特集　自然言語による情報アクセス技術

2 Web 検索の技術動向と評価手法

江口 浩二
国立情報学研究所
eguchi@nii.ac.jp

Web文書は，従来の情報検索が扱ってきた新聞記事，特許，学術論文などとは異なる特性を備えており，これまでWeb文書の特性を利用したさまざまな技術が提案・開発されてきた．また，Web検索の有効性評価という点では，実用を意識した評価が難しいとされてきたが，最近になってさまざまな試みがなされているところである．本稿では，Web文書を対象とした検索技術の最新動向を，とりわけテキスト処理手法およびリンク解析手法を中心に概観するとともに，Web検索に関する評価ワークショップを中心とした評価の取り組みについて紹介する．

Web検索の諸問題

World-Wide Webにおいて，人間の知的活動のさまざまな領域に関する情報が豊富に提供されるに伴って，Web情報アクセスシステムの代表例であるWebサーチエンジンはWeb上の情報にアクセスするための手段としてなくてはならないものとなっている．Web上の情報の単位となるのがWeb文書であり，主にこれを対象にした情報検索をWeb検索と呼ぶ．従来の情報検索が扱ってきた新聞記事，特許，論文などと異なり，Web文書には検索の観点から見ると次のような特徴がある．

- 作成者と作成目的の多様性：情報の信頼性，記述の専門性，想定読者など
- ジャンルの多様性：論文，カタログ，議事録などから個人のプロフィール，日記などまでが区別なく混在
- 表現の多様性：タグを用いたレイアウトや構造化，フレーム，表や画像などの視覚効果
- 情報の粒度：複数文書から構成される情報，複数情報が記載された文書（図-1）
- リンクによる参照：参照・被参照の情報の活用が可能
- 変化の速度：文書の追加，削除，更新が常時発生

また，Web検索において効果的な検索を難しくしている要因として，特に重要な点を以下に指摘する．

■ Web情報空間の規模

Web情報空間の規模については，年々増加の一途をたどっており，総務省平成15年度版情報通信白書によれば，Webコンテンツの総データ量は.JPドメインだけでも平成14年末の時点で10,150ギガバイトと推計されている．これに伴って，Web検索の研究開発も，全世界のWebを対象とした汎用的な検索を目指す方向性と，特定組織のWebサイトに限定，もしくはジャンルやドメインを限定するといった方向性に分かれるようになってきた．

大規模なWeb文書データに対応した汎用的な検索を実現するためには，並列化による処理の高速化，あるいは分散化による管理コストの軽減などが必要になるだけでなく，Webページの価値を判定する仕組みがより重要となる．1つの解決策が，後ほど詳述するリンク構造の解析に基づくトピック・ディスティレーション技術である（「トピック・ディスティレーション」の章参照）．また，ジャンルやドメインに特化した検索として興味深い研究事例を後述する（「ジャンル・ドメインに特化した検索」の章参照）．

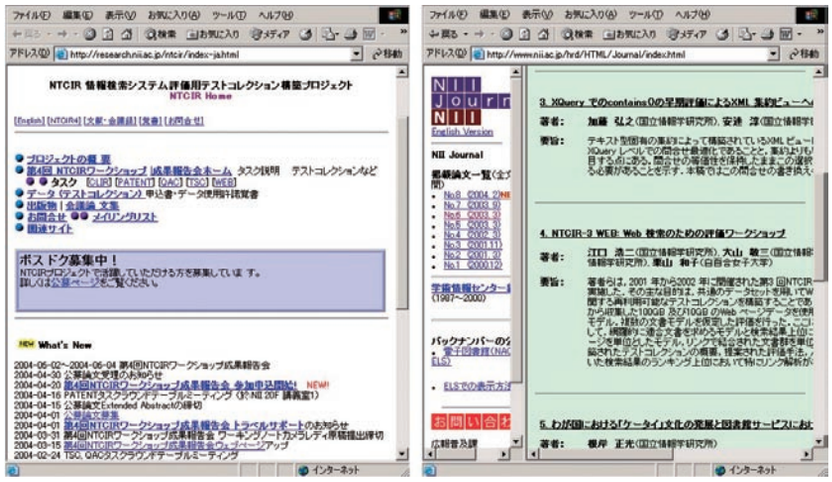

図-1　複数文書から構成される情報（左）と複数情報が記載された文書（右）の例

■検索に関する情報量の不足

メリーランド大学のJansenらは，実際に広く利用されているWebサーチエンジンのログに基づく分析結果として，ユーザがWebサーチエンジンに与えるクエリの長さは平均して2単語程度であり，大半のユーザは検索結果の1ページ目（上位10件程度）までしか閲覧しないと報告している．このように不足しがちな検索に関する情報を補間する手段として，本稿ではユーザに関する情報やユーザのコンテクストを活用した技術と，ユーザの置かれた環境を考慮した技術について触れる（「個人や環境に適応した検索」の章参照）．

