DEWS2004 I-2-03

ユーザの感覚を考慮したWeb検索システムの評価手法

大塚　崇志† 　江口　浩二†† 　山名　早人†

† 早稲田大学大学院理工学研究科　〒169-855 東京都新宿区大久保 3-4-1
†† 国立情報学研究所　〒101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: †{t-oh,yamana}@yama.info.waseda.ac.jp, ††eguchi@nii.ac.jp

あらまし　Web 空間は増加の一途を辿り，目的の情報を得ることは容易ではない．そのため Web 検索システムの必要性，更には性能の向上が求められており，検索エンジンの評価手法に関する研究が重要となってきている．しかし，既存の評価手法では，必ずしもユーザ自身の評価とは一致しない．これは，精度・再現率といった指標のみに基づき，ユーザの検索エンジンに対する感覚を考慮して来なかったことに原因がある．本報告では，Web 検索時にユーザがとる行動を考慮することにより Web 検索システムの性能を評価する“ ユーザ指向の評価尺度 ”を提案する．また，ユーザの満足度として検索に必要となる時間を計測することで，提案尺度の評価と従来の尺度との比較を行う．

キーワード　検索エンジン，情報検索，評価手法

An Evaluation Method of Web Search Engines based on Users' Sense

Takashi OHTSUKA†, Koji EGUCHI††, and Hayato YAMANA†

† Graduate School of Science and Engineering, Waseda University　3-4-1 Okubo Shinjuku-ku Tokyo, 169-8555 JAPAN
†† National Institute of Informatics　2-1-2 Hitotsubashi Chiyoda-ku Tokyo, 101-8430 JAPAN
E-mail: †{t-oh,yamana}@yama.info.waseda.ac.jp, ††eguchi@nii.ac.jp

Abstract　It is not easy to obtain the useful information from the Web space because the space is increasing and large. Therefore, the Web search system becomes indispensable and the improvement of its performance is required. Then, the researches on search engine's evaluation methods are being important. The evaluation based on the conventional evaluation methods, however, is not always equal to the user's evaluation. The reason is the conventional evaluation methods are based on precision and recall, that don't take user's sense for search engines into consideration. In this paper, we propose "a user oriented evaluation criterion" which evaluates the performance of Web search systems by considering user's actions when they retrieve Web pages. We also evaluate the proposed criterion in comparison with the conventional methods by measuring the spent time on search as the user's satisfaction degree.

Key words　Search Engine, Information Retrieval, Evaluation Method

1. はじめに

World Wide Web の普及に伴い，Web 空間の総容量は増加し続けている．そのためユーザが目的の情報を得るために，Web 検索エンジンの重要性が非常に高まってきた．

Web 上の情報の単位となるものが Web 文書であり，主に Web 文書を対象とした情報検索が Web 検索と呼ばれている．Web 文書には従来の情報検索が扱ってきたような新聞記事や論文などとは異なる特徴がある．具体例としては，ジャンルの多様性（論文，カタログ，日記などが混在），表現の多様性（タグを用いたレイアウト，表や画像），リンクによる参照（ハイパーリンクによるページ間の参照）などである．Web 検索システムではこれらの特徴に対応する様々な手法が提案され，検索エンジンに応用されているが，研究課題がまだ多く残されている．

特に Web 検索システムの評価に対する課題として，Web 検索に相応しい評価基準の欠如を挙げることができる．これはユーザの検索質問と検索結果に対する評価のモデルが十分に研究されていないためである．具体例として，検索結果の適合判定ではテキストだけを見るのか画像なども評価するのか，リンク先のページも見るのか，内容の信憑性や重要性をどのように評価するのかというように，従来の情報検索の評価尺度では評価できない事項が多く存在する．

そこで本稿では，Web 検索の特徴を考慮した新しい評価尺度を提案する．具体的にはユーザが Web 検索を行う際の特徴を評価基準に加えることを行う．更にその新尺度を用いることで，Web 検索システムの改善点を明確化する方法の応用を考察する．

本稿では，第 2 節において従来の評価手法とその特徴を述べ，Web 検索を評価するために更に必要と考えられる評価基準を述べる．第 3 節において Web 検索の特徴を考慮し，ユーザの感覚を評価基準として用いることによって検索システムの評価を行う新しい尺度を提案する．そして第 4 節において，従来の評価手法と提案手法を比較するための評価実験を行う．次に第 5 節において従来手法と提案手法の考察を行い，ユーザの感覚から得られる求めるべき検索結果とはどのようなものかを明確にする．最後に第 6 節でまとめを述べる．

