テキスト処理に基づくWeb情報アクセス支援
―検索から分類・追跡へ―
Text Processing in Support of Web Information Access

江口 浩二[†]
Koji Eguchi

1. まえがき

World-Wide Web において，人間の知的活動の様々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 文書である．本稿では，**2.** で Web 文書にアクセスする上で問題となり得る諸々の問題について述べ，**3.** で Web 文書にアクセスするための基本的な手段である Web 検索技術について説明する．また，Web 情報アクセスをより高度に実現するためのテキスト処理として，**4.** で Web 文書分類技術，**5.** でトピック検出・追跡技術を取り上げる．さらに，**6.** で，Web 情報アクセスの評価の試みとして，TREC，NTCIR，TDT の動向を紹介する．

2. Web 情報アクセスの諸問題

本稿では特に断らない限り，Web 文書は主に HTML 形式で記述された Web ページを指すことにする．Web 文書にアクセスするための基本となる手段が検索である．Web 文書に対する効果的な検索を難しくしている要因として特に重要なものに，Web 情報空間の規模，検索に関する情報量の不足，情報ニーズの多様性が挙げられる [1]．それぞれについて **2.1**，**2.2**，**2.3** で説明する．

2.1 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，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 [2]．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．それぞれの研究動向を **3.1**，**3.2** で紹介する．

2.2 検索に関する情報量の不足

Jansen ら [3] 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，ユーザに関する情報やユーザの情報アクセスのコンテクストを活用した個人化検索と，ユーザの位置する地理的状況を考慮した位置指向検索などがある．本稿では，**3.3** において個人化検索について述べる．位置指向検索については文献 [4, 5] を参考にされたい．

[†]国立情報学研究所

2.3 情報ニーズの多様性

Broder [6] は Web 検索における情報ニーズ（あるいはタスク）を次の三つのカテゴリに分類しており，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一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．

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一つである．

現在の多くの Web サーチエンジンは，インターネットに接続されたデータベース等の内容自体を直接の検索対象としていない（そのような Web 上の情報は Deep Web などと呼ばれる）．従って，Web サーチエンジンの多く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と言える．

トランザクション指向の検索を実現するには，(1) プロービング処理等によりデータベースに含まれる内容を特徴づける技術，(2) クエリで表現された情報ニーズに適合したデータベースを選択する技術，(3) データベースから取り出した情報を統合する技術（検索結果の統合やラッパー処理）などが必要になってくる．より詳細については文献 [7] などを参照されたい．

3. Web 検索技術

3.1 大規模かつ一般的な検索

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．一つの解決策が，HITS [8] や PageRank [9] に代表されるリンク構造の解析に基づく手法である．これらの手法ではトピック・ドリフト問題 [10] が起こり得るため，その解決が研究課題の一つとなっている．トピック・ドリフト問題とは，例えば，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．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．当該研究課題は文献 [11] などに詳しいため，本稿では詳細を割愛する．

3.2 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）[‡]が知られている．ドメインやジャンルに特化した Web サーチエンジンを実現するための技術として，一般の Web サーチエンジンに対し特定のドメインやジャンルに関連する語を投入して Web ページ群を獲得する方法 [12]，特定のドメインやジャンルに Web ページ集合を分類する技術 [13, 14]，あるいは特定のドメインやジャンルについて集中的に Web ページを収集する技術 [15] などが必要になってくる．

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 [16] 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を自動的に判別する研究がなされつつある [17, 18]．国内でも関連する研究が行われており，例えば，立石ら [19] 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，予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，Web 上の日記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）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 [20] 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．

3.3 個人化検索

Web 検索を高度化するための一つの方向として，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 [21]．その結果，同じクエリを入力しても，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，また，他者のプロファイルを利用する方法がある．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．例えば，Google は当該サービス[§]を試験的に提供しており，ユーザが自ら設定したプロファイルに基づいて，最適な検索結果を提示することを試みている．また，Eurekster[¶]は，ユーザの検索履歴からプロファイルを自動生成するほか，他者のプロファイル（例えば，同じトピックに関心を持つグループ）を利用可能にするソーシャルネットワーキング機能が付与されている．