■情報ニーズの多様性

IBMのBroderはWeb検索における情報ニーズ（あるいはタスク）を次の3つのカテゴリーに分類しており，後述するTREC WebトラックやNTCIR WEBタスクに対しても方向性を与えてきた．

- 情報指向（informational）：特定のトピックに関する1件もしくは複数件のWebページを獲得することを要求する．
- ナビゲーション指向（navigational）：ある特定のWebサイト（またはある対象物の代表的なページ）に到達することを要求する．
- トランザクション指向（transactional）：インタラクションを伴うようなWebサイト（オンライン・ショッピング，Webが仲介する種々のサービス，特定のデータベース等）に到達することを要求する．

現在の多くのWebサーチエンジンは情報指向もしくはナビゲーション指向の要求に対応しており，トランザクション指向の要求には間接的に答えるのみである．上記のような情報ニーズの種類はクエリとして明らかに示されないことも多い．前述の少ない情報しか与えないクエリからその背後に潜むユーザの情報ニーズを理解し，それに即した結果を提示することが，Webサーチエンジンの課題の1つである．

ところで，Web検索の有効性評価はWeb検索技術の発展になくてはならないものであるが，諸々の技術的な理由により容易ではない．後述する評価ワークショップおよびテストコレクションはそのような問題に対する有望な解決手段である．ただし，それらはWebに適したものでなければならない．本稿では，Web検索の評価手法に関する動向についても誌面を割きたい（「Web検索の評価の取り組み」の章参照）．

トピック・ディスティレーション

大規模なWeb情報空間に対して，クエリが与える限定された情報から適切な検索を実現するには，Webページの価値を判断し，価値の高いWebページを優先して提示するような仕組みが必要となってくる．1つのアプローチとして，ハイパーリンクの構造を解析することでWeb文書の価値を判定する技術について，これまで盛んに研究開発がなされてきた．Web検索のためのリンク構造解析技術として代表的なものに，特定のトピックのページに関するランキング手法であるHITSと，トピックに依存しないランキング手法であるPageRankが挙げられる．

IBMのKleinberg（現在コーネル大学）らが提案したHITSは，特定のトピックに関する情報の豊富さを表すオーソリティ（authority）と，オーソリティへのハイパーリンクの豊富さを表すハブ（hub）という概念を導入し，良いオーソリティは多くの良いハブからリンクされ，良いハブは多くの重要なオーソリティをリンクするという関係を求めることで，検索結果の質を改善している．また，この過程における計算はトピック・ディスティレー

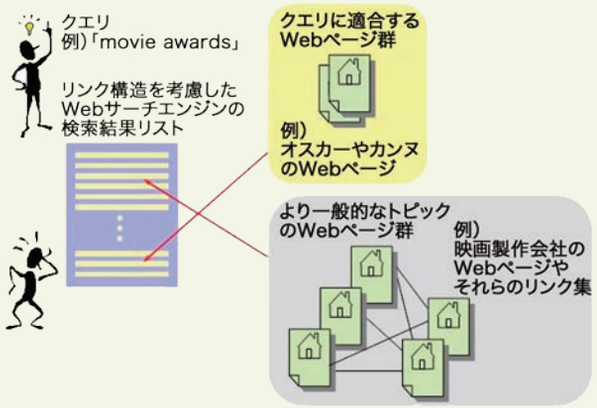

図 -2　トピック・ドリフトの例

ションと呼ばれている．

スタンフォード大学の Page（現在 Google）らが提案した PageRank は，多くの良質な Web ページからリンクされている Web ページは良質な Web ページであるという仮説に基づくもので，トピックに依存せずに計算される．PageRank は Web サーチエンジン Google[☆1] におけるランキングに取り入れられていることで知られている．

