解　説

# サイト・アウトライニング
## ―インターネットからの情報収集と可視化技術―

日本アイ・ビー・エム（株）東京基礎研究所
武田 浩一　takedasu@jp.ibm.com
野美山 浩　nomiyama@jp.ibm.com

### インターネットからの情報収集と可視化

インターネット上に存在するWebページ数の増加は驚異的である．昨年7月のCyveillance社調査結果では，ユニーク・ページ数で21億ページ，1日700万ページの割合で増加中と報告されている☆1．ということは2001年3月頃には40億ページになっていたはずで，NEC北米研究所のSteve Lawrenceらが報告して[1]話題になった1997年末の3.2億ページの頃から，WWWは実に約3年で10倍以上に成長したことになる．また，昨年7月の同時期にBrightPlanet社が発表したホワイトペーパー[2]では，データベースなどから動的に生成されるWebページ数は5,500億ページにものぼると主張しており，彼らはこれをdeep Webと名付けている．

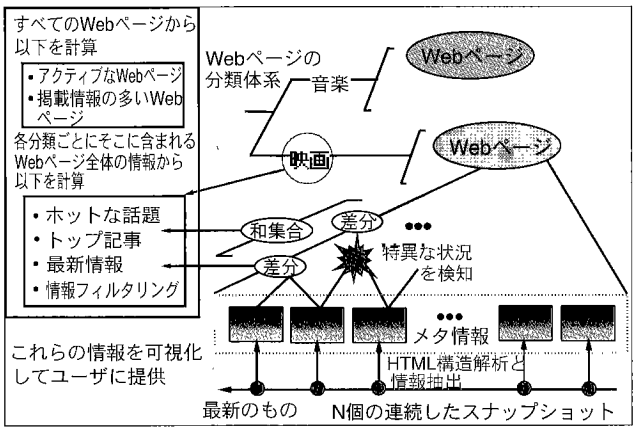
このような状況の中で，個人がインターネット，特にWWWから情報を収集するためにはどのような手段が考えられるであろうか？　必要な情報がキーワードで表現できるなら，インターネット検索エンジンによって候補となるWebページが検索できる．ただし，単純なキーワードによる検索や高頻度のキーワードによる検索では，検索結果が数万件以上になったり，検索結果に無関係なWebページが多数含まれていたりして，求める情報の発見は容易ではない．またポータル・サイトのように1個所のサイトで多くのユーザが興味を持つ情報やサービスを提供する形態が広く普及しているが，人手による情報の編集や分類・嗜好情報等の付与には限界があるため，専門的知識やパーソナライズ可能な情報の提供の場として過度な期待はできない．

個々のWebサイトで非同期に随時更新・掲示される情報に注目したPointCast社は，インターネットで最新情報を配信するプッシュ技術を1996年にPointCast Networkとして製品化し一世を風靡した．1997年には100万人以上の視聴者を獲得したが，ダイヤルアップ利用者へのアピールが不十分だったためか2000年にはサービスを停止してしまい，現在ではInfoGate社のパーソナル情報配信ツールとしてその名残りをとどめている．最新情報の配信は，現在では電子メールやメールマガジンなどのサービスが中心となっている．また，個人が特定のWebサイトの最新情報を知りたい場合には，いわゆるWeb自動巡回ソフトという製品が利用できる．これらの製品では，各サイトの更新の有無，差分情報の検出，音声・画像などのマルチメディア情報の保存といった機能が提供されている．

このようにインターネットから情報を収集するためには，目的に応じた多様な機能が必要である．具体的には，WWWの広大な情報の海から以下のような多種な情報を収集し，これを特定の尺度によって順序づける（ランキング）技術が中心となる．

