

Web 検索チャレンジの課題
— NTCIR ワークショップ3の新タスク —

大山 敬三, 神門 典子, 江口 浩二, 栗山 和子

国立情報学研究所

E-mail: {oyama, kando, eguchi, kuriyama}@nii.ac.jp

第3回 NTCIR ワークショップの新タスクとして計画をしている Web 検索用テストコレクションについて、テストコレクションの構築とタスクの設定に関する方針を示し、検索課題、文書集合、正解集合のそれぞれに関して検討中のさまざまな課題について議論を行っている。検索課題については、Web 検索の特徴を考慮して精度重視型や再現率重視型などについて検討している。文書集合については予備収集実験の結果についての分析も行い、サイト数、収集文書数およびリンク数など検討に必要な基礎的なデータを提示している。正解判定および正解集合については、リンクで結合された文書群やハブの扱いについて検討を行っている。

Consideration on Web Information Retrieval Challenge
— A New Task of NTCIR Workshop-3 —

Keizo OYAMA, Noriko KANDO, Koji EGUCHI, Kazuko KURIYAMA

National Institute of Informatics

This paper describes about a test collection for web information retrieval which is planned as a new task of the 3rd NTCIR workshop. The policy of construction of the collection and design of the task is shown, and various issues are discussed concerning the search topics, the document collection, and the relevant document sets. Characteristics of web retrieval are considered in the search topics. Preliminary crawling experiments and its analysis result are shown concerning the document collection. Handling of linked document groups and hubs is discussed for the relevance judgement and the relevant document sets.

1. はじめに

World Wide Web においては、さまざまな重要な情報が提供されるようになり、既に情報流通の基盤となっている。Web 情報を対象とした情報検索 (Web 検索) 技術はその重要な要素であり、Web 検索システムの代表例であるサーチエンジンは Web 利用者になくってはならないものとなっている。

Web 文書は、従来の情報検索が扱ってきた新聞記事、特許、論文などとは異なる多くの属性をもっており、従来の情報検索技術だけでなく、Web 検索固有のアプローチを組み合わせることが重要である。これ

まで、Web 検索に対してはさまざまな技術が提案・開発されてきた。

しかし、多くの研究では、各研究者が個別に収集した文書群を対象として実験を行ったり、既存のサーチエンジンのインタフェースを通して実験を行ったりしているため、手法間の相互比較や各手法の詳細な分析を行なうことが困難であった。また、実験の規模も小規模なものが多く、実用性も含めた評価が難しくなっていた。

今後、Web 検索技術の研究を効率化し、より高度化していくためには、多くの研究者が、実用的規模で

体系的な研究開発を行えるようにすることが必要であり、共通基盤としてのテストコレクションが重要となる。

そこで、国立情報学研究所では情報検索システム評価用テストコレクション構築プロジェクトにおいて、第3回 NTCIR ワークショップの新タスクとして Web 検索を取り上げることとした。

本テストコレクションでは、Web から実際に収集した実用規模に近いデータを用意し、情報学分野の多くの研究者が一定の条件の下で共通に利用できるようにする予定である。

以下では、まず、Web 検索の特徴とそれに対応するテストコレクションの一般的な課題、海外における Web 検索用テストコレクションの状況について述べ、次に NTCIR における Web 検索タスクについて、その概要、検索課題、文書集合、正解集合のそれぞれについて議論する。

2. Web 検索の特徴とテストコレクション

本章では Web 検索の特徴を述べ、テストコレクションとして考慮すべき点を示す。Web 検索の最も一般的な応用はサーチエンジンであるので、特に断らない限りこれを念頭に置いて議論する。なお本稿では、実際のシステムや技術として確立されているか否かにかかわらず、テストコレクション構築の観点から重要と考えるものを取り上げることとする。

2. 1. Web 文書の特徴と検索技術への要求

従来の情報検索が対象としていた新聞、特許、論文などの文書と比較した場合に、Web 文書の特徴と考えられる点をいくつか挙げ、それに対応して検索技術に求められる機能を例示する。

(1) 作成目的

Web 文書の作成者は専門家、営業担当から児童・生徒までさまざまであり、機械が自動生成している場合も少なくない。目的も、広報、販売促進、趣味、技術サポート、ショッピングなど多様であり、一般・特定グループ・個人向けの情報などが混在している。