2. 関 連 研 究

本節では，情報検索において伝統的に使用されてきた評価手法の特徴を述べる．また，現在の情報検索研究として新しく考案された評価尺度を説明する．

2.1 精度と再現率

精度と再現率は 1950 年代中期の Cranfield 実験 [1] で考案された伝統的な評価尺度である．

ある文書集合と検索質問集合を仮定する．ここで検索質問を 1 つ固定したとき，文書集合中に含まれる検索質問に対する適合文書の総数を R とする．ここである検索システムを使用して検索を実行する．その時 n 件の文書を出力し，その中に含まれる適合文書の数が r 件であったとする．このとき精度は r/n，再現率は r/R と定義される．

精度と再現率を用いた評価を Web 検索システムの評価として使用する場合には以下のような問題点が存在する．

- 再現率の算出が困難である：再現率を求めるためには適合文書の総数が必要であるが，データベースの規模が大きすぎるため，人手で適合・不適合の判断を下すことができず，適合文書総数が分からない．更に実際の検索エンジンにおいては，Web 上全てのデータを収集することができないので，実際の再現率を算出することは不可能である．
- 適合文書の数だけを評価するため，順位付け出力の評価ができない：データベースの増加により適合文書も増加しているため，Web 検索における順位付け出力の評価は行われるべきである．
- 適合または不適合の 2 値による評価のため，多段階の適合判定ができない：Web 文書のような多様なジャンルや表現法を持つ文書を，適合または不適合だけで判断することは難しく，有効であるとは考えにくい．
- そもそも文書集合における検索質問の適合・不適合の判定が仮定されている必要がある：検索対象となる文書集合，検索質問集合，各検索質問に対する適合文書集合の 3 要素を持つテストコレクションが必要である．テストコレクションは人工的に作成されるものであり，実際の Web 環境における評価法への適用が難しい．

これらの問題点を補うため，精度と再現率に基づく評価尺度の変形として次のような尺度が存在する [2]．

- 精度（λ）：検索結果上位 λ 件の文書集合における精度
- 再現率（λ）：検索結果上位 λ 件の文書集合における再現率

これらは，検索結果の上位部分だけを用いて，システム評価を行う尺度である．この概念は Web 検索においては有効であると考えられるが，λ の値を決定する明確な方法がない．検索質問ごとに相応しい λ を決定できなければ，システム評価に大きな影響を与えてしまうと考えられる．更に，大規模文書集合であることが予想される Web 検索において，検索結果上位 λ 件のみによる再現率を求めることの有効性自体問題である．

- R 精度：検索結果上位 R 件の文書集合における精度（R は適合文書の総数）

R 精度は，精度（λ）の λ の値として，適合文書の総数を取った尺度である．テストコレクションを用いた状況においては，上位に順位付けられた検索結果の有効性を示す尺度となる．しかし Web 検索においては普通，適合文書の総数を知ることはできない．

- （非補間）平均精度：最上位の文書から順に調べ，適合文書が出現した時点までの文書集合での精度を順次計算し，得られた全ての精度の平均値
- n 点平均精度 [3]：あらかじめ決められた n 点の再現率における精度の平均

n 点平均精度において，n としてよく用いられるのは，再現率 $0.0, 0.1, \cdots 0.9, 1.0$ の 11 点である(注1)．これらの尺度は R 精度のような尺度と比べ，検索結果のより下位の部分を評価する．つまりシステムをマクロ的な指標で評価することになる．これらの尺度でもテストコレクションが必要となるため，Web 検索への適用は難しい．

上記以外にも精度と再現率の両方に基づく尺度が存在している．しかし Web 検索においては再現率を求めることができないことから，精度と再現率に基づく評価尺度は Web 検索システムの評価に対しては妥当ではない．

次節において，これらの欠点を補うべく考案された評価尺度を述べる．

2.2 DCG [4]

DCG（Discounted Cumulative Gain）は Järvelin，K. と Kekäläinen，J. によって 2000 年に考案された新しい評価尺度である．

(注1)：適合文書の総数によっては 11 点での精度が正確に決定できない場合がある．このため補間精度と呼ばれる精度を用いる．

DCG は非 2 値による適合判定の使用に基づいた評価が可能である．そのため多段階適合性を用いた評価基準の尺度として使用できる．更に適合文書のランキングの位置を考慮することにより，適合度順出力の評価も可能である．以下で DCG を説明する．

