition)*, Psychology Press (1988).

[11] 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).

[12] Krippendorff, K.: *Content Analysis: An Introduction to Its Methodology (Fourth Edition)*, SAGE Publications (2018).

[13] Moffat, A., Bailey, P., Scholer, F. and Thomas, P.: Incorporating User Expectations and Behavior into the Measurement of Search Effectiveness, *ACM TOIS*, Vol. 35, No. 3 (2017).

[14] Moffat, A. and Zobel, J.: Rank-Biased Precision for Measurement of Retrieval Effectiveness, *ACM TOIS*, Vol. 27, No. 1 (2008).

[15] Sakai, T.: Ranking the NTCIR Systems based on Multigrade Relevance, *Proceedings of AIRS 2004 (LNCS 3411)*, pp. 251–262 (2005).

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

[17] Sakai, T.: Evaluation with Informational and Navigational Intents, *Proceedings of WWW 2012*, pp. 499–508 (2012).

[18] Sakai, T.: Metrics, Statistics, Tests, *PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173)*, pp. 116–163 (2014).

[19] Sakai, T.: *Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power*, Springer (2018).

[20] Sakai, T.: How to Run an Evaluation Task: with a Primary Focus on Ad Hoc Information Retrieval, *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF* (Ferro, N. and Peters, C., eds.), Springer, pp. 71–102 (2019).

[21] Sakai, T. and Robertson, S.: Modelling A User Population for Designing Information Retrieval Metrics, *Proceedings of EVIA 2008*, pp. 30–41 (2008).

[22] Sakai, T. and Song, R.: Evaluating Diversified Search Results Using Per-Intent Graded Relevance, *Proceedings of ACM SIGIR 2011* (2011).

[23] Sakai, T. and Song, R.: Diversified Search Evaluation: Lessons from the NTCIR-9 INTENT Task, *Information Retrieval*, Vol. 16, No. 4, pp. 504–529 (2013).

[24] Sakai, T. and Zeng, Z.: Which Diversity Evaluation Measures are "Good"?, *Proceedings of ACM SIGIR 2019*, pp. 595–604 (2019).

[25] Sanderson, M., Paramita, M. L., Clough, P. and Kanoulas, E.: Do user preferences and evaluation measures line up?, *Proceedings of ACM SIGIR 2010*, pp. 555–562 (2010).

[26] Santos, R. L. T., Macdonald, C. and Ounis, I.: Search Result Diversification, *Foundations and Trends in Information Retrieval*, Vol. 9, No. 1, pp. 1–90 (2015).

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

[28] Voorhees, E. M. and Buckley, C.: The Effect of Topic Set Size on Retrieval Experiment Error, *Proceedings of ACM SIGIR 2002*, pp. 316–323 (2002).

[29] Zhou, K., Lalmas, M., Sakai, T., Cummins, R. and Jose, J. M.: On the Reliability and Intuitiveness of Aggregated Search Metrics, *Proceedings of ACM CIKM 2013*, pp. 689–698 (2013).

[30] 酒井哲也: 情報アクセス評価方法論: 検索エンジンの進歩のために，コロナ社 (2015).

[31] 川崎真未，Kang, I.，酒井哲也: 放棄セッションにおけるユーザ操作に着目したモバイル検索カードの順位付け，*IPSJ TOD*, Vol. 11, No. 3, pp. 1–11 (2018).

[32] 永田　靖，吉田道弘: 統計的多重比較法の基礎，サイエンティスト社 (1997).

[*6] http://waseda.box.com/fit2020sakaizeng