

Web 検索のための評価ワークショップに適したシステム評価手法 System Evaluation Methods Suitable for Evaluation Workshops on Web Retrieval

江口 浩二^{*} 大山 敬三 石田 栄美[▽]
神門 典子 栗山 和子^{*}

Koji EGUCHI Keizo OYAMA Emi ISHIDA
Noriko KANDO Kazuko KURIYAMA

著者らは、2001年から2002年に開催された第3回NTCIRワークショップにおいてWeb検索タスク（以下、NTCIR-3 WEB）を実施した。その主な目的は、共通のデータセットを用いてWeb検索エンジンの検索有効性を評価するとともに、Web検索に関する再利用可能なテストコレクションを構築することであった。上記の目的のもと、検索対象の文書データとしてJPドメインから収集した100GB及び10GBのWebページデータを使用し、多様な入力形態による検索実行結果に対して、複数のユーザモデル、複数の文書モデルを仮定した評価を行った。ここに入力形態として、語、文、文書を設定した。また、ユーザモデルとして、網羅的に適合文書を求めるモデルと検索結果上位における精度を重視したモデルを想定した。文書モデルとしては、ページを単位としたモデル、リンクで結合された文書群を単位としたモデルを仮定した。本稿では、NTCIR-3 WEBにおいて提案された評価手法について述べている。

The authors conducted the Web Retrieval Task ('NTCIR-3 WEB') from 2001 to 2002 at the Third NTCIR Workshop. In the NTCIR-3 WEB, they attempted to assess the retrieval effectiveness of Web search engine systems using a common data set, and to build re-usable test collections that are suitable for evaluating Web information retrieval systems. With these objectives, they evaluated on searches using various types of user input, user models and document models. As the document data sets, they constructed 100-gigabyte and 10-gigabyte document collections that were gathered from the 'jp' domain. The user input were given as query term(s), sentence(s), and document(s). They assumed two user models where comprehensive relevant documents are required, and where precision of the top-ranked results is emphasized. They also assumed two document models, a document as an individual page, and a document as a page set connected by hyperlinks. This paper discusses evaluation methods proposed in

^{*} 正会員 国立情報学研究所
eguchi@nii.ac.jp, oyama@nii.ac.jp
[▽] 非会員 国立情報学研究所
emi@nii.ac.jp, kando@nii.ac.jp
^{*} 非会員 白百合女子大学
kuriyama@shirayuri.ac.jp

NTCIR-3 WEB.

1. はじめに

我々は、2001年8月から2002年10月にかけて開催された第3回NTCIRワークショップにおいて、Web検索タスク（以下、NTCIR-3 WEB）を実施した[1,2]。これは、共通のデータセットを用いてWeb検索エンジンの検索有効性を評価するとともに、Web検索に関する再利用可能な実験評価用テストコレクション（以下、Webテストコレクション）を構築することを目的とした評価ワークショップ[3]である。一般にテストコレクションは次の三つにより構成される。すなわち、(i) 文書データセット、(ii) 情報ニーズを記述した検索課題、(iii) 各検索課題に対する適合判定結果リストである。NTCIR-3 WEBは、これら三つの構成要素を用いて検索システムの有効性を比較評価するという点では、従来のテキスト検索システムに関する評価ワークショップ[3]と同様であるものの、上記(i)～(iii)の各々についてWeb検索に特徴的な観点を採り入れた。本タスクは以下の手順で実施された。

- オーガナイザ（本稿の著者ら）は、JPドメインから独自にWebページを収集し、100GB及び10GBの文書データ及びそれらの関連データを構築した。
- 各参加グループは、オーガナイザが設計した数種のタスクからいくつかを選択して、オーガナイザが提供した検索課題をもとにしてクエリを生成し、前記の文書データに対する検索実行結果をオーガナイザに提出した。
- オーガナイザは、検索実行結果に対して適合判定を実施し、その判定結果をもとに種々の観点からの評価尺度を用いてシステムごとの最終的な評価値を算出した。