DEC の Bharat（現在 Google）らによれば，これらの手法ではトピック・ドリフト問題が起こり得る．トピック・ドリフト問題とは，たとえば，一般的な語を含むクエリが与えられ，その一般的な語によって検索された Web ページがリンク集等により密に結合されていた場合などで，ユーザが本来求めていたトピックとは関連性の低いはずの Web ページが検索結果の上位にランキングされる問題である（図 -2）．なお，HITS や PageRank を改善する手法や，これらとは異なる観点からリンク構造を解析する手法も提案されており，上に示したトピック・ドリフト問題が部分的に改善されているものの，検討の余地が残されていると思われる．トピック・ドリフト問題については本稿において後ほど触れたい．

ジャンル・ドメインに特化した検索

Web の規模の拡大に伴って，ドメインやジャンルに特化した Web サーチエンジンの研究開発が行われてきた．代表的なものとして，情報系分野の学術論文を検索するための ResearchIndex（CiteSeer）[☆2] が知られている．

また，最近になって，ネットワーク上に公開された意見や評価，評判，感情などの主観的な情報を活用するための研究が行われるようになり，今年 3 月には当該研究領域に関する国際シンポジウム[☆3] が AAAI 主催で開かれた．これらの研究は，ユーザが意思決定の材料として他者の主観に関する情報を参照することを目的としたもので，そういったジャンルに特化した検索とも位置づけられよう．製品等に関する評価情報を収集するとともに，それらがポジティブな見方を示しているかネガティブであるか，またその程度がどれくらいであるかといったことを自動的に判別する研究がなされつつある．国内でも関連する研究が行われており，たとえば，NEC の立石らは商品名とそれに関してある観点から見た評価を示す表現を，あらかじめ用意した評価表現辞書をもとに Web ページから抽出することで，Web 上に存在する評価情報の効果的な収集を試みている．

Web 上に存在する主観情報は，個人の Web ページ，電子掲示板，専用サイト[☆4]，Web 上の日記などとして提供されていることが多く，個人による動的な更新やコミュニケーションに適した Blog（Weblog）[☆5] と呼ばれる発信形態で提供されることも少なくない．その意味で，主観情報の活用技術は Blog に関する研究とも密接に関連すると思われる．この種の研究の今後の展開が期待されるところである．

個人や環境に適応した検索

ユーザから与えられた限定的な情報に基づいて効果的な検索を実現する方法に，ユーザに関する情報やユーザのコンテクストを活用した検索技術，ユーザの置かれた環境に応じて適切な Web ページを提示する技術などが挙げられる（図 -3）．本章では，個人の検索履歴などを活用した個人化検索と，地理情報に基づく情報アクセスについて説明する．

個人化検索

Web 検索を高度化するための 1 つの方向として，個人化検索（personalized search）が挙げられる．従来の Web 検索では，多くのユーザのために適合であると計算された Web ページは各々のユーザにとっても適合であることを仮定していた．それに対して，個人適応型検索では，各ユーザのインタラクションのコンテクストにおいて適合性が決定される[1)]．その結果，同じクエリを

☆1 http://www.google.com/
☆2 http://citeseer.ist.psu.edu/cis/
☆3 http://www.clairvoyancecorp.com/research/workshops/AAAI-EAAT-2004/home.html
☆4 たとえば，http://www.epinions.com/
☆5 たとえば，http://www.cocolog-nifty.com/

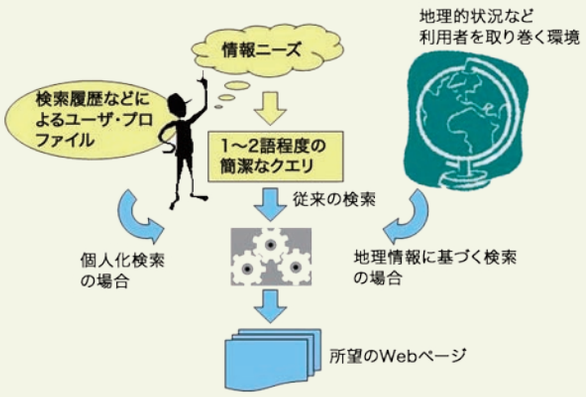

図 -3　個人化検索と地理情報に基づく検索

入力しても，検索結果がユーザにとって異なることになる．所望の情報を獲得する時間と手間の軽減が期待される．個人化検索の実現方法としては，ユーザがプロファイル（興味のあるトピックの集合）を設定する方法と，ユーザの設定を伴わずに検索履歴等を利用してプロファイルを自動生成する方法がある．個人化検索は一部のWebサーチエンジンにおいても実現されている．たとえば，Googleは当該サービス☆6を試験的に提供しており，ユーザが設定したプロファイル等に基づいて，最適な検索結果を提示することを試みている．