4. Web 情報アクセスのための分類技術

Web 文書を対象にした分類技術として，様々な観点から研究が行われているが，本稿では特に Web 情報アクセスを目的としたものに焦点を当てる．本節では分類操作の対象が検索結果文書群の場合と大規模 Web 文書集合の場合に分けて，それぞれ **4.1** と **4.2** で説明する．

4.1 検索結果文書群の分類

利用者が入力する検索質問に対する検索結果を高精度に分類し，利用者のブラウジングにおける認知的負荷を軽減することを目的とした技術について概要を述べる．このような場合の分類の観点としては，Web 文書のトピック [22, 23, 24]，ジャンルやタイプ [13, 25]，URL 文字列等による発信元に関する情報 [26] のほか，ハイパーリンクの接続関係によるコミュニティ，Web 文書に記述された地物の地理的配置などがある．

なかでもトピックによって検索結果 Web ページ群の分類を実現するためのアプローチとしては種々のものが提案されており，(1) 文書内の内容語の分布に基づく文書クラスタリング [27, 23]，(2) 上記に加えてハイパーリンクの接続関係をも考慮した文書クラスタリング [28]，(3) 文書内に共通に使用される文字列に基づく文書クラスタリング [22]，(4) 検索語と文書内の特定のタグに含まれる語の共起関係に基づく文書クラスタリング（Vivisimo[‖]，[29]），(5) 既知の主題体系あるいは分類体系に基づくテキスト分類（text categorization）[30]，これらのうちの複数を組み合わせたものなどが提案されている．なお，これらの分類処理においては，Web 文書そのものでなく，Web サーチエンジンの検索結果リストにおいて付与される文書の抜粋が処理の対象にされることもある．階層的分類と非階層的分類，排他的分類と非排他的分類といった点もアプローチによって異なる．

以上のようなタスクは一般に，有効性の評価が容易でないことが多い．また，種々のアプローチの比較評価についてこれまで十分に行われてこなかった．後述する NTCIR WEB タスクでは，検索結果 Web ページ群をトピックによって分類する技術の評価方法について考察を行っている [24]．

[‡] http://citeseer.ist.psu.edu/cis/

[§] http://labs.google.com/personalized/

[¶] http://www.eurekster.com/

[‖] http://vivisimo.com/

4.2 大規模 Web 文書集合の分類

大規模な Web ページ集合を自動分類する研究が行われている．これらは例えば，**3.2** に紹介したジャンルやトピック，ドメインに特化した検索を実現するためのインデクシングの要素技術として用いられ得る．

本稿では，不均質な Web 情報を扱う上で特に重要となるジャンルへの自動分類について取り上げる．ここでいうジャンルは，類似した形式で記述された文書のグループであり，トピックとは直交するとする，Finn ら [31] の定義に従うものとする．いくつかの研究事例では同様の概念を文書タイプと呼んでいる [32, 33]．Web ページのジャンル体系については種々の提案がなされているところであり [34, 32, 35, 33, 25]，例えば，個人のページ，政府や組織などの公共のページ，リンク集，掲示板などの入力を伴うページなどのジャンルが挙げられる．

ジャンル自動分類のアプローチとしては，(1)Web 文書に特徴的なタグやリンク，URL 文字列等を利用する方法 [32, 25] と，(2) 代名詞の数や受動態の頻度といったテキストの機能的特徴を利用する方法 [13]，(3) テキスト中の内容語の分布に基づく方法 [36] に大別される．

また，Web 文書に限ったものではないが，文書が著者の意見を表現するものであるか，あるいは事実を報告するものであるかというジャンルの捉え方があり，その観点から文書を自動分類する研究がなされている [31, 37]．さらには，**3.2** で言及した通り，著者の意見がポジティブであるかネガティブであるかという観点で，文書を自動分類する研究がなされている [17, 18]．誰もが気軽に情報を発信できる Web の特徴を鑑みて，この種の研究は今後さらに活発になるものと考える．