1) 最新の情報
2) 特定の話題やキーワードに関連する情報

☆1 http://www.cyveillance.com/web/us/newsroom/releases/2000/2000-07-10.htm




ユーザが指定するWebページを定期的にモニタし，そのスナップショットから最新情報などを求めてダイジェストする．

図-1　サイト・アウトライニングの仕組み

3) 特定の型（動画や音楽）を持つ情報
4) 特定のパターンを持つ情報（求人情報や株価など）

また大量の情報をそのままで提示しても理解できない場合には，これを俯瞰できるような情報可視化（information visualization）技術[4]が要求される．さらに，最近のインターネット・ユーザの年齢層の広がり[5]や，携帯端末などの出力機器の多様性を考慮した情報の提示手法が求められている．ここには，テキスト情報の要約や言い換えといった自然言語処理も含まれている．

## サイト・アウトライニング

Webページで提供される情報の大きな特徴の1つは，その内容が随時更新されることである．したがって，あるWebページのある時点でのスナップショットを見ただけでは，そのページのどこがどれぐらい新しいのかを知ることができない．AT&TのDouglisらは，バージョン管理の手法を利用して，更新前後のWebページの差分を検出し，それを同じページ上で可視化する手法を提案した[3]．この差分検出による最新情報の自動収集と可視化は，Tierra Highlights2という製品にも採用され，現在では，DiffWeb☆2などの更新検知サービスでも見かけることができるし，昨年は米国証券取引委員会（SEC）がインターネット上での詐欺行為を同技術によってモニタしようとしたことが報道されて話題になったこともある．日本でも，フレッシュアイ☆3が，過去30日内に更新されたWebページのみを対象として独自の最新情報検索サービスを提供しているし，ANATAGONOMY☆4のような情報可視化とパーソナライゼーションを駆使した最新情報提供のフリーソフトウェアもある．

単なる差分検出であればHTMLファイルの記述上の更新部分が分かるだけだが，以下のようにその時点のHTMLファイルに情報抽出を先に適用することで，要素レベルの差分検出を行うことが可能になる．

1) そのHTMLファイルのテキスト構造の分析によって，見出し，アンカー，パラグラフ，画像ファイル，音声ファイル，といった要素を特定する．
2) 特にテキスト情報からは自然言語処理によってそこに含まれる単語，句などを抽出する．
3) これらの情報をそのWebページの更新日時およびURL等の情報とともに，特定の時点でのスナップショットのメタ情報として保存する．

1つのWebページに対して，このようなスナップショットのメタ情報を連続して対応づけることにより，要素単位の更新日時，掲載期間，掲載頻度などが2次的なメタ情報として計算できる．さらに単語や句レベルの情報を利用して，そのWebページに高頻度で掲載される話題や最近のホットな話題といった情報を提供することもできる．これにより，大量に発生するインターネットの新規情報をさまざまな観点からダイジェストして，前章で述べたような多目的な情報提供が実現できる．筆者らは，このようなアプローチを「サイト・アウトライニング」と呼び，研究開発を行ってきた．これまでにmySiteOutlinerという名称でコンシューマ版ソフトウェアとして製品化されており，昨年よりWindows/ME搭載のAptiva/ThinkPadに同梱されている☆5．

サイト・アウトライニングでは，図-1に概要を示したように，以下の4つのステップでWWWのテキスト情報

☆2 http://www.diffweb.com/
☆3 http://www.fresheye.co.jp/
☆4 http://www.labs.nec.co.jp/Freesoft/anatagonomy/index.html
☆5 30日間の体験版が以下のサイトから入手可能である：http://www.ibm.com/jp/pc/personal/msol/download/

```
<anchor>
  <DC:title>ビジネス・ユーザ　-あなたのビジネスに最強の
e-businessをご提供します。</DC:title>
  <url>http://www.jp.ibm.com/business.shtml </url>
   <kwds>
<kwd><word> e-business</word><class> ACT </class></kwd>
   <kwd><word>ビジネス</word> <class> ACT</class></kwd>
   <kwd><word>ユーザ </word><class> HUM </class></kwd>
   <kwd><word> 最強　</word><class> ABS </class></kwd>
   <kwd><word> 提供　</word><class> ACT </class></kwd>
  </kwds>
</anchor>
```