このような多様な目的の文書に対しては、文書内容や掲載サーバの性質、多数の利用者のアクセス傾向などを分析し、文書のカテゴリーを判定して検索結果の絞込みや分類する技術が望まれる。

(2) 利用目的

個々の利用者も作成者同様多様であり、目的も、広く情報を集める、特定の情報を得る、注文や申し込みをするなどさまざまである。

しかし、サーチエンジンでは一般に、利用者が与える検索条件の情報量が少ない、個別利用者に関する情報を別途入手することが困難である、といった制約がある。

これに対し、個々の利用者情報の蓄積や検索シーケンスを解析し、利用者属性や検索意図を把握して、検索要求の詳細化と情報提示方法の最適化に利用したり、また、多くの利用者の情報要求を検索ログやアクセスログから分析して適用することにより、未知の利用者の検索要求に対する精度を高める技術が望まれる。

(3) 格付け

Web では情報の信頼性、網羅性、公平性などのばらつきが大きい。そこで、参照・被参照関係を統計的に処理することにより、ハブやオーソリティといったような、各文書の性質や重要性、信頼性などを判断する指標を算出し、検索結果のランキングに利用する技術が望まれる。

また、安定性、最新性や情報の価値の時間的な変化も重要な情報である。文書の作成・変更・削除やアクセスなどの時間的な変化を分析することによりこれらを判定し、検索意図に応じてランキングに利用できる技術も有効である。

(4) 表現

文章だけでなく表、画像などが多用されており、文書サイズのばらつきも極めて大きく、用語や表記も全く制御されていないため、キーワードの統計的性質や分野ごとの概念体系を活用することが難しく、既存の情報検索技術をそのまま適用しても必ずしも効率のよい検索は期待できない。表示時の利用者の認知特性などを部分文書の重み付けなどに利用できれば精度の向上に貢献できる可能性がある。

(5) 情報の粒度

ハイパーリンクを用いることにより、一つのまとまった情報を複数に分割して記述し相互にリンクすることが可能である。逆に単独の文書だけからでは情報の

内容が把握できない文書が多い。反対に、複数の関連の薄い情報が一つの文書に混在して記述されている場合も多い。検索意図に対して最適な粒度(例えば部分文書、単文書、文書集合、サイトなど)に情報を再構成して提示できれば、利用者による検索結果の利用効率を高められると考えられる。

(6) Web 空間の動的变化

Web 空間は巨大で常に拡大しており、また、Web 文書は常に作成・変更・削除が行われており、検索対象が動的に変化するため、Web 空間の全ての文書を検索対象とすることは不可能である。このため、多くの利用者のアクセス傾向や参照関係から将来のアクセスや文書の重要性を推定して収集戦略を立てたり、リンクのアンカーテキストを用いて未収集文書を検索対象に含めることにより検索可能な空間を広げたりする技術が有効であると考えられる。

2. 2. テストコレクションの構築と制約

一般的な情報検索用テストコレクションは主に、文書集合、検索課題、正解集合の要素から構成される。しかし、Web 検索用のテストコレクション構築において前節に述べたようなさまざまな検索技術に対する評価を行えるようにするためには、これらの構成そのものから見直す必要が出てこよう。

(a) 文書集合

公平で実用的な検索性能の評価を行うためには、従来の情報検索に比べて格段に大きな文書集合が必要である。実際の Web 空間の一定時間間隔のスナップショットによって構成することが理想であるが、実際は Web ロボットにより文書を収集するため、ネットワークやサーバへの負荷、あるいは収集期間などの制約を受ける。

リンク情報も重要であるため、ミクロな参照関係と同時にマクロな統計的性質を維持するようにサブ空間を切り出すことが理想的であるが、これもかなり難しい。

また、実際のシステムでは文書の重要度や更新頻度に応じた収集戦略が検索性能に大きな影響を与えるため、この評価も重要であるが、テストコレクションの中でこれを行うことは極めて難しい。

そこで、実現可能な収集方法の中で、評価結果へ

の影響ができるだけ小さくなるものを選ぶとともに、その影響の度合いを評価することでテストコレクションの公平性や客観性を確保する努力が必要となる。

(b) 検索課題

Web 検索では検索ログや利用者情報などを収集し、検索シーケンスと組み合わせることにより検索意図を推定し、検索に反映させることが重要な課題となる。しかし、テストコレクションにおいては、検索ログなども含めてこれに対応する検索課題を用意することは困難である。