5. トピック検出・追跡技術

これまでに述べてきた話題はいずれも Web をスナップショットとして捉えてきたが，現実の Web は時間とともにその構造や内容が変化している．その点で，時系列文書から新しいトピック（ここでは特にイベント）を検出し，それに関連する文書を追跡する技術が重要と考える．この種の研究はこれまで主にニュース記事やニュース放送に対して行われてきた [38]．以下の問題に分けることができる．(1) 文書からトピックに関してまとまりのある単位を切り出す問題，(2) 時系列文書から新しいトピックを検出する問題，(3) ある時点の文書群をトピックによってクラスタリングする問題，(4) ある特定のトピックに関連する文書を追跡する問題，(5) 任意の二つの文書が同じトピックについて記述されたものであるかどうかを判別する問題．また，時系列文書におけるトピックのトレンド分析についても研究がなされているところである [39]．本稿では詳細には触れないが，文献 [38, 40] が参考になる．

TDT（Topic Detection and Tracking）[**]は，トピック検出・追跡技術についての評価を目的とした評価ワークショップ（**6.** 参照）である．TDT 1997 のパイロットスタディを経て，TDT 1998 から評価ワークショップとして参加グループを集めるようになり，現在，TDT 2004 が進行しているところである．ニュース記事やニュース放送の英語と中国語（マンダリン）を使用したテキストデータと音声データが収集され，前記 (1)〜(5) の評価に用いられている．

[**] http://www.nist.gov/speech/tests/tdt/

6. Web 情報アクセスの評価の取り組み

Web 情報アクセス技術の有効性評価は，諸々の技術的な理由により，容易でない [41]．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[††]と NTCIR WEB タスク[‡‡]が知られている．それぞれについて **6.1**，**6.2** で概要を述べる．なお，NTCIR WEB タスクでは Web 検索だけでなく，**4.1** で述べたような，Web 文書の分類についても評価を試みている．また，トピック検索・追跡に焦点を当てたものとして TDT がある．これは現時点では必ずしも Web を対象にしたものではないものの，**5.** で取り上げた話題に関連する．**5.** において概略を説明したので本節では省略する．

6.1 TREC における試み

TREC Web トラックは，一年周期で実施されている TREC の一環として，1999 年に開始され，現在，通算六回目が TREC 2004 の一環として進行している．

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（**2.3** で述べた情報指向に対応），所与の名称を用いて該当する特定の Web ページまたは特定の Web サイトのトップページを検索する状況（**2.3** のナビゲーション指向に対応）などが想定された．また，リンク構造解析に基づく検索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．

新たな試みとして，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 年に開始されている．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，World Wide Web Consortium（W3C）の Web サイトにおける Web ページを集中的に収集し，Web 文書データセットを構築することが進められている．テラバイト・トラックと Web トラックの関係は，**2.1** で言及した通り，大規模かつ汎用的な検索を目指す方向と特定的な検索を目指す方向に，研究開発が分化しつつあることの表れと言えよう．また，タスク設計としては，ユーザの情報ニーズの種類（例えば，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．

[††] http://trec.nist.gov/, http://es.csiro.au/TRECWeb/
[‡‡] http://research.nii.ac.jp/ntcweb/

これは **2.3** 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言える．

6.2 NTCIR における試み

NTCIR WEB タスクは，一年半の周期で行われている NTCIR ワークショップの一環として，2001 年から 2002 年にかけてと，2003 年から 2004 年にかけて実施された．現在，第 5 回 NTCIR ワークショップの一環として三回目の WEB タスクを実施すべく準備がなされつつある．

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．タスク設計としては，所与のトピックに適合した Web ページを検索する課題（**2.3** で述べた情報指向に対応），施設等の代表的な Web ページを検索する課題（ナビゲーション指向に対応）などが設定された．

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効果的とは言えない結果が確認されている [42]．これは **3.1** で言及したリンク構造解析におけるトピック・ドリフト現象を裏付ける観察結果と見なすことができ，更なる分析が期待される．

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．例えば，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 [24]，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 [43]，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 [44] についてである．

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・トラックにおいて上記の観点で検討が行われており（2003 年からインタラクティブ・トラックは Web トラックと一体となって運営されている），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 [45]．

Web 情報アクセス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求められるところであるが，山名ら [46] 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．

6.3 おわりに

不均質な Web 情報を対象とした情報アクセス手法について，主にテキスト処理に関係するいくつかの話題を中心に概観した．紙面の都合上，詳細を省略したが，本稿に引用した文献を参照されたい．本稿で紹介した話題のいくつかは，情報検索，自然言語処理，データマイニング，機械学習などの研究領域が関わるものであり，これらの連携がより重要となってくると思われる．

参考文献

[1] 江口：“Web 検索の技術動向と評価手法”, 情報処理, **45**, 6, pp. 569–573 (2004).