見出しとURL以外に，その見出しに含まれるキーワードとその意味分類を記述している（ACTは行為，HUMは人を示す）．

図-2　見出しつきリンクのメタ情報

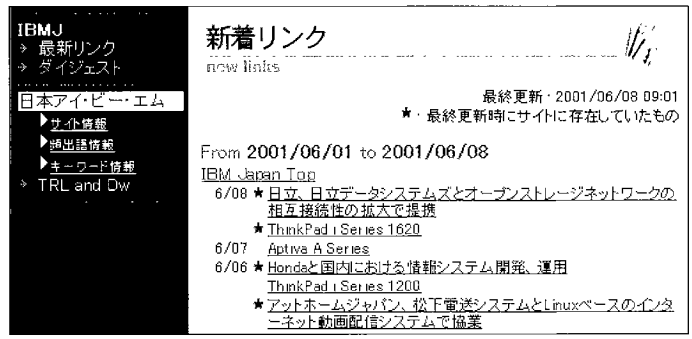
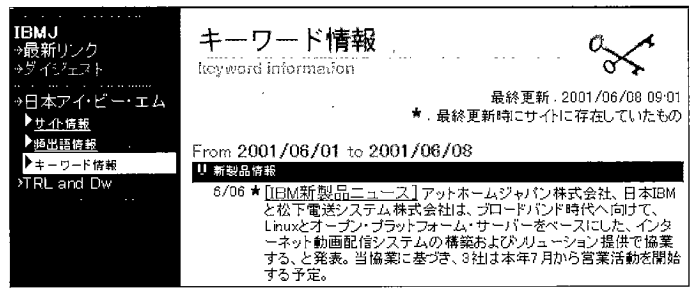



図-3　日本アイ・ビー・エム関連Webページの最新情報の表示




黄色の部分がマッチした文字列を示す

図-4　文字列の指定による最新情報からの情報フィルタリング

を中心とした情報収集・提供を行う．

[ステップ1]　ユーザが興味を持つWebページのURL集合（たとえば，Webブラウザの「お気に入り」や「ブックマーク」にあるURLなど）をもとにWebページの巡回（クローリング）を行い，スナップショットからメタ情報を獲得する．

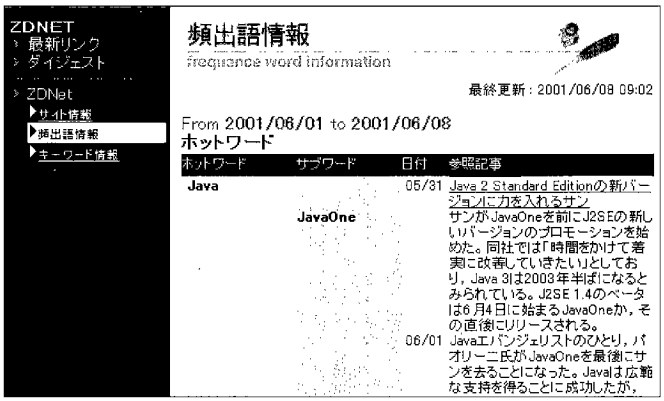
URL集合はサイト・タクソノミーと呼ぶ木状の構造に体系化でき，分野別の情報提示が可能である．あるWebページから獲得される情報とは，必ずしもそのページのスナップショットに含まれる情報だけに限定する必要はない．たとえば，フレームを使って記述されているWebページであれば，フレームの各構成要素のWebページまでを対象とする方が実用的である．一般には，あるURLで指定されるWebページに対するメタ情報は，そのWebページからk個以内のリンクによって到達可能なWebページのメタ情報の集合和に拡張することが可能である．我々が「Webサイト」として認識している情報源は，このようなWebページ集合に対応するメタ情報として定義できよう．アノテーションのついたリンク（アンカー）や本文のパラグラフに含まれるテキスト情報からは，そこに含まれる単語や句を抽出し，人名・組織名といった意味分類とともに保存する（図-2に1つのリンクのメタ情報を例示している）．メタ情報の記述にはXMLを利用している．ニュース記事の記述形式としてはNewsML☆6やRSS（Rich Site Summary）☆7などが利用され始めた．各Webサイトがこのような形式での情報提供を行うことで，より正確でかつ豊富なダイジェスト情報の生成が可能となる．