まず，$d(i)$ は i 番目にランクされた文書を示し，$g(i)$ は文書 $d(i)$ の得点，$dcg(i)$ は上位 i 番目までにランクされているページの累積得点を示すとする．このとき，DCG は以下のように定義される．

$$dcg(i) = \begin{cases} g(1) & if\ i = 1 \\ dcg(i-1) + g(i)/\log_c(i) & otherwise \end{cases} \quad (1)$$

$$g(i) = \begin{cases} h & if\ d(i) \in H \\ a & if\ d(i) \in A \\ b & if\ d(i) \in B \end{cases} \quad (2)$$

ここで，H, A, B はそれぞれ高適合文書，適合文書，部分適合文書の集合であり，多段階の適合判定を用いることが可能である．h, a, b はそれぞれの適合度に割り当てられている重みである．また log の底 c はランクの位置に対する重み係数である．$c > 1$ を仮定すれば，c の値を大きくすればするほど，ランキング下位における文書への得点を高めることになる．

DCG は適合度の評価を行うと同時に，適合文書のランクを考慮している．適合文書が検索結果上位に出現している場合はより得点が高く，下位に出現するにつれ得点が減少していく．このようにして与えられた得点を累積することで，検索システムを評価する．

DCG は検索結果の上位部分だけでも評価が可能であり，適合文書の総数を用いる必要もない．Web 検索システムを評価するためには優れた尺度である．問題点としては，適合文書への重み付けを決定する適切な方法が無いことが挙げられる．また主にシステムの性能改善を目的とした評価というシステム指向の評価尺度であり，テストコレクションが必要であるため，実際の Web 検索の評価尺度として応用することが困難である．更に DCG は純粋な評価尺度であり，なぜ得点に差が生じるのかを特定することが困難であると考えられるため，検索システムの改善のために応用することが難しいと考えられる．

2.3 WRR [5]

WRR（Weighted Reciprocal Rank）は，1998 年に開始された日本における検索実験プロジェクトである NTCIR（NII/NACSIS Test Collection for Information Retrieval）[6] で採用された新しい評価基準である．NTCIR は TREC（Text REtrieval Conference）[7] における検索実験を参考にしており，現在までに 3 回の実験が実施されている．NTCIR-3 において研究課題として Web 検索が初めて導入され，その評価のため 2001 年に WRR が考案された．現在実行中の NTCIR-4 WEB タスクにおいて Web 検索の研究が行われている．以下で WRR を説明する．

まず，2.2 節と同様に $d(i)$ は i 番目にランクされた文書を示すとする．また H, A, B も 2.2 節と同様の定義である．更に，m は評価対象となる検索結果の上位 m 件を示す．このとき WRR は以下のように定義される．

$$wrr(m) = max(r(i)) \quad (3)$$

$$r(i) = \begin{cases} \delta_h/(i - 1/\beta_h) & if\ (d(i) \in H \wedge 1 \leqq i \leqq m) \\ \delta_a/(i - 1/\beta_a) & if\ (d(i) \in A \wedge 1 \leqq i \leqq m) \\ \delta_b/(i - 1/\beta_b) & if\ (d(i) \in B \wedge 1 \leqq i \leqq m) \\ 0 & otherwise \end{cases} \quad (4)$$

ここで $\delta_h \in \{0,1\}, \delta_a \in \{0,1\}, \delta_b \in \{0,1\}, \beta_b \geqq \beta_a \geqq \beta_h > 1$ は，それぞれが満たす重み係数である．

NTCIR における評価では m の値として $5, 10, 15, 20$ が使用された．また δ_x と β_x の組み合わせとして次の 2 レベルが使用された．

レベル 1：$(\delta_h, \delta_a, \delta_b) = (1, 1, 0), (\beta_h, \beta_a, \beta_b) = (\infty, \infty, \infty)$

レベル 2：$(\delta_h, \delta_a, \delta_b) = (1, 1, 1), (\beta_h, \beta_a, \beta_b) = (\infty, \infty, \infty)$

WRR は主に初出の適合文書が検索結果のどの程度上位に現れるかを評価する尺度である．そして WRR を複数の質問にわたって平均することによって評価を行う．検索結果の上位部分だけで評価可能であるため，ユーザが自分の情報要求を満たすシステムかどうかを決定するというユーザ指向性の評価と言える．また複数の質問間での平均を取ることにより，検索質問に依存するシステム間の性能差を緩和することができると考えられる．