そこで、複数の利用者に対してアンケートを行うなどの代替手段により代表的な検索意図を複数設定し、これらに対応した尺度によって評価することで、多様な利用目的に対応できるようにすることが必要になる。

(c) 正解集合

同一の検索要求に対しても多様な検索意図があり得るため、正解集合も複数の尺度に対応したものをを用意することが望まれる。しかし、正解判定には多大なコストがかかるため実現が難しい。そこで、判定理由などを記録にとどめ、評価結果を事後に分析するために利用できるようにするなどの工夫をすることが重要になる。

また、正解集合を用いたシステムの評価も、単に個別正解文書の再現率や精度だけでなく、リンクで結ばれた複数文書の集約や、単一文書中の該当部分の抽出などを考慮できることが望ましい。

3. Web 検索用テストコレクションの状況

Web 検索用テストコレクションの試みとして、米国商務省国立標準・技術院(NIST: National Institute of Standards and Technology)が主催する TREC(Text Retrieval Conference)での、Web トラックが挙げられる[1]。TREC-8 Web トラックでは、大規模 Web タスクと小規模 Web タスクが実施され、それぞれが使用する文書集合は、100G バイトの VLC2 コレクションと、VLC2 から抽出された 2G バイトのサブセット(以下、WT2g コレクション)であった。以下に、TREC-8 Web トラックを中心に Web 検索用テストコレクションの状況について記す。

3. 1. 大規模 Web タスク

TREC-8 大規模 Web タスクは、後述の小規模 Web タスクほど狭く焦点が絞られておらず、個々の参加者の多様な目的のもとに実施された。

(1) 大規模 Web タスクでの文書集合

TREC-8 大規模 Web タスクでは、VLC2 (Very Large Collection, second edition) [2, 3] として知られるコレクションが用いられた。これは、1850 万ページ、100.426G バイトからなる Web のスナップショット¹であり、Internet Archive [4] が 1997 年始めに Web から収集したデータが基礎となっている。

VLC2 は、116,102 の異なるホストからのデータを含んでいる。それぞれのホストが寄与する割合の平均は、約 160 ページである。全ホスト中 24,814 ホストについては、1 ページを提供するのみである。

(2) 大規模 Web タスクの検索課題と正解判定

実際に運用されているサーチエンジンから提供された、10 万の自然言語クエリをもとに、有害なクエリあるいは有害な答を導くであろうクエリが除去され、その残りからランダムに選択された 1 万クエリが、検索課題とされた。

参加者は 1 万クエリのすべてを処理し、評価のために検索結果上位 20 件を提出した。提出を受け付けた後、検索課題が 60 件に絞り込まれるまで次の処理が繰り返された。(a)ランダムに検索課題を選択する。(b)たかだか二語の非ストップワードを含む場合、検索課題は受理/棄却の判断のため提示される。判定者には、クエリを所有していた利用者が求めていることを理解したと感じた場合、かつ、その検索課題に関する文書の適合性を判定できると感じた場合、その検索課題を受理するように要請する。

更に、答が同一となる検索課題、正解文書が 5 件に満たない検索課題のすべてが除去され、結果として 50 の検索課題が評価に用いられた。正解判定時には、正解判定者に対して、検索課題ごとにプールされた文書のテキストのみが文書長の昇順に提示さ

れた。

(3) 大規模 Web タスクでの評価

効率と有効性のトレードオフは、実際には次の五つの次元、すなわち、(a)インデクシングの速さ、(b)インデクスの大きさ、(c)検索処理の速さ、(d)検索処理の有効性、(e)処理コストにわたって発生する。大規模 Web タスクでは、これらの五つの尺度のほか、スケラビリティ、ハードウェア資源などに関しても、システム間の比較評価がなされた。

3. 2. 小規模 Web タスク

小規模 Web タスクは次の問いに答えるという目的に焦点が絞られた。(a) TREC 随時検索トラックにおける最良の手法が、Web データの WT2g コレクションに対しても同様に有効であるのか？(b) Web データにおけるリンク情報を用いることにより、ページコンテンツのみを用いるよりも、有効な検索結果を得ることができるのか？