[ステップ2]　個々のWebページについて，最新のメタ情報と，前回得られたメタ情報の差分を求め，最新情報を計算する．ここで動的な情報源としてのWebページの特性が，最新リンク・パラグラフ数，更新頻度，最近の話題といった2次的なメタ情報の形で算出される．この計算は，一定期間（たとえば1週間）や過去の一定スナップショット数といった複数の基準に基づいて行うことができる．これにより，最近特にアクティブなWebページやある時期に大量の情報が掲載されたWebページなどを求めることも可能である．

[ステップ3]　メタ情報に基づいて，最新情報，トップ・ニュース，ユーザ指定のキーワードに基づくフィルター情報，ホット・トピックといったさまざまなダイジェスト情報を生成する．

図-3から図-6にこれらのダイジェスト情報の例を示す．図-3は日本アイ・ビー・エム関連Webページの今週の新着リンクを掲載日ごとに表示している．赤印のついたリンクは最新のスナップショットに掲載されているが，それ以外のリンクはすでに削除されたことを示している．図-4は，新着リンクと新着テキスト（パラグラフ）のうちで，特定の文字列を含むものをフィルタリングした結果を表示している．たとえば，「新製品情報」という見出しでは，「開発」，「新製品」，「開始」といった文字列を含むリンクやテキストを選択する．ユー

☆6 http://www.newsml.org/
☆7 http://www.xml.com/pub/rg/16．RSS1.0以降はRDF Site Summaryとして知られている．




ZDNETニュース（ZDNN, http://www.zdnet.co.jp/news/）から抽出されたホットな話題（"Java"）の例．Javaというキーワードを含む記事が右側に掲載日時順に並べられている．中央のサブワードとはJava関連記事の中分類に相当する．

図-5　ホットな話題の抽出例

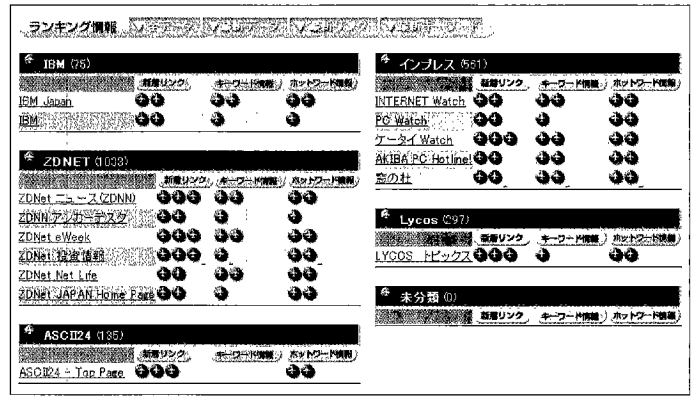
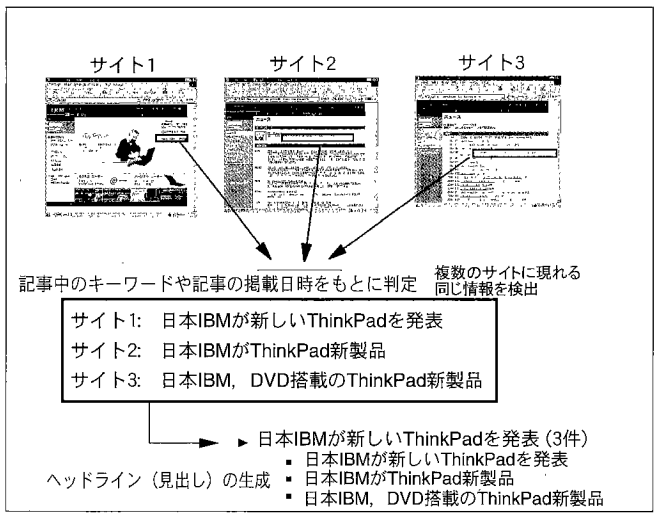