しかし，検索結果の極めて上位部分の結果についてを対象とするため，適合文書の有無といった 2 値的な評価になり易い．またパラメータの重み付け調整が困難であると考えられる上，検索質問間毎のパラメータ変更も考慮する必要がある．

3. ユーザの感覚を考慮した評価手法

第 2 節で述べたとおり，現在の Web 検索システムの性能を評価する指標としては，主に精度と再現率，精度と再現率に基づいた尺度，そして NTCIR-3 で使用された DCG と WRR が使用されている．しかし精度と再現率に基づく評価指標は，検索対象の文書集合が Web 文書であるという特徴を考慮していない．また DCG を用いる場合は，Web 検索に相当するだけの大規模テストコレクションが無ければ，DCG を用いる意味がなくなってしまう．更に WRR ではパラメータの設定が困難であり，ユーザ指向の評価を妥当性を持って行うことが難しい．

そこで，実際の非常に巨大な Web 空間での性能評価が可能となるような評価指標が必要であると考えられる．本節では Web 検索の特徴を述べ，ユーザの Web 検索時の行動を考慮した，ユーザの感覚に基づくユーザ指向の評価尺度を提案する．

3.1 Web 検索の特徴

ユーザが実際に Web 空間での検索を行う場合，通常は検索結果の上位 30 件程度までしか検索結果を参照しないと考えられる [8]．つまり上位 30 件以下に検索された結果は意味を持たないと考えることができる．そこで検索結果の評価としては，上位にランキングされた文書の精度を評価対象とする．

またユーザが検索結果を参照する場合，ランキングの最上位から降順に結果を見ていくのが普通である [9]．そのためユーザがより参照し易い，ランキング上位部分に適合文書が並ぶことが望ましい．また検索結果の適合文書と不適合文書が混在してランキングされている場合よりも，適合文書は適合文書同士，不適合文書は不適合文書どうし隣り合っていたほうが，ユーザとしては検索結果の満足度が高くなると考えられる．例えば適合文書が連続して出力されれば，ユーザの入力した検索クエリが有効であると判断可能である．更に適合した関連ページそのものを参考にできることから，より良い検索語を思い付くことにつながると考えられる．また不適合文書が隣り合っていれば，入力した検索語を直ぐに変更することができると考えられる．

以上のような考えから，次のような評価尺度を提案する．

3.2 提 案 尺 度

Web 検索の特徴を評価の尺度に加えるために，新しい評価尺度 UCS（User's Character Score）を提案する．UCS を次のように定義する．

2.2 節と同様に，$d(i)$ は i 番目にランクされた文書であり，$s(i)$ を $d(i)$ の得点とする．また $s(1) = 1$ とする．更に 2.3 節と同様に m を定義する．X を適合文書集合としたとき，以下のように得点を定義する．

$$UCS(m) = \sum_{i=1}^{m} s(i) \tag{5}$$

$$s(i) = \begin{cases} 1 & if\ (d(i-1) \in X \wedge d(i) \notin X) \\ 1 & if\ (d(i) \in X \wedge d(i-1) \notin X) \\ a \cdot s(i-1) & if\ (d(i-1), d(i) \in X) \\ a \cdot s(i-1) & if\ (d(i-1), d(i) \notin X) \end{cases} \tag{6}$$

ここで a は $a > 1$ である重み係数である．

UCS では，検索結果が連続的に適合文書・不適合文書が並んだ場合に，各文書に大きな得点を与える．つまり検索結果の $d(i)$ と $d(i+1)$ が適合文書または，不適合文書として連続した場合に $d(i+1)$ の得点が高くなる．最終的に上位 m 件の累積和が UCS となる．UCS は，m の値を 30 程度にすることによって，ユーザが参照するだけの上位検索結果を評価対象とすることが可能である．つまりユーザが参照しているであろう文書に対しては同等の価値があり，参照しないであろう文書には価値が無いとみなすことになる．また m の範囲内においては，各文書の得点が検索結果のランクに影響されず，適合・不適合の連続性によってのみ影響される．これは参照している文書内では，文書の並び順が重要であることを考慮したものである．これによって UCS は，3.1 節に述べたようなユーザの満足度を測る尺度とみなすことが可能である．

3.3 提案する評価手法

ユーザの感覚を考慮するための評価手法として次の手法を提案する．

UCS とは別の評価基準として，検索に要した時間の評価を行う．ここで言う検索に要した時間とは，ユーザが検索結果の参照を開始した時から，検索結果に満足である，または不満足であると判断するまでの時間とする．また同時に，各ユーザ毎に検索結果に対して満足であるのか，不満足であるのかを判断してもらう．ここで満足または不満足という 2 値による判定を導入した理由として，以下の 2 つを挙げる．