(1) 小規模 Web タスクでの文書集合

小規模 Web タスクにおける文書集合は、以下の要件を満たすことに留意して作成された。

- TREC 随時検索トラックのコレクションに相当する大きさであること
- TREC-8 の随時検索トラックの検索課題に関連するデータを、適当な量だけ含み得ること
- 恣意性がなく、自然に定義されたサブコレクションであること
- 適度な量の閉じたハイパーリンクを含むこと。

WT2g コレクションの作成方法としては、次の発見の手法が用いられた。(a)まず、100G バイトの VLC2 コレクションを代表する、異なるホストを特定する。(b)次に、VLC2 コレクションにおいて、発見された正解文書の件数を求め、正解文書密度の降順にホストを順位付ける。(c)最後に、データが 2G バイトをわずかに越えるまで、高順位に位置するホストに由来するすべての文書を、VLC2 コレクションから抽出する。

以上のような WT2g コレクションが、コレクション内のページ間リンク情報とともに参加者に配布された。

(2) 小規模 Web タスクの検索課題と正解判定

小規模 Web タスクでは、TREC-8 随時検索トラックの検索課題が使用された。また、提出された検索結

¹ 本論文では、「スナップショット」という表現を用いているが、実際の Web ページ収集は一定時間にわたるという事実留意されたい。

果の判定に用いたツールおよび文書表示は、随時検索トラックと同様のものが用いられた。

また、小規模 Web タスクと随時検索トラックでは、同一の正解判定者により判定され、正解判定者のほとんどは検索課題の作成者であった。

正解判定においては正解文書(以下、直接的正解文書)だけでなく、正解文書へのリンクを含んだ文書(以下、間接的正解文書)の判定も実施された。

(3) 小規模 Web タスクでの評価

TREC-8 の小規模 Web タスクにおいて、標準的な TREC 評価尺度では、リンクの使用による効果が測定されなかった。

ただし、少数のリンクベース法に関してのみ、間接的正解文書を直接的正解文書と同じだけの重みを与えて評価を行なった結果、再現率あるいは上位 20 件精度に関して格段の効果を示した。

3. 3. Web 検索用テストコレクションの課題

大規模 Web タスクでは大規模なデータを扱うための技術を備えていることが参加の要件となるため、多様な参加者の協調を図るのは容易でない。その意味で小規模 Web タスクは意義深い。しかしながら、TREC-8 小規模 Web タスクにおいては、評価結果の分析から以下のように指摘されている。

- WT2g コレクションが十分な大きさであったか？
また、リンクベース法が有効に機能するに十分なリンクを含んでいたか？
- リンクベース法の有効性を評価するためには、Web で現実に発生し得る情報ニーズや、検索結果の適合性でない「検索結果の価値」に基づく評価モデルが必要であろう。

この他にも、Web 検索システムに不可欠なページ収集あるいは利用者の情報ニーズに対する充足度などに関する評価も必要と思われ、Web 検索用テストコレクションには未だ多くの困難な課題が残されている。

4. NTCIR における Web 検索タスク

4. 1. 概要：リンク構造をもった文書の検索

今回の Web 検索タスクの目的は、タグとリンク構造を持った Web 文書の検索に関する研究である。タスク内容と評価方式は、Web 検索の特徴的な観点をと

りいれる一方で、従来の成果との比較を可能にするため、検索課題の形式、評価尺度などは、従来方式も継承する。

NTCIR プロジェクトでは、伝統的実験室型実験を基本としつつ、検索対象の文書集合やその使われ方の特性を考慮した、より現実的なテストコレクション構築を目指してきた。

Web 文書は、前述のように、情報検索システムが従来、扱ってきた文書とはまったく異なる特徴がある。しかも、多くの場合 1-2 語の短い検索質問で、大量の文書群中から高速で高精度な検索が求められる。そのため、2. 節に述べたようなさまざまな情報を用いて文書のランキングを行なう手法が提案されている。

しかしながら、これらの多くは検索ログなど、文書外から得られるものである。そこで、今回は「できるところから取り組む」を基本とし、文書自体に含まれる情報として、タグとリンク構造に着目する。ただし、逆リンクについては、さらに検討が必要である。