図-7　サイト・アウトライニングのトップ画面例




一定期間内に内容語が共通して含まれる類似記事集合を求める．次にそれらの記事を掲載する相異なるWebページの数をスコアとし，スコアの多い記事集合から順に表示する．

図-6　最新記事からのトップ記事の作成

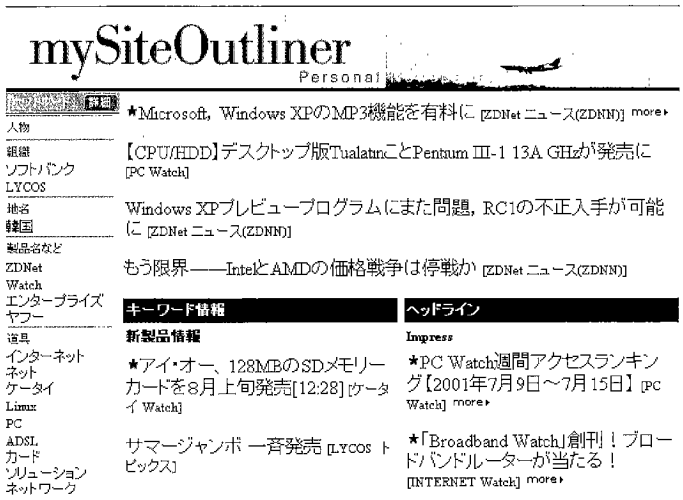
ザは，キーワード・エディタと呼ばれるツールを用いて，このような文字列のAND/ORで表現されるフィルタリング条件式を定義できる．個々の条件式にはラベルをつけることができ，これが見出しとして使われる．図-5は，ホットな話題の抽出例を示している．ホットな話題とは，今週の出現頻度が平均的な出現頻度を大きく上回る単語を抽出し，それを含むすべての記事（リンクやテキスト）を掲載時刻順に表示したものである．図-6は，最新リンクの中からその日のトップ記事生成を説明したものである．ここでのトップ記事とは，一定の期間内に掲載され，見出し内容の類似性が高く，ほぼ同一と考えられるリンク集合のうち，それらのリンクを掲載するWebページ数が閾値を超えるものである．

［ステップ4］　ダイジェスト情報から，その日あるいはその週の注目部分を決定し，ナビゲーションを容易にする．

図-7は，このようなダイジェスト情報の量を一覧するための画面を示している．画面では，各カテゴリに含まれるWebページから集計された新着リンク数，フィルター記事数，ホットな話題の記事数をアイコンの数で10未満，100未満，100以上の3段階に表示している．ダイジェスト情報ごとに上位のWebページのランキングも分かるようになっている．

Windows/2000が稼動するPC（Pentium III-667MHz）上のプロトタイプによる実験では，社内外の約600Webページを登録し，1日1回のクローリングを行ったところ，平均約25分で上記のステップが実行できた．1週間あたりの新着リンク・テキスト数は約4,000件であった．所要時間のほとんどは単一スレッドでWebページの内容をネットワークから取得するために費やされており，更新があったWebページからのメタ情報の抽出は毎時6,000ページ程度，1週間単位のダイジェスト情報の生成に関しては毎時3,000ページ程度を処理できるものと推定している．毎日更新されるWebページの比率は約30%であった．この比率が維持され，十分高速なクローリング環境が提供されるとすれば，同プロトタイプは毎日10万ページ以上のWebページにサイト・アウトライニングを適用できるものと考えている．

神戸大学の馬らは，新着記事の持つ3つの特徴量（新鮮度，流行度，緊急度）によって，各記事を重み付ける手法を提案しており[6]，このような手法を用いれば，サイト・アウトライニングによって得られた最新情報を重要度順にランキングすることも可能である．複数の特徴量とユーザの動作（ある記事をクリックする，など）をもとに重み付けを学習することで，柔軟性のあるランキングによる高度なパーソナライゼーションも実現できよう．