- 多段階評価を行うための判断基準の設定が困難である（これは個人差による影響が大きくなることを緩和するためである）
- 評価の判断がどちらともつかない中間的なものに偏ってしまうことを避ける

評価の方法として，判断までの時間が長い検索結果の評価を低くすることと，不満足と判断された検索結果の評価を低くするという 2 方向からの評価を行う．つまり，検索時間が長く，不満足だと判断された検索結果が悪い評価であり，検索時間が短く，満足だと判断された検索結果が良い評価を受けるというものである．

検索に要した時間と，検索結果に対する満足・不満足という判断と UCS による評価を用いて検索システムの評価を行う手法とする．

4. 評 価 実 験

提案した尺度の性質を評価するための実験を行う．本節ではまず実験で使用したデータについてを述べる．次に実験の目的と内容を説明し，実験の結果を示す．

4.1 対象データ

本実験で使用したデータについてを述べる．

4.1.1 検 索 対 象

検索対象として使用されたデータは，NTCIR-3 Web 検索タスク用大規模文書データセット（NW100G-01）と同一であり，2001 年 8 月から 11 月にかけて "*.jp" ドメインの Web サーバから収集されたデータセットである．合計で約 100GB のテキストファイルとそのメタデータである．

検索結果は，実際に NTCIR-4 WEB へ提出された NTCIR-4 参加チームの異なる 5 種類の検索結果を使用した．NTCIR-4 は NW100G-01 を検索対象データとしているため，検索結果は NTCIR-3 Web 検索タスク用大規模文書データセット（NW100G-01）内のデータと同一である．

4.1.2 検 索 質 問

検索質問については NTCIR-4 [10] で使用されたデータ [11] を利用する．具体的には，20 名により作成された 267 課題の候補が作成された．検索質問の作成に当たり，検索対象のデータが 2001 年 8 月から 11 月のものであるため，その時点で存在しなかったと思われる事項を検索質問から除外した．不適切なものを除外した結果，153 課題を検索質問として選択した．NTCIR-4 参加チームが検索実行結果を提出した後，オーガナイザはそれらを分析することにより検索質問を 2 つのグループに分割する．一方が (a) 網羅的な適合判定に基づく評価用，他方が (b) 検索結果上位のみに対する適合判定に基づく評価用である．それぞれの課題数は，(a) に 50 課題前後，(b) に 100–150 課題程度（ただし (a) の課題群を含む）を見積もっている．(a) の検索質問から適合文書数が存在しないまたは数件程度しかな

いものと，適合文書が数十万件といった極端に多いものを除外した全 53 題の検索質問を今回の実験では使用した．図 1 に検索質問例を示す．

<TOPIC>
<NUM>0002</NUM>
<TITLE CASE="b"> トランペット, 価格, 特徴 </TITLE>
<DESC> トランペットの特徴およびその価格が記述されている文書を探したい。</DESC>
<NARR>
<BACK> 店舗ごとにどのような特徴をしてどれ位の値段のトランペットが売られているのかを知りたい。</BACK>
<TERM>「トランペット」とは、真鍮製でオーケストラや吹奏楽、ジャズの演奏会などで一般的に使用される種類の楽器を指す。プラスチック製などの、玩具としてその名が付いたものは該当しない。</TERM>
<RELE> トランペットの製品番号・その形状や音色などの特徴・店頭価格が全て記載されている文書が適合する。</RELE>
</NARR>
<ALT0 CASE="b"> トランペット </ALT0>
<ALT1 CASE="b"> トランペット, 特徴, 価格 </ALT1>
<ALT2 CASE="b"> トランペット, 特徴, 値段 </ALT2>
<ALT3 CASE="c" RELAT="2-3"> トランペット, 特徴, 価格 </ALT3>
<USER> 大学 2 年, 女性, 検索歴 6 年, 熟練度 3, 精通度 4</USER>
</TOPIC>

図 1　検索質問例

検索質問は以下の項目で構成されている．

- <NUM> ('topic number') 検索質問 ID の番号を示す．
- <TITLE> ('title') は，課題作成者が現実のサーチエンジンへ投入することを模擬した 1–3 語からなる検索質問である．これらは検索において重要な語から順に並べられている．<TITLE> は「CASE」属性を伴っており，これは以下のような検索戦略の型を示すものである．