課題としては，ユーザの興味が時間とともに変化することへの対処などが挙げられる．このような現象はWeb検索に限らず，従来の情報検索システムの利用行動においても経験することであり，カリフォルニア大学のBatesが提案した検索行動のモデル（berrypicking model）においても考慮されている．また，筆者らはアドホックな検索におけるクエリ拡張手法においてユーザの興味変化に対応する手法を提案している[2]．個人化検索においても以上に述べた観点で検討を行う余地がある．

地理情報に基づく情報アクセス

地理情報は我々の日常生活と密着しており，Web情報へのアクセス手段として実用的な側面を持つ．外出先において携帯端末を用いてユーザの物理的位置から距離的に近い店舗や施設などに関するWebページを優先して検索するという利用状況が典型的である．ほかにも，Webサーチエンジンの検索結果を地理的な配置に基づいて分類したり，地図にWebページを配置したりするといった活用が検討されている．一部の研究者により基礎的な技術が研究されつつあり，重要となる要素技術として，Webページ中の住所表記から経緯度を特定する技術などが挙げられる[3), 4)]．

Web検索の評価の取り組み

Web検索の有効性評価は，諸々の技術的な理由により，容易ではない[5)]．評価ワークショップおよびテストコレクションはそのような問題に対する有望な解決手段である．ただし，それらはWebに適したものでなければならない．評価ワークショップとは，多くの研究グループが共通のデータセット（テストコレクション）を構築し，それを用いてタスク遂行し，成果を相互比較するものであり，Web検索に焦点を当てたものとして，TREC Webトラック☆7とNTCIR WEBタスク☆8が知られている．

TREC Webトラックでは，.GOVドメインのWeb文書からなる18ギガバイトのデータセット，非営利団体のInternet Archiveが収集したデータを元にした100ギガバイトのデータセットおよびそのサブセットが構築され，評価に用いられてきた．タスク設計としては，所与のトピックに適合したWebページを検索する状況，所与の名称を用いて該当する特定のWebページ（あるいは特定のWebサイトのトップページ）を検索する状況などが想定された．また，トピック・ディスティレーション技術の評価を想定し，所与の比較的広い意味を持つトピックについて，最も関連するWebサイトのトップページ群を検索するという設定でも評価が行われた．

NTCIR WEBタスクでは，.JPドメインからHTMLファイルおよびプレーン・テキストファイルを収集することで，約100ギガバイトのWeb文書データセット（NW100G-01）が構築された（**表-1**，**表-2**）．また，Web検索手法の評価を目的として，Webに特徴的なハイパーリンク構造などの特性を勘案し，評価モデルの構築が行われた．筆者らの評価分析の結果，ユーザが簡潔で曖昧性を含むクエリを使用し，上位10件程度の検索結果のみを閲覧することを前提とした評価モデル（すなわちWebサーチエンジンの典型的な利用状況）においては，リンク構造を考慮した検索手法が有効であり，それ以外のモデルではリンク構造が考慮されていたとしても有意な効果は現れないことが確認されている[6)]（**図-4**）．これは「トピック・ディスティレーショ

☆6 http://labs.google.com/personalized/
☆7 http://trec.nist.gov/
☆8 http://research.nii.ac.jp/ntcweb/

(a)	収集元サイト数	97,561
(b)	サイト内ページ数の上限	1,300
(c)	収集ページ数	11,038,720
(d)	検索対象ページ数 *	15,364,404
(e)	(c) から出て行くリンクの数	78,175,556
(f)	(c) から出て (d) へ入るリンクの数	64,365,554

* (c) の各 Web ページから出て行くリンク先においてインターネット上で存在が確認された Web ページの数

表 -1　NW100G-01 の特徴

言語	比率 *
日本語	90%
英語	8.3%
中国語（簡体字）	0.05%
韓国語	0.03%
中国語（繁体字）	0.02%
西ヨーロッパ言語	0.01%
その他の言語	0.01%
テキスト内容を含まない	0.78%
特定不能	0.02%

* 各 Web ページの「content-type」フィールドに示された文字コードセットに基づく概算見積りである.