本稿では特に、NTCIR-3 WEBにおいて導入された評価手法のうち、特徴的な提案について論ずる。

2. タスク設計

Web検索システムは従来のテキスト検索システムとは異なる点が多い。NTCIR-3 WEBではWeb検索の特徴を鑑みて、(a)サーベイ検索タスク（(a-1)トピック検索タスク、(a-2)類書検索タスク）、(b)ターゲット検索タスク、(c)自由課題タスク（(c-1)分類出力タスク、(c-2)音声入力タスク）を設計した。本論文では(a-1)及び(b)を中心に記述する。目的及び手順をそれぞれ2.1節と2.2節に示す¹。これらは、後述の通り、前提とするユーザモデルが互いに異なる。

2.1 サーベイ検索タスク

サーベイ検索タスクは、ユーザが網羅的に適合文書を探すようなユーザモデルを仮定している。サーベイ検索タスクはクエリの表現形式に応じて、(i)語や文を用いて検索するトピック検索タスク、(ii)文書を用いて検索する類書検索タスクに分けて実施した。特に(i)について以下に述べる。

トピック検索タスクは、従来のAd Hoc型検索[3,4]に相当するものであるが、3.3節や4章で後述する通り、適合判定や評価尺度等についてWeb検索の種々の特徴を反映させた。各参加グループは、検索課題ごとに実行した検索結果の上位1,000件に対応する文書IDのランク付きリストを提出した。また、参加グループは、検索課題を構成する<TITLE>のみを用いた検索と、<DESC>のみを用いた検索の実行結果リストを少なくともも提出することが期待された。検索課題の詳細については3.2節に示す。

2.2 ターゲット検索タスク

ターゲット検索タスクは、事実を確認し得る記述や疑問に

¹ その他のタスクの詳細に関しては文献[1]を参照されたい。

対する答えを含んだ少数の文書を求めるようなユーザモデルを仮定しており、この場合、検索結果上位における精度が強調される。実行結果リストの提出に関わる条件は、2.1節に述べたサーベイ検索タスクとほぼ同様であるが、提出すべき実行結果リストは各検索課題に対して上位20件のみでよい。4章に述べる通り、検索結果の上位のみに着目した評価尺度が適用された。

3. Webテストコレクション

Webテストコレクションは、(i)文書データセット、(ii)情報ニーズを記述した検索課題、(iii)各検索課題に対する適合判定結果リストにより構成される。これら構成要素のそれぞれは、現実のWebの環境に適合したものとなることを目指した。上記(i)～(iii)のそれぞれを3.1節、3.2節及び3.3節に示す。

3.1 文書データセット

NTCIR-3 WEBにおいては、2001年8月から同年11月にかけてJPドメインから収集された次の二種類の文書データを用意した。すなわち、(i)100GBの文書データ(NW100G-01)と、(ii)その部分からなる10GBの文書データ(NW10G-01)である²。これらを構成する大部分のWebページは日本語もしくは英語で記述されているが、それら以外の言語の記述も含まれる。さらに、(i)と(ii)のそれぞれに対して、それらを構成する個々の文書からリンクが張られたページ群のリストを提供した。ただし、これらはJPドメインに限定しなかった。NW100G-01を構築するにあたって、JPドメインのHTTPサーバの任意のポートを対象に、HTMLまたはPlain Text形式のファイルに限定してクローリングを行った。

個々の文書はページコンテンツとそのメタデータで構成される。メタデータとしては文書ID、クローリング日時、Content-Type、URL、HTTPヘッダを用いた。

3.2 検索課題

オーガナイザは情報ニーズが記述された検索課題を提供した。その形式は過去のNTCIRワークショップのものを参考にした[3]が、<TITLE>の定義、<RDOC>及び<USER>の新たな導入、<NARR>の形式については独自に提案した。検索処理での使用が必須となるタグ部分と、使用が許可されるタグ部分はタスクごとに指定された。検索課題は下記のタグで示された部分により構成された。図1に検索課題例を示す。

- <TOPIC> は検索課題の境界を示す。
- <NUM> は検索課題IDを示す。
- <TITLE> はサーチエンジンへ投入されるクエリを模擬して、以下のように定義した。すなわち、検索課題作成者が、三つの戦略、(a)同義語・類義語を並べる場合、(b)異なる観点の語を並べる場合、(c)前記の(a)と(b)の複合のなかから、適切と思われる戦略を選択し、それにそってサーチエンジンに入力するであろう三語以内の語を並置した³。ただし、重要な語から順に左から並べることとした。なお、図1の例における<TITLE CASE="c" RELAT="2-3">は戦略(c)においてタイトル中の二番目の語と三番目の語が同義あるいは類義関係にあることを示している。
- <DESC> は一文で表現される情報ニーズの最も基本的な記述である。