(a)　すべての語間の関係に OR 演算子を用いることができる場合

(b)　すべての語間の関係に AND 演算子を用いることができる場合

(c)　3 語のうち 2 語のみの関係に OR 演算子を用いることができる場合

＊ それらは「RELAT」属性で特定される．

- <DESC> ('description') は情報ニーズの最も基本的な記述であり，一文程度で表現される．
- <NARR> ('narrative') は，検索の背景・目的，語の定義，および，適合判定基準の補足を数段落で記述したものである．これらはそれぞれ <BACK>，<TERM> および <RELE> タグで示される．ただし，<BACK> は常に記述されるが，残り 2 つのタグが常に記述されるとは限らない．
- <ALT0> ('alternative query 0') は，検索質問における TITLE の先頭一語のみを機械的に抽出することにより作成した検索質問である．ただし，TITLE がもともと一語のみである場合は省略した．TITLE は課題作成者がサーチエンジンに入力することを想定した 1–3 語からなる検索質問であり，TITLE を構成する語は課題作成者が検索において重要と判断したものから順に並べられている．したがって，ALT0 には検索において最も重要であると判断された語が定義されていることになる．
- <ALT1>,<ALT2>,<ALT3> ('alternative query 1, 2 and 3') は，課題作成者とは別人の 3 名が（元々の TITLE 部分が予め削除された）検索質問を眺めて，独自に付与した検索質問である．ただし，TITLE と語の並び順も含めてまったく同一のタグは省略したため，3 名分のすべてが定義されているとは限らない．ALTn の様式はタグ名を除いては TITLE と同一である．
- <USER> ('user attributes') には，検索質問作成者の属性として，職業，性別，検索歴，検索熟練度，話題精通度を記す．

4.2 実験説明

提案した評価手法と，実際のユーザの感覚とのマッチングを検証することを目的に評価実験を行った．

実験では，全 53 題の検索質問それぞれに対し，NTCIR-4 の異なる参加チームの検索システムによって得られた 5 種類の検索結果を評価者に提示する．5 種類の検索結果は各検索質問ごとに表示する順番を変え，表示する順番による検索時間や，検索評価への影響を緩和した．

評価者には，検索課題を理解してもらった後，検索結果を提示する．検索結果としては，ユーザが参照すると考えられる上位 30 件を提示する．提示した検索結果に対して，評価者にそれぞれの検索結果の評価を実行してもらう．評価対象者は日常的にインターネットを使用している人物で，全 26 人により評価実験が実行された．

評価者には，提示された検索結果を参照してもらうことにより，実際に検索結果の満足・不満足を評価してもらう．また，評価者の検索方法についてであるが，全ての検索結果を参照する必要はない．評価者が満足または不満足だと即断すれば，検索結果の一部分のみの参照でも問題はない．これは文書の連続性に基づく満足度の上昇を評価するとともに，検索過程とユーザインタラクションの評価へもつながると考えたためである．

4.3 評価方法

検索結果を評価する方法として，提案した評価手法を用いる．

また UCS と DCG,WRR との比較を行い，総合的な評価を行う．それぞれの評価尺度に対するパラメータは次のように設定した．

DCG の計算には 2.2 節における各パラメータを $(h, a, b) = (3, 2, 1)$ かつ，$c = 2$ とした．

WRR の計算には 2.3 節における各パラメータを $(\delta_h, \delta_a, \delta_b) = (1, 1, 1), (\beta_h, \beta_a, \beta_b) = (\infty, \infty, \infty)$ とした．

UCS の計算には 3.2 節におけるパラメータ a として $a = 1.1$ を使用した．

4.4 実験結果

はじめに，UCS，DCG，WRR と全 53 検索課題に対する平均検索時間とを 5 種類の各検索結果に対して比較したものを

表 1 に提示する．表 1 における DCG と UCS の得点付けにお

表 1　UCS，DCG，WRR と平均検索時間の比較

	DCG		WRR	UCS		平均検索時間
	id_lack	id_poss		id_lack	id_poss	
検索結果 1	7.70	16.39	0.67	47.66	50.51	74.09
検索結果 2	4.97	11.09	0.44	56.89	56.20	90.42
検索結果 3	3.71	7.59	0.38	76.60	76.62	79.15
検索結果 4	1.45	3.38	0.17	111.93	113.60	80.94
検索結果 5	7.33	15.73	0.64	47.73	54.32	70.30
平均	5.03	10.84	0.46	68.16	70.25	78.98
平均検索時間との相関係数	-0.51	-0.49	-0.54	0.28	0.20	