文書集合の規模は、参加者が扱いやすい規模(10GB 程度)と、ある程度現実に近い規模(100GB 程度)を想定している。ただし、データ容量が大きいく取り回しが大変であることや、データをそのまま配布することに関する制約が多いことなどを配慮して、当面は文書データそのものを配布することはない。参加者は国立情報学研究所内のオープンラボに用意した計算機資源を用いてデータの処理を行い、その結果を持ち帰って検索実験を行う形を取る。

4. 2. サブタスク

Web 検索のサブタスクは以下のとおりである。

- A. サーベイ検索(再現率も重視)
 - A1. 検索課題検索 A2. 類書検索
- B. ターゲット検索(精度重視)
- C. 自由課題

4. 2. 1. サーベイ検索

従来の学術文書や新聞記事を対象とした Ad Hoc 型検索に相当する。固定した文書集合に対して、新たな検索課題で検索を行なう。

検索課題: A1 は、従来の書式の検索課題を用いる。A2 の類書検索では、検索課題中の<TITLE>(中

心的な概念を表わす 1-2 語)と適合文書をキーとして検索を行なう。

検索結果の提出と評価: 適合度順の検索結果の上位 1000 文書を提出し, 上位文書を集めて正解候補とし, 人間の判定者が判定する。従来の 4 段階適合判定(高度に適合, 適合, 部分的適合, 不適合)に加え, top relevant「一番よいもの」を判定する。評価は, 従来の trec_eval のほか, 重み付き平均精度など適合性のレベルや順位を考慮した尺度を使用する予定である。検索結果と正解判定は, ページ単位を基本とし, 1クリック先のページで正解文書候補プール中にあるものも判断材料に含める。

根拠パッセージ: Web ページは長さが不均質なため, 根拠となるパッセージ(文書の部分)も提出する。評価の基本はページ単位とし, 根拠パッセージ提出は必須ではないが, 補助的評価として検討する。根拠パッセージを提出しない場合は, ページ全体を根拠とするとみなす。

4. 2. 2. ターゲット検索

答えが一つみつければよいもの, ファクト型の検索で, 精度が重視される検索タスクである。検索結果の提出と評価: 検索結果は, 根拠パッセージ(必須ではない)付きで, 上位 5 件を順位付きで提出する。評価は, 以下のいずれか, あるいは, いくつかの方式で行なうことを検討している。

B1. TREC の Q&A 式。最初に検索された正解文書の順位の逆数。

B2. 有用性。正解は+1, 不正解は-1 を与える。

B3. 信頼性。正解は+1, 関連があるが間違っているものは-1, 関連がないものは 0 点。

4. 2. 3. 自由タスク

文書データを配布し, 自由に研究課題を登録し, 研究を進める。分類出力タスクは, 一例としての提案である。複数の参加者が得られた場合, ラウンドテーブルディスカッションを含め, タスクとして取り上げる。

分類出力タスク: 検索課題の<TITLE>のみで検索をし, 上位 20 件の結果をいくつかのラベル付きグループに分類して返す。たとえば, 「中田英寿」の検索結果が, 「サイト」, 「日程」, 「雑誌」, 「TV」, 「写真」, 「応援日記」など別に返すなど。

これは, 利用者が入力する非常に短い検索質問に対し, 高精度で, かつ, 利用者のナビゲーションを助けるための技術の評価の検討が目的である。評価は, 下記の参考指標を算出し, ラウンドテーブルディスカッションを通じて, 比較や検討を進める。

- ・分類がわかりやすいか(他との比較)
- ・各クラスの文書数
- ・各クラスと文書内容の適合性
- ・所望のページが見つかったか

5. 検索課題

複数の大学で Web 検索とその検索意図に関するアンケート調査を行ない, 検索課題の書式の設計に利用した。検索課題は, 従来と同様の書式とする。タスクにより, 使用項目と必須項目は異なる。以下に例を示す。

例 1

<title>中田英寿</title>
<description>中田英寿の今後の試合予定を知りたい。</description>
<narrative>適合文献は, 中田英寿の今後の試合予定を示しているもの。チケット予約とは連動してなくてよい。ファンの個人的な HP などでも具体的な日程, 場所, 時間がわかるのであれば, 正解とする。今後は, ページが作成された時点からみて「今後」。試合の印象記などは正解ではない。</narrative>