表 -2　NW100G-01 における言語ごとの Web ページの比率

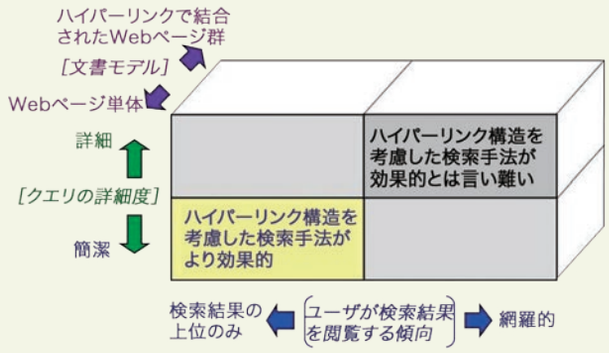

図 -4　ハイパーリンク構造を考慮した検索手法の振る舞い

ン」の章で述べたリンク構造解析におけるトピック・ドリフトの現象を裏付ける観察結果と見なすことができ，さらなる分析が期待される．

ユーザの実際の利用行動や満足度を考慮することも，Web 検索の評価において重要な観点である．TREC ではインタラクティブ・トラック☆9 において上記の観点で検討が行われてきた．NTCIR WEB タスクにおいても検索結果の閲覧時間の計測に基づく評価が検討されている．

新たな試みとして，TREC ではテラバイト級の Web 文書データセットを用いたテラバイト・トラックが，2004 年から開始された．ここでは，検索の有効性だけでなく効率性が特に強調される．これとは別途に Web トラックとして，ある特定の企業等組織が提供する Web ページのみを集中的に収集し，Web 文書データセットを構築することが進められている．また，タスク設計としては，ユーザの情報ニーズの種類（たとえば，情報指向なのかナビゲーション指向なのか）が所与でない状況で適切な検索を実現することに焦点を当てて議論されているところである．これは「Web 検索の諸問題」の章で述べた Web 検索における情報ニーズの多様性の問題に焦点を当てたものといえる．

また，NTCIR WEB タスクでは，Web 検索に関連する多面的な技術にも焦点を当ててきた．たとえば，クラスタリング等の技術を用いて検索結果を分類提示する技術，Web ページに自然言語で記述された住所等の情報を元にして地理的状況を反映したアクセス技術，音声で入力されたクエリを用いて Web 文書を検索する技術についてである．Web 検索手法の研究を行う上で，より Web の現状に即した文書データセットが求められるところであるが，早稲田大学の山名らの研究グループは全世界的規模の Web ページを分散して収集することを試みており，今後の展開が大いに期待される．

参考文献

1) Pitkow, J., Schütze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E. and Breuel, T.: Personalized Search, Communications of the ACM , Vol.45, No.9, pp.50-55 (2002).
2) Eguchi, K., Ito, H., Kumamoto, A. and Kanata, Y.: Adaptive Query Expansion Based on Clustering Search Results, 情報処理学会論文誌, Vol.40, No.5, pp.2439-2449 (May 1999).
3) 横路誠司，高橋克巳，三浦信幸，島　健一: 位置指向の情報の収集，構造化および検索手法，情報処理学会論文誌，Vol.41, No.7, pp.1987-1998 (July 2000).
4) 相良　毅，有川正俊，坂内正夫: ジオリファレンス情報を用いた空間情報抽出システム，情報処理学会論文誌：データベース，Vol.41, No.SIG6 (TOD 7), pp.69-80 (2000).
5) 神門典子，安達　淳他: 評価ワークショップによるテキスト処理研究：第3回NTCIRワークショップを例として，人工知能学会誌，Vol.17, No.3, pp. 312-319 (2002).
6) Eguchi, K., Oyama, K., Ishida, E., Kando, N. and Kuriyama, K.: Evaluation Methods for Web Retrieval Tasks Considering Hyperlink Structure, IEICE Transactions on Information and Systems, Vol.E86-D, No.9, pp.1804-1813 (2003).

（平成16年5月10日受付）

☆9　2003 年からは Web Track と一体となって運営されている．