いて，id_lack と id_poss の 2 種類の得点付けを行った．ここで id_lack とは，検索結果における重複ページを不適合文書として適合判定を行った場合である．また id_poss とは，検索結果における重複ページを考慮せず，各文書の適合度に従った場合である．

表 1 の結果から，平均検索時間は検索結果 5 が最も短く，全検索結果の検索時間の平均よりも 8.6 秒短い．このことから，検索結果 5 が最も参照しやすい検索結果であると判断することができる．逆に検索結果 2 は平均よりも 11.4 秒も長い時間検索が行われていることから，検索結果 2 は最も参照し難い検索結果であると判断することができる．各尺度の得点と，平均検索時間の相関係数を求め，各尺度と平均検索時間との相関を求めたが，明確な相関は見られなかった．

次に UCS，DCG，WRR と平均満足得点とを 5 種類の各検索結果に対して比較したものを表 2 に提示する．ここでいう平均満足得点とは，各検索質問に対して満足と判断した評価者の人数である満足得点（26 点満点）の平均である．表 2 の結果か

表 2　UCS，DCG，WRR と平均満足得点の比較

	DCG		WRR	UCS		平均満足得点
	id_lack	id_poss		id_lack	id_poss	
検索結果 1	7.70	16.39	0.67	47.66	50.51	17.81
検索結果 2	4.97	11.09	0.44	56.89	56.20	12.17
検索結果 3	3.71	7.59	0.38	76.60	76.62	7.58
検索結果 4	1.45	3.38	0.17	111.93	113.60	3.72
検索結果 5	7.33	15.73	0.64	47.73	54.32	17.08
平均	5.03	10.84	0.46	68.16	70.25	11.67
平均満足点数との相関係数	0.99	1.00	0.98	-0.95	-0.92	

ら平均満足得点は検索結果 1 が最も高く，ユーザが最も良いと評価した検索結果であるということができる．逆に検索結果 4 は平均満足得点が最も低いことから，ユーザにとって最も悪い検索結果であると判断できる．

表 1 と同様に，各尺度の得点と平均満足得点との相関係数を求めた．表 2 より，DCG の id_lack，id_poss と WRR は平均満足得点と高い相関を持つことが分かる．また，UCS の id_lack，id_poss は平均満足得点と高い負の相関を持つことが分かった．

5. 考　　察

UCS の得点は，DCG や平均満足度と比較しても分かるように，不適合文書の連続に対して高い得点が与えられる傾向にある．そのため，UCS と DCG，WRR を直接比較することができない．そこで UCS のパラメータを変更し，直接比較可能な尺度を UCS2 として提案する．

UCS の得点付けにおいて，不適合文書の連続に高い得点が与えられることから，UCS2 では，不適合文書の連続に対して適合文書の連続とは異なる得点を与える．具体的には 3.2 節のパラメータ a を $a = 0.9$ とした（適合文書の連続には $a = 1.1$ を使用）．これは，文書の得点を下げることから，不適合文書の連続に対しペナルティを与えることと同義である．UCS2 を DCG，WRR と比較し，表 3 に示す．表 3 における UCS2 と

表 3　DCG，WRR と UCS2 の比較

	DCG		WRR		UCS2	
	id_lack	id_poss	id_lack	id_poss	id_lack	id_poss
検索結果 1	7.70	16.39	0.67	0.67	27.99	31.83
検索結果 2	4.97	11.09	0.44	0.44	23.59	25.08
検索結果 3	3.71	7.59	0.38	0.38	21.23	24.04
検索結果 4	1.45	3.38	0.17	0.17	18.90	21.42
検索結果 5	7.33	15.73	0.64	0.64	27.18	35.14
平均	5.03	10.84	0.46	0.46	23.78	27.50
UCS2 との相関係数	0.99	0.93	0.99	0.93		

の相関係数は次のようにして求められている．DCG の id_lack に対して UCS2 の id_lack，DCG の id_poss に対して，UCS2 の id_poss の相関を求めている．WRR の id_lack，id_poss に対しても同様である．

表 3 から分かるように，UCS2 は DCG，WRR と非常に高い相関を持っている．

このことから，UCS2 は DCG の代用として利用することも可能であると考えられる．また，UCS2 と DCG が高い相関を持つことから，適合度と文書の連続性にも相関があると考えることができる．UCS2 の利点として次の特徴が挙げられる．