例 2

<title>エルニーニョ</title>
<description>「エルニーニョ」現象とその世界の気象への影響(海水温, 気圧, 降雨量などへの影響を含む)について説明している文書を探したい。</description>
<narrative>適合文献は, 「エルニーニョ」の影響についての情報を提供するもの。海と陸上の大気との相互作用は, エルニーニョ現象に関連するものならば, 関心がある。「エルニーニョ」は, 世界の気候に影響を及ぼすので, 特に南太平洋で重要である。</narrative>

DESCRIPTION は, もっとも基本的な記述である。中田の試合予定, ヘルシーなお菓子のレシピなど,

「①」の「②」のような形を基本とする。それに対し、TITLE はもっとも中心的な概念を表わす「①」に相当する語 1-2 語のみで、検索要求のすべての側面を表わしているとは限らない。NARRATIVE は、背景、検索目的、判定基準などの詳しい説明である。

検索課題は、分野と検索目的に関するバランスを考慮して、選定する予定である。

6. 文書集合

6. 1. 文書集合の定義と提供方法

テストコレクションでは文書集合を明確に与える必要がある。それには以下のようにいくつかの方法があり得るが、それぞれに得失がある。

- (1) 収集した文書集合の一部を取り出し、それらの url を検索文書集合として定義するとともに、検索実行に必要な文書データを提供する。この方法では、検索実行のための敷居が低く一般の研究者が使用しやすい。従来のテストコレクションと同様の方法であり、正解集合の作成やシステム評価に既知の多くの手法が使える。
- (2) 収集した文書集合の部分集合を取り出し、そこから得た url のみを文書集合として定義して提供する。検索実行のためには Web 文書を各研究者が取得する必要がある、実行のための敷居がかなり高い。検索実行と正解判定に用いた文書内容に差異が生じ、システム評価の精度が時間とともに低下するが、使用する手法は(1)と同様となる。
- (3) url のパターンを与え、それに一致する Web 文書を文書集合として定義する。検索実行のためには Web 空間を各研究者が crawling する必要がある、ごく一部の研究者しか実行できないと考えられるが、収集戦略自体の評価が可能になる可能性もある。検索実行と正解判定では文書集合自体に差異が生じ、正解集合の作成やシステム評価に未知の要素が多い。

我々にとって Web 検索タスクは初めての取り組みであるため、未知の要素が多い(3)は断念した。また、ワークショップ参加者がある程度集まらなないと正解集合の網羅性が確保できないこと、作成したテストコレクションの有効性を作成後もできるだけ長期間維持

できること、などから、(1)の方法を採用することとした。

6. 2. 予備収集実験

文書集合の作成に用いる文書は Web から実際にロボットで crawling して収集する。テストコレクションの有効性を確保するために必要な、収集範囲や文書量を検討するため、予備収集実験を 2001 年 6 月から行っている。

本稿ではその途中経過として、.ac.jp, .go.jp, .co.jp の範囲の26, 879サイトを対象として収集した文書の統計的な傾向を検討する。なお、6月29日の時点で、上記の範囲で121, 309サイトが発見されている。

図1にサイトあたりの収集文書数の上限値(F)に対する収集文書数(P)を示す。

$F=a$ におけるグラフの傾きは、サイト内の文書数が a 以上であるサイト数になる。P は F が 1000 近くなると傾きが鈍くなることから、1000 を越える文書を保有しているサイトの比率がかなり小さいことがわかる。

なお、平均文書サイズはFが10から1000に増加するに従って 6.6Kbyte から 9.5Kbyte へ徐々に増大している。また、Fが1000のときのサイトあたりの収集文書数は142.7である。