² 参加グループは、国立情報学研究所に設置されたオープンラボラトリにおいてのみ、文書データを処理できるとする制限が課せられた。ただし、インデックスファイルなどの加工済みデータを持ち出すことが許可された。また、文書データの処理を目的とするリモートアクセスが許可された。

³ 参加グループがブル型検索を用いる場合、(a)と(b)に対してはそれぞれOR演算とAND演算、(c)はOR演算とAND演算の組合せとみなすことができる。

```
<TOPIC>
<NUM>0004</NUM>
<TITLE CASE="c" RELAT="2-3">コンピューターウイルス、予防、
対策</TITLE>
<DESC>コンピューターウイルスの予防方法や対策法について説明
している文章を探したい</DESC>
<NARR><BACK>インターネット利用が爆発的に普及する中でコ
ンピュータウイルスは日常的な問題にまで近づいてきている。そこ
でどのような予防法をとり、もし感染してしまった際にはどのよう
な対処をすればよいのか知っておきたい。</BACK><RELE>適合文
書は、コンピューターウイルスへの予防・対策についての情報を提供
するもの。被害届出やウイルスの種類についてのみ述べているもの
は適合としない。特定のウイルスについてのみ情報を提供するペー
ジは部分的適合とする。</RELE></NARR>
<CONC>コンピューターウイルス、ワーム、情報セキュリティ、不正
アクセス、予防、対策、感染</CONC>
<RDOC>NW003214039, NW013338047, NW013315769</RDOC>
<USER>大学院修士1年、男性、検索歴5年</USER>
</TOPIC>
```

図1 検索課題の例
Fig. 1 A Sample Topic

- <NARR> は、検索の背景・目的、語の定義、及び、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>、<RELE>タグで示される。ただし、これら三つが常に記述されるとは限らない。
- <CONC> は同義語や類義語、関連語からなるリストを示す。これらは検索課題作成者により定義された。
- <RDOC> は三件以下の適合文書のIDを示す。これらは類書検索タスクで使用される。
- <USER> には、検索課題作成者の属性が簡潔に記された。

3.3 適合判定

プーリングと適合判定

情報検索に関する評価ワークショップ[3]において、各実行結果リスト上位の一定件数を取り出し併合する処理をプーリングと呼ぶ。プーリングによって得られた文書集合(以下、文書プール)は、所与の検索課題に対する適合文書集合を見積るのに用いられる。NTCIR-3 WEBでは従来のプーリング手法[3]を用いたが、文書プールに対してメタサーチエンジン手法を適用してランク付けを行った[1]。また、各実行結果リストの上位100件をプーリングに使用した。

判定者が文書プール中の各文書の適合性を判定する際は、後述する複数の文書モデルの仮定のもと、多値適合レベルすなわち高適合、適合、部分的適合、不適合のいずれかに判定した。

文書モデル

Webページの表現形式は多様であり、Webにおける情報の基本単位は、ハイパーリンクで結合されたページ群である場合、個々のページの場合、あるいはページ中のパッセージである場合がありうる。TREC Web Track[4]では個々のページが情報の単位と仮定された。この仮定のもとでは、複数のAuthority的ページへのリンクを備えたHub的ページ[5]は、そのページ自体に適合情報を含まない限り、不適合とみなされる。しかしながら、Webの環境においては、しばしばHubのページが上記の仮定に基づく適合ページよりもユーザにとって有用である。以上の考えのもと、NTCIR-3 WEBでは、適合判定に際して次の三つの文書モデルを仮定した。

- **1クリック距離文書モデル**: 判定者が所与のページの適合

性を判定する際に、そのページの内容だけでなく、それからリンクが張られたページ群のうち文書プールに含まれるものの内容を参考にすることを許可するものである。これは適合文書の大部分が文書プールに含まれるという仮定に基づく。