[2] 総務省：“平成 15 年度版情報通信白書” (2003).

[3] B. J. Jansen, A. Spink and T. Saracevic: “Real life, real users, and real needs: A study and analysis of user queries on the web”, Information Processing and Management, **36**, pp. 207–227 (2000).

[4] 横路, 高橋, 三浦, 島：“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, **41**, 7, pp. 1987–1998 (2000).

[5] 相良, 有川, 坂内：“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, **41**, SIG6(TOD 7), pp. 69–80 (2000).

[6] A. Broder: “A taxonomy of web search”, SIGIR Forum, **36**, 2, pp. 3–10 (2002).

[7] J. Callan, F. Crestani and M. Sanderson Eds.: “Distributed Multimedia Information Retrieval”, LNCS 2924, Springer-Verlag (2004).

[8] J. Kleinberg: “Authoritative sources in a hyperlinked environment”, Proceedings of the 9th ACM SIAM Symposium on Discrete Algorithms, San Francisco, California, USA (1998).

[9] S. Brin and L. Page: “The anatomy of a large-scale hypertexual web search engine”, Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia (1998).

[10] K. Bharat and M. R. Henzinger: “Improved algorithms for topic distillation in hyperlinked environments”, Proceedings of the 21st Annual International ACM SIGIR Conference, Melbourne, Australia, pp. 104–111 (1998).

[11] S. Chakrabarti: “Mining the Web: Discovering Knowledge from Hypertext Data”, Morgan Kaufmann Publishers (2003).

[12] S. Lawrence, C. L. Giles and K. Bollacker: “Digital libraries and autonomous citation indexing”, IEEE Computer, **32**, 6, pp. 67–71 (1999).

[13] J. Karlgren, I. Bretan, J. Dewe, A. Hallberg and N. Wolkert: “Iterative information retrieval using fast clustering and usage-specific genres”, Proceedings of the 8th DELOS Workshop on User Interfaces in Digital Libraries, Stockholm, Sweden, pp. 85–92 (1998).

[14] J. M. Pierre: “Practical issues for automated categorization of web sites”, ECDL 2000 Workshop on the Semantic Web, Lisbon, Portugal (2000).

[15] S. Chakrabarti, M. van den Berg and B. Dom: “Focused crawling: A new approach to topic-specific web resource discovery”, Proceedings of the 8th International World Wide Web Conference, Toronto, Canada (1999).

[16] “AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications”, Stanford, California, USA (2004).

[17] P. D. Turney: “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews”, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), Philadelphia, Pennsylvania, USA, pp. 417–424 (2002).

[18] B. Pang, L. Lee and S. Vaithyanathan: “Thumbs up? Sentiment classification using machine learning techniques”, Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002), Philadelphia, Pennsylvania, USA, pp. 79–86 (2002).

[19] 立石, 石黒, 福島：“インターネットからの評判情報検索”, 情処研報, NL144-11, pp. 75–82 (2001).

[20] 武田, 大向：“Weblog の現在と展望：セマンティック Web およびソーシャルネットワーキングの基盤として”, 情報処理, **45**, 6, pp. 586–593 (2004).

[21] J. Pitkow, H. Schutze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar and T. Breuel: “Personalized search”, Communications of the ACM, **45**, 9 (2002).

[22] O. Zamir and O. Etzioni: “Web document clustering: A feasibility demonstration”, Proceedings of the 21st Annual International ACM SIGIR Conference, Melbourne, Australia, pp. 46–54 (1998).

[23] 江口, 伊藤, 隈元, 金田：“漸次的に拡張されたクエリを用いた適応的文書クラスタリング法”, 信学論 (D-I), **J82-D-I**, 1, pp. 140–149 (1999).

[24] K. Eguchi: “Overview of the topical classification task at NTCIR-4 WEB”, Proceedings of the 4th NTCIR Workshop, Tokyo, Japan (2004).