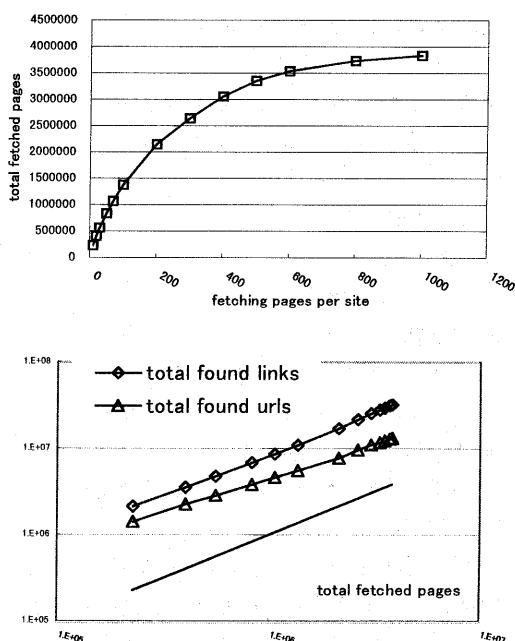


図2. 収集文書数に対するリンク数, url数

図2に収集文書数(P)に対する出リンク数(L), および発見 url 数(U)を示す. L/P は8.3程度でP(およびF)にはあまり依存しない.

一方, U/P はPとともに減少し, $P=1000$ では3.4程度となる. リンク先が収集文書内で閉じていればこれは1となるが, 実際には対応する文書が存在しない url, 1文書に複数対応する url, 収集サイト以外の url があるため1になることはない.

6. 3. ワークショップ用文書集合

ワークショップで用いる文書集合に関しては, 以下のような検討事項が残っており, 予備収集実験を継続して分析を行う必要がある.

(1) 収集範囲

- 含める方向: .ne.jp, .or.jp, .gr.jp, 地域名.jp
- 要検討: .com, .org, .to, .tv...の日本語サイト
- 要検討: .com, .org, .edu...の参照の多いサイト

(2) 収集量

- 検索用文書のリンク先を文書集合に含めるか?
→最低でも1ホップ先まで収集しておく必要あり
- 収集サイト数: 全てにするか選択するか?
- サイトごとの収集文書数: 1000 程度. 大規模サイトの分析が必要.

(3) 収集対象外のサイトへのリンクの扱い

(4) サイト内の収集戦略

- リンク解析に基づく収集の優先付けを行うか?
→収集にバイアスがかからないか?

(5) データフォーマット

- データ形式
- メタデータ
- 文字コード変換

(6) 文書集合の抽出

- 不要文書の選別・排除を行うか?
機械的に生成されたデータの文書
- 不適切な文書の選別・排除を行うか?
個人情報を含む文書, 反社会的な文書

7. 正解集合

正解判定は, 検索された文書の上位を集めて正解候補とし, それを人間の判定者が判定をする. 正解の網羅性を高めるため, サーチャによる検索結果を正解候補に追加する. 可能ならば, 複数の異なる判

定者を採用したい. 従来, 正解判定は, 検索課題を作成した本人が最適任であるといわれていたが, Web 検索は, 多数の人が, 異なる目的(あるいは同様の目的)で, 同様の検索質問を行なう可能性が高いからである.

8. おわりに

国立情報学研究所では本稿で述べたような考え方に基づいて Web 検索用テストコレクションの構築を進めている. 本稿執筆時点では検討の不十分な点も多く残されており, 実際のワークショップにおいては本稿と異なるデータや手法を採ることもあり得る.

より良いテストコレクションの構築のためにはワークショップ参加者の利用者の積極的な貢献が重要であり, 関係研究者からの要望やアドバイスなどを大いに期待している.

なお, 最新の情報についてはプロジェクトホームページ[5]を参照されたい. また, タスクの内容, 評価方式については, 随時, ML を通じて検討する予定である.

参考文献

- [1] D.Hawking, E.Voorhees, N.Craswell and P.Bailey: "Overview of the TREC-8 web track", Proceedings of the 8th Text REtrieval Conference (TREC-8) (Eds. by E.M.Voorhees and D.K.Harman) (1999).
- [2] D.Hawking, N.Craswell and P.Thistlewaite: "Overview of the {TREC-7} very large collection track", Proceedings of the 7th Text REtrieval Conference (TREC-7) (Eds. by E.M.Voorhees and D.K.Harman) (1998).
- [3] D.Hawking, N.Craswell, P.Thistlewaite and D.Harman: "Results and challenges in Web search evaluation", Proceedings of the 8th International World Wide Web Conference (WWW8) (1998).
- [4] "The internet archive", <<http://www.archive.org/>>.
- [5] NTCIR Project, <<http://research.nii.ac.jp/ntcir/>>.