- **ページ単位文書モデル:** 判定者が所与のページの適合性を判定する際、そのページの内容にのみ基づくものである。これは従来と同様のモデルである。
- **パッセージ単位文書モデル:** 判定者が、ページ中で適合性の根拠を与えるようなパッセージを特定するものである。

4. 評価尺度

各参加グループのサーチエンジン・システムによる実行結果リストを評価するため、サーベイ検索タスクでは上位1,000件以下、ターゲット検索タスクでは上位10件に着目した。評価尺度としては、サーベイ検索タスクに対しては、(i) 精度と再現率に基づく尺度、及び、(ii) DCG [6]を使用した。ターゲット検索タスクに対しては、(i)と(ii)に加えて、(iii)WRR [2]を使用した。上記の評価尺度はページを基本単位として設計されたものであるが、適合判定の前提となる文書モデルが1クリック距離文書モデルかページ単位文書モデルのいずれであるかによって適合文書集合が異なる。

4.1 精度と再現率

精度と再現率に基づく評価尺度として、サーベイ検索タスクにおいては、非補間平均精度 (aprec)、及び、R精度⁴ (rprec) を求めた[7]。上記の評価尺度は、情報検索に関する従来の評価ワークショップ[3]で用いられたものと同様である。一方、ターゲット検索タスクにおいては、検索結果上位10件の精度 (prec(10)) を求めた。

ところで、上記のような精度と再現率に基づく評価尺度は、しばしば多値適合レベルを二値の適合レベルに縮退させることを要求する。従って、以下の適合レベルを仮定した。

- 適合レベル1: ある文書が多値適合レベルにおいて高適合あるいは適合であるとき、二値レベルにおいては適合であるとみなし、そうでなければ不適合であるとみなす。
 - 適合レベル2: ある文書が多値適合レベルにおいて高適合、適合あるいは部分的適合のとき、二値レベルにおいては適合であるとみなし、そうでなければ不適合であるとみなす。
- 適合レベル1及び2のそれぞれに従って、実行結果リストごとに、前記の各評価尺度の全検索課題に関する平均値を求めた。

4.2 DCG

DCG (Discounted Cumulative Gain) は多値適合レベルに適した評価尺度であり、次式で表される[6]。

$$dgc(i) = \begin{cases} g(1) & \text{if } i=1 \\ dgc(i-1) + g(i) / \log b(i) & \text{otherwise,} \end{cases}$$

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

ここに $d(i)$ は上位 i ランクの文書を示し、H、A及びBはそれぞれ高適合、適合、部分的適合と判定された文書集合を示す。式(1)で定義される利得の度合いとしては、4.1節に示した適合レベル1に対して $(h, a, b) = (3, 2, 0)$ 、適合レベル2に対して $(h, a, b) = (3, 2, 1)$ とした。また、式(1)における対数関数の底は $b = 2$ とした。NTCIR-3 WEBのサーベイ検索タスクにおいては実行結果リストの上位1,000ランクまで、ターゲット

検索タスクにおいては上位10ランクまでのDCGを求め、ランクごとに全検索課題にわたる平均値を計算した。

4.3 WRR

MRR (Mean Reciprocal Rank) は、しばしば質問応答システムの評価に用いられ、各質問に対する実行結果リストにおける初出の正解のランクの逆数を、全質問にわたって平均した値で定義される[8]。我々は、MRRの発想を多値適合レベルに対応するよう拡張したWRR (Weighted Reciprocal Rank) を提案し、ターゲット検索タスクの評価に適用した[2]。WRRは次式で定義される $wrr(m)$ の全検索課題にわたる平均値として求められる。

$$wrr(m) = \max(r(m)),$$

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

ここに m は実行結果リストにおいて評価の対象とするランクの最大値を示す。また、重み係数は $\delta_h \in \{1, 0\}$, $\delta_a \in \{1, 0\}$, $\delta_b \in \{1, 0\}$ 、及び $\beta_h, \beta_a, \beta_b > 1$ のそれぞれを満たすものとする。なお、式(2)において β_x (ただし $x \in \{h, a, b\}$) の値が十分大きいとき、 $(-1/\beta_x)$ の項を省略することができる。式(2)において $m=10$ としてWRRを計算した。このとき、4.1節に示した適合レベル1に対して $(\delta_h, \delta_a, \delta_b) = (1, 1, 0)$ 、適合レベル2に対して $(\delta_h, \delta_a, \delta_b) = (1, 1, 1)$ とした。また、本稿では簡単のため、いずれの適合レベルにおいても $(\beta_h, \beta_a, \beta_b) = (\infty, \infty, \infty)$ と仮定した。