[25] K.-J. Lee: “Document genre classification for user interface of web search engine”, IEICE Transactions on Information and Systems, **E87-D**, 7, pp. 1982–1986 (2004).

[26] 風間, 原田, 佐藤：“サーチエンジンの検索結果のマルチレベルグルーピング”, 第 2 回インターネットテクノロジーワークショップ論文集 (1999).

[27] D. R. Cutting, D. Karger, J. O. Pedersen and J. W. Tukey: “Scatter/Gather: A cluster-based approach to browsing large document collections”, Proceedings of the 15th Annual International ACM SIGIR Conference, Copenhagen, Denmark, pp. 318–329 (1992).

[28] Y. Wang and M. Kitsuregawa: “Use link-based clustering to improve web search results”, Proceedings of the 2nd International Conference on Web Information Systems Engineering (WISE'01), Kyoto, Japan, pp. 115–124 (2001).

[29] 成田, 太田, 片山, 石川：“階層的クラスタリングを利用したメタ検索エンジンの提案— METAL —”, 情処研報, DBS128-50, pp. 375–382 (2002).

[30] H. Chen and S. Dumais: “Bringing order to the web: Automatically categorizing search results”, Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'00), The Hague, The Netherlands, pp. 145–152 (2000).

[31] A. Finn, N. Kushmerick and B. Smyth: “Genre Classification and Domain Transfer for Information Filtering”, LNCS 2291, Springer-Verlag (2002).

[32] 松田, 福島：“文書タイプ分類による問題解決向き WWW 検索システムの開発と評価”, 情処研報, 99-FI-53, pp. 9–22 (1999).

[33] 久野, 石田, 安形, 上田：“Web ページのタイプ判定法”, 2000 年度日本図書館情報学会春季研究大会発表要綱, pp. 55–58 (2000).

[34] J. Dewe, I. Bretan and J. Karlgren: “Assembling a balanced corpus from the internet”, Proceedings of the 11th Nordic Computational Linguistics Conference, Copenhagen, Denmark (1998).

[35] S. W. Haas and E. S. Grams: “Readers, authors, and page structure: A discussion of four questions arising from a content analysis of web pages”, Journal of the American Society for Information Science, **51**, 2, pp. 181–192 (2000).

[36] Y.-B. Lee and S. H. Myaeng: “Text genre classification with genre-revealing and subject-revealing features”, Proceedings of the 25th Annual International ACM SIGIR Conference, Tampere, Finland, pp. 145–150 (2002).

[37] Y. Seki, K. Eguchi and N. Kando: “Analysis of multi-document viewpoint summarization using multi-dimensional genres”, AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Stanford, California, USA, pp. 150–153 (2004).

[38] J. Allan Ed.: “Topic Detection and Tracking: Event-based Information Organization”, Kluwer Academic Publishers (2002).

[39] J. Kleinberg: “Bursty and hierarchical structure in streams”, Proceedings of the 8th ACM SIGKDD International Conference, Edmonton, Canada, pp. 91–101 (2002).

[40] 高間：“Web 情報ストリーム”, 情報処理, **44**, 7, pp. 720–725 (2003).

[41] 安達, 神門, 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, **17**, 3, pp. 312–319 (2002).

[42] K. Eguchi, K. Oyama, E. Ishida, N. Kando and K. Kuriyama: “Evaluation methods for web retrieval tasks considering hyperlink structure”, IEICE Transactions on Information and Systems, **E86-D**, 9, pp. 1804–1813 (2003).

[43] M. Arikawa, T. Sagara, K. Noaki and H. Fujita: “Preliminary workshop on evaluation of geographic information retrieval systems for web documents”, Proceedings of the 4th NTCIR Workshop, Tokyo, Japan (2004).

[44] A. Fujii and K. Itou: “Evaluating speech-driven IR in the NTCIR-3 Web Retrieval Task”, Proceedings of the 3rd NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering (to appear).

[45] T. Ohtsuka, K. Eguchi and H. Yamana: “An evaluation method of web search engines based on users' sense”, Proceedings of the 4th NTCIR Workshop, Tokyo, Japan (2004).

[46] 山名：“Web データの新しい利用法の開拓を目指して”, 情報研報, DBS133-13, pp. 107–110 (2004).