- 上位 m 件のみを適合または不適合の 2 値で判定するだけのコストで検索システムの評価を行うことが可能
- DCG に要求されるような大規模テストコレクションや判断基準の困難な多段階適合性を必要としない

次に UCS2 と平均検索時間との相関を検証した結果を表 4 に示す．表 1 と比較して，id_lack と平均検索時間の相関は DCG や WRR と同程度であるが，id_poss と平均検索時間の相関が強くなっていることが分かる．この特徴は，UCS が文書の連続性に基づいた，ユーザの参照のしやすさを考慮したものであることに由来すると考えることができる．id_lack では重複文書を不適合としていることから，連続性の評価という影響が弱まっているため，DCG や WRR と同程度の相関になったと推測できる．

表 4　UCS2 と平均検索時間の比較

	UCS2		平均検索時間
	id_lack	id_poss	
検索結果 1	27.99	31.83	74.09
検索結果 2	23.59	25.08	90.42
検索結果 3	21.23	24.04	79.15
検索結果 4	18.90	21.42	80.94
検索結果 5	27.18	35.14	70.30
平均	23.78	27.50	78.98
平均検索時間との相関係数	-0.53	-0.73	

6. 今後の課題とまとめ

今後の課題として，今回のような一回の検索語に対する検索結果の評価を行うだけではなく，複数回の検索語の変更を考慮したよりインタラクティブな評価方法を考案する必要がある．

本稿ではユーザの感覚を考慮した Web 検索システムの評価手法を提案した．実験結果から，提案した尺度と検索時間との相関を認めることはできなかった．提案尺度のパラメータを変更することにより，ユーザの満足得点と DCG との相関を認めることができた．またそれによって，ユーザの感覚を評価に加えることができた．テストコレクションの作成に関するコストを低減できることから，提案した尺度は実際の検索システム評価へ応用することができると考えられる．

謝　　辞

本研究の一部は，科学技術振興費「e-Society」，及び，文科省 21 世紀 COE「プロダクティブ ICT アカデミア」によるものである．また本研究では，国立情報学研究所が主催する NTCIR-4 WEB タスクのデータの一部を，同プロジェクトのオーガナイザの一員として使用した．

文　　献

[1] C.W. Cleverdon, “The significance of the Cranfield tests on index languages, ” In A. Bookstein, Y. Chiaramella, G. Salton, and V.V. Raghavan, editors, Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3–12, ACM Press, Chicago, October 1991.

[2] David Hawking, Ellen Voorhees, Nick Craswell, Peter Bailey, “Overview of the TREC-8 Web Track, ”Proceedings of the 8th Text REtrieval Conference, NIST Special Publication 500–246, pp.131–149, 1999.

[3] R.Baeza-Yates, “Modern Information Retrieval, ”Addison Wesley Longman Publishing, 1999.

[4] Kalervo Järvelin & Jaana Kekäläinen, “IR evaluation methods for retrieving highly relevant documents, ”Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.41–48, New York, 2000.

[5] Koji Eguchi, Keizo Oyama, Emi Ishida, Noriko Kando, and Kazuko Kuriyama, “Evaluation Methods for Web Retrieval Tasks Considering Hyperlink Structure, ”IEICE Transactions on Information and Systems, Vol.E86-D, No.9, pp.1804–1813 , Sep. 2003.

[6] -:“NII-Test Collection for IR Home Page, ”⟨*http* : *//research.nii.ac.jp/ntcir/index − ja.html*⟩.

[7] -:“Text REtrieval Conference (TREC) Home Page, ”⟨*http* : *//trec.nist.gov/*⟩.

[8] Amanda Spink, B. J. Jansen, D. Wolfram & T. Saracevic, “From E-Sex to E-Commerce: Web Search Changes,”IEEE Computer, 35(3), pp.107–109, Pennsylvania, Mar.2002.

[9] -:“Japan internet.com デイリーリサーチ - 検索結果は上から順に．上位表示サイトのクリック率高まる，87% -, ”⟨*http* : *//japan.internet.com/research/*20020320/1.*html*⟩.

[10] -:“NTCIR-WEB, ”⟨*http* : *//research.nii.ac.jp/ntcweb/index−ja.html*⟩.

[11] Koji Eguchi, et al., “Overview of the Informational Retrieval Task at NTCIR-4 WEB,” Working Notes of the 4th NTCIR Workshop Meeting, Tokyo, June 2004 (to appear).