さらに、実行結果リストにおいて着目するランクの最大値が10位の場合において、適合文書が一つも見つからなかった検索課題の比率を適合レベルごとに求めた (%nf(10))。

4.4 ページ内容の重複を考慮した評価手法

多くのユーザはサーチエンジンの検索結果に重複したページが出現することを望まないと考え、以下のようにページ内容の重複を考慮した評価手法を提案した。

- 各実行結果リストにおいて、重複文書群のうち初出の文書については、特別な措置をとらずに他の文書と同様に扱う。
- 重複文書群で二番目以降に出現する文書については、たとえそれが適合と判定された文書であっても、不適合(もしくは部分的適合)として扱う。

この評価手法は、精度・再現率に基づく評価尺度やDCGと組み合わせることを前提とし、結果として重複文書を含む実行結果リストにペナルティが課せられる。重複文書群を特定するため、判定者は適合文書間の重複性について可能な範囲で判定した。また、ページ内容が完全に一致する文書群を機械的に検出し、判定者の重複判定を補完した。なお、主要な部分が一一致する類似文書群や、ハイパーリンクで何らかの関連性をもつて接続された関連文書群を用いることも可能であり、それらを1クリック距離文書モデルに基づく適合判定結果と組み合わせることで、トピック・ディステーション手法[5]の妥当な評価手法になり得ると期待する。

5. 評価結果

個々の参加グループは、多様な研究目的のもとNTCIR-3 WEBに従事し⁵、六つの参加グループが不備なく実行結果リストを提出した。また、オーガナイザが、文書プールの網羅性を改善することを目的として、実行結果リストを提出した。

⁴ 検索結果のランク $|R|$ 位における精度として定義される。ここに $|R|$ は各検索課題に対する適合文書数を示す。

⁵ 個々の検索システムの詳細については、以下の論文集を参照されたい。
<<http://research.nii.ac.jp/ntcir/publication1-en.html>>。