画面中央にトップ記事，画面下部左にはフィルタリング記事，画面下部右にはトップ記事以外の最新重要記事を配置している．また画面左側にはカテゴリー別のホットな話題が表示されている．

図-8　XSLTを用いたパーソナル新聞の生成

## 個人向けポータル・サイトの構築

サイト・アウトライニングのような技術が普及することで，個人向けのポータル・サイトの構築がきわめて身近なものになる．ユーザには，大別して次の3つの自由度がある．

- 情報を収集するWebページの組織化（ディレクトリなど）
- そこから得られる情報をダイジェストする観点
- 情報の編集とレイアウトの指定

これに以下のような主にサーバ型のサービスによって提供されている機能を組合わせることで本格的な情報収集のカスタマイズが可能になる．

1) 与えられたWebページに関連するWebページや情報の推薦機能（Alexa☆8など）
2) テキスト中の任意の文字列に対するリンクの対応づけ（flyswat☆9など）
3) Webページの任意の部分をクリッピングする機能（octopus☆10など）
4) 特定の条件に合致する最新情報を即時にアラートする機能（メールによる携帯端末への通知なども含む）

検索エンジンやWeb自動巡回ソフトを利用することで情報収集の対象となる新たなWebページのURL追加は容易になる．一定期間Webページをサイト・アウトライニングによってモニタしていると，各Webページを

a) 最新情報を定期的に掲載する動的な情報源
b) ほとんど更新のないアーカイブ的な情報源
c) 一時的に存在しただけの非永続的な情報源

という3種類に大別することができる．サイト・アウトライニングでは（a）の情報源を中心に最新情報の収集を行い，（b）の情報源はリンク集として管理するとよい．また（c）の情報源は，そのスナップショットをローカルにコピーしておくことで情報が突然失われることに対応しておけばよい．さらに（b）の情報源は，他のWebページからのリンクによる参照度を利用したランキング[7]によって，その相対的な重要度を評価することもできる．（a）のように更新回数の多いWebページは，1週間あるいは1カ月間の新着情報をまとめて仮想的な週刊誌あるいは月刊誌のようにダイジェストすることも有効である．これにより，ユーザはWebページの更新頻度とは独立に情報を閲覧することができる．

サイト・アウトライニングで生成されるダイジェスト情報はXMLで記述しておくことで，それをXSL（eXtensible Stylesheet Language）やXSLT☆11などのマッピング記述言語によって自在にHTML（あるいは別のXML文書）に変換することができる．特に複数のダイジェスト情報を1画面に配置して個人ポータル・サイトのトップページを作成したい場合には，このような柔軟性が必要となる．ANATAGONOMYや読売COLiNS☆12でのパーソナル新聞といった情報可視化手法も同様にして扱

☆8 http://www.alexa.com/
☆9 http://www.flyswat.com/
☆10 http://www.octopus.com/home.htm
☆11 http://www.w3.org/TR/xslt．XML文書からXML文書への変換を記述でき，XSLの一部として使用できる．
☆12 http://pnews.cplaza.ne.jp/

えることになる．既存ページの一部にダイジェスト情報を組み込みたい場合にも，この手法が適用できる．図-8には，筆者らが試作しているパーソナル新聞の例を示している．ここでは，個人別のサイト・タクソノミーと記事のフィルタリング情報から生成されるダイジェスト情報をXMLで表現し，これをXSLTでパーソナル新聞に変換している．