表1 種々の評価尺度に基づくシステム順位
Table 1 System Ranking Based on Several Evaluation Measures

Topic Part	aprec	rprec	dcg(100)	dcg(1K)	prec(10)	dcg(10)	wrrt(10)	%nfl(10)
<TITLE>	GRACE-LA1-1	<u>GRACE-LA1-2</u>	GRACE-LA1-1	GRACE-LA1-1	GRACE-LB-1	GRACE-LB-1	<u>K3100-13</u>	<u>K3100-13</u>
<TITLE>	<u>GRACE-LA1-2</u>	GRACE-LA1-1	<u>GRACE-LA1-2</u>	<u>GRACE-LA1-2</u>	<u>GRACE-LB-2</u>	<u>GRACE-LB-2</u>	GRACE-LB-1	<u>K3100-14</u>
<TITLE>	OKSAT-WEB-F-04	ORGFREF-LA1-6	OKSAT-F-04	OKSAT-F-04	<u>K3100-14</u>	<u>K3100-14</u>	<u>GRACE-LB-2</u>	<u>GRACE-LB-2</u>
<TITLE>	ORGFREF-LA1-6	OKSAT-WEB-F-04	ORGFREF-LA1-6	ORGFREF-LA1-6	<u>K3100-13</u>	<u>K3100-13</u>	<u>K3100-14</u>	GRACE-LB-1
<TITLE>	<u>K3100-05</u>	<u>K3100-05</u>	<u>K3100-05</u>	<u>K3100-05</u>	ORGFREF-LB-6	ORGFREF-LB-6	ORGFREF-LB-6	ORGFREF-LB-6
<TITLE>	<u>K3100-06</u>	<u>K3100-06</u>	<u>K3100-06</u>	<u>K3100-06</u>	UAIF8	NAICR-I-B-1	ORGFREF-LB-3	UAIF8
<TITLE>	NAICR-I-A1-4	NAICR-I-A1-4	NAICR-I-A1-4	UAIF4	UAIF7	UAIF8	NAICR-I-B-1	UAIF7
<TITLE>	ORGFREF-LA1-5	UAIF3	UAIF3	UAIF3	NAICR-I-B-1	ORGFREF-LB-5	ORGFREF-LB-5	NAICR-I-B-1
<TITLE>	UAIF3	ORGFREF-LA1-5	UAIF4	NAICR-I-A1-4	ORGFREF-LB-5	UAIF7	UAIF8	ORGFREF-LB-3
<TITLE>	UAIF4	UAIF4	ORGFREF-LA1-5	ORGFREF-LA1-5	ORGFREF-LB-3	ORGFREF-LB-3	ORGFREF-LB-4	ORGFREF-LB-5
<TITLE>	ORGFREF-LA1-3	ORGFREF-LA1-3	ORGFREF-LA1-3	ORGFREF-LA1-3	ORGFREF-LB-4	ORGFREF-LB-4	UAIF7	ORGFREF-LB-4
<TITLE>	ORGFREF-LA1-1	ORGFREF-LA1-1	ORGFREF-LA1-1	ORGFREF-LA1-1	ORGFREF-LB-1	ORGFREF-LB-1	ORGFREF-LB-1	ORGFREF-LB-1
<TITLE>	ORGFREF-LA1-4	ORGFREF-LA1-4	ORGFREF-LA1-4	ORGFREF-LA1-4	ORGFREF-LB-2	ORGFREF-LB-2	ORGFREF-LB-2	ORGFREF-LB-2
<TITLE>	ORGFREF-LA1-2	ORGFREF-LA1-2	ORGFREF-LA1-2	ORGFREF-LA1-2				
<DESC>	<u>GRACE-LA1-4</u>	<u>GRACE-LA1-4</u>	<u>GRACE-LA1-4</u>	GRACE-LA1-3	<u>GRACE-LB-4</u>	GRACE-LB-3	<u>GRACE-LB-4</u>	<u>GRACE-LB-4</u>
<DESC>	GRACE-LA1-3	GRACE-LA1-3	GRACE-LA1-3	<u>GRACE-LA1-4</u>	GRACE-LB-3	<u>GRACE-LB-4</u>	GRACE-LB-3	NAICR-I-B-2
<DESC>	OKSAT-WEB-F-06	OKSAT-WEB-F-06	OKSAT-F-06	OKSAT-F-06	UAIF5	UAIF6	<u>K3100-15</u>	GRACE-LB-3
<DESC>	NAICR-I-A1-3	NAICR-I-A1-3	NAICR-I-A1-3	NAICR-I-A1-3	UAIF6	UAIF5	NAICR-I-B-2	UAIF5
<DESC>	<u>K3100-07</u>	UAIF2	UAIF1	UAIF2	NAICR-I-B-2	<u>K3100-15</u>	UAIF6	<u>K3100-15</u>
<DESC>	UAIF1	UAIF1	UAIF2	UAIF1	<u>K3100-15</u>	NAICR-I-B-2	<u>K3100-16</u>	UAIF6
<DESC>	UAIF2	<u>K3100-07</u>	<u>K3100-07</u>	<u>K3100-08</u>	NAICR-I-B-3	NAICR-I-B-3	UAIF5	<u>K3100-16</u>
<DESC>	NAICR-I-A1-2	<u>K3100-08</u>	<u>K3100-08</u>	<u>K3100-07</u>	NAICR-I-B-4	NAICR-I-B-4	NAICR-I-B-3	NAICR-I-B-3
<DESC>	<u>K3100-08</u>	NAICR-I-A1-2	NAICR-I-A1-2	NAICR-I-A1-2	<u>K3100-16</u>	<u>K3100-16</u>	NAICR-I-B-4	NAICR-I-B-4
<DESC>	NAICR-I-A1-1	NAICR-I-A1-1	NAICR-I-A1-1	NAICR-I-A1-1				

本稿では、100GBの文書データを用いた検索実行結果に対して、100GB・10GBデータに対する全実行結果に基づいたプリーングを行い、1クリック距離文書モデルと適合レベル1の仮定のもと、4章に述べた種々の尺度で評価値を算出した。評価尺度ごとにシステムを性能の高いものから順位付けた結果を表1に示す。表1の左側はサーベイ検索タスク（トピック検索タスク）、右側はターゲット検索タスクを示している。

表1のターゲット検索タスクの評価結果について、検索課題の<TITLE>のみを用いた場合と<DESC>のみを用いた場合で、リンク情報を用いた実行結果（下線）の順位の分布を比較したところ、<TITLE>のように短いクエリを用いた検索実行の方が、リンク解析が有効であることが確認される。また、同表で<TITLE>を用いた場合のサーベイ検索タスクとターゲット検索タスクにおいて、リンク情報を用いた実行結果（下線）の順位分布を比較したところ、ターゲット検索タスクの方が、リンク解析が効果的であることが確認される⁶。

6. おわりに

Web検索のための評価ワークショップに適したシステム評価手法を提案し、それらをNTCIR-3 WEBにおいて導入した。特徴的な点としては、Webに特有なユーザモデルや、リンク構造を考慮した文書モデルなどが挙げられる。詳細な分析は今後の課題である。

[謝辞]

本研究は、文部科学省科学研究費補助金特定領域研究「情報学」（課題番号13224087）及び若手研究（課題番号14780339）による。NTCIR-3 WEBのすべての参加者の協力に感謝する。また、同アドバイザー委員諸氏及び国立情報学研究所・安達淳教授からは有益な助言を得た。

[文献]

[1] K. Eguchi, K. Oyama, E. Ishida, N. Kando and K. Kuriyama: "Overview of the Web Retrieval Task at the Third NTCIR Workshop", Proc. of 3rd NTCIR Workshop on Research in Information Retrieval, Automatic Text

Summarization and Question Answering (2003).
 [2] "NTCIR-WEB", <http://research.nii.ac.jp/ntcweb/>.
 [3] 吉岡, 神門 (編): "(特集) NTCIR: 情報アクセスに関わるテキスト処理技術の評価ワークショップ", 人工知能学会誌, 第 17 卷 (2002).
 [4] D. Hawking and N. Craswell: "Overview of the TREC-2001 Web Track", Proc. of TREC-2001, pp. 61-68 (2001).
 [5] J. Kleinberg: "Authoritative sources in a hyperlinked environment", Proc. of 9th ACM SIAM Symposium on Discrete Algorithms (1998).
 [6] K. J"arvelin and J. Kek"al"ainen: "IR evaluation methods for retrieving highly relevant documents", Proc. of ACM SIGIR 2000, pp. 41-48 (2000).
 [7] R. Baeza-Yates Ed.: "Modern Information Retrieval", Addison-Wesley (1999).
 [8] E. Voorhees: "The TREC-8 Question Answering Track report", Proc. of TREC-8, pp. 77-82 (1999).

江口 浩二 Koji EGUCHI

国立情報学研究所 / 総合研究大学院大学助手 . 情報検索 , Web 情報処理及びそれらの評価に関する研究に従事 . ACM, 情報処理学会, 電子情報通信学会, 人工知能学会各正会員 .

大山 敬三 Keizo OYAMA

国立情報学研究所 / 総合研究大学院大学教授 . 構造化テキスト処理システムの研究に従事 . 電子情報通信学会, 情報処理学会, 情報メディア学会各正会員 .

石田 栄美 Emi ISHIDA

国立情報学研究所 COE 研究員 . 情報検索, テキスト自動分類の研究に従事 . 情報処理学会, 日本図書館情報学会, 三田図書館・情報学会各正会員 .

神門 典子 Noriko KANDO

国立情報学研究所 / 総合研究大学院大学助教授 . 情報検索システムの評価, 言語横断情報アクセスの研究に従事 . 情報処理学会, 言語処理学会各正会員 .

栗山 和子 Kazuko KURIYAMA

白百合女子大学文学部講師 . 情報検索システムの評価, 言語横断検索の研究に従事 . 情報処理学会, ACM, 図書館情報学会, 応用数理学会各正会員 .

⁶ 実行結果 GRACE-LA1-1 及び GRACE-LB-1 が、リンク解析を伴わないものの上位に順位付けられているが、これらは他の「GRACE」で始まる同グループの実行結果とはシステムのパラメータが異なるため、例外と言える。