## 今後のインターネット情報収集技術

インターネットの技術動向を考慮すると，今後のインターネット情報収集技術は以下のような展開が期待できる．

### 1) 常時接続とブロードバンド化

自宅のPCを大容量のネットワークに常時接続することで，最新情報の即時通知やマルチメディア・コンテンツの情報収集への需要が高まるものと考えられる．このことは，特にマルチメディア・コンテンツにテキスト情報を人手であるいは自動的に付与するアノテーション技術や，メタ情報の記述方式の普及を促進し，マルチメディア情報の検索や分類の高機能化をもたらすであろう．また，常時接続されたPCをサーバとして，モバイル機器との情報交換を行う機能や，メールを介したアド・ホックな情報収集コマンドの実行とその結果の送付といった機能への要求が高まるであろう．特に，自動巡回機能を持つWebブラウザAirWeb☆13で採用されているようなエージェント技術は，常時接続環境下での個人の情報収集・活用能力を飛躍的に高めることができ，かつ公開されているエージェント開発キットによる粒度の細かな情報収集エージェントが広く共有される可能性がある．

### 2) テキスト・マイニング技術[8]との融合

Webページはテキスト情報だけでなく，表形式や定型データの宝庫でもある．すでにWhizBang! Labs社☆14などで製品化されているように，Webクローリングに情報抽出技術を含めることで，パターン化された情報の収集と分析といった機能が追加できる．これにより，インターネットにおける誹謗・中傷記事のモニタや製品価格，求人情報，企業業績情報だけの収集といったソリューション指向のシステムが構築できる．

### 3) ナレッジマネジメントと企業ポータル

個人の情報収集・活用のノウハウ（情報源，情報フィルタリング用プロファイル，Webページ分類体系）は共有可能である．特にWebページの分類体系は，Open Directory Project☆15のように大量のボランティアが協力して265万サイトにものぼる巨大な共有リソースに成長したものもあれば，About.com☆16のように，さらに専門的な知識をWebページとともに体系化することで，高度な情報サービスを提供するものも存在する．URL以外の他者のWebページのコンテンツ自体は著作権の問題もあって一般に情報サービスの対象にはできないが，企業内のWebページの情報を社員に提供する企業ポータルは，ナレッジマネジメントのITインフラとして成長が見込まれる．また，最近注目を集めているP2P（peer-to-peer）方式によるクライアント間の直接の情報開示・交換も，企業内であれば導入が容易である．Webサイトだけではなく，個々のクライアントの情報共有ディレクトリなどを情報源と考えてサイト・アウトライニング手法を適用することで，こういった新しい環境にも対応することができるであろう．


参考文献

1) Lawrence, S. and Giles, C. L.: Searching the World Wide Web, Science, Vol.280, No.5360, pp.98-100 (1998).
2) BrightPlanet.com: The Deep Web: Surfacing Hidden Value, White Paper, http://www.completeplanet.com/Tutorials/DeepWeb/index.asp (2000).
3) Douglis, F., Ball, T., Chen, Y.-F., and Koutsofios, E.: WebGUIDE: Querying and Navigating Changes in Web Repositories, Proc. of 5th Int'l World-Wide Web Conf., pp.1335-1344 (1996).
4) 武田浩一, 野美山浩: テキスト情報の可視化を利用した情報検索, 情報処理, Vol.41, No.4, pp.343-350 (Apr. 2000).
5) ITry Project: http://www.ibm.com/jp/pc/itry/, 日本アイ・ビー・エム (株) (2001).
6) 馬　強, 角谷和俊, 田中克巳: 放送型情報配信システムのための時系列特性を考慮した情報フィルタリング, 情報処理学会論文誌：データベースVol.41, No.SIG6 (TOD7), pp.46-57 (2000).
7) Kleinberg, J.: Authoritative Sources in a Hyper-linked Environment, Proc. of ACM-SIAM Symp. on Discrete Algorithms, pp.668-677 (Jan. 1998).
8) 那須川哲哉, 河野浩之, 有村博紀: テキストマイニング基盤技術, 人工知能学会誌, Vol.16, No.2, pp.201-211 (2001).


(平成13年6月13日受付)

☆13 http://www.airclub.org/．情報処理振興事業協会（IPA）の「未踏ソフトウェア創造事業」にも採択されている．
☆14 http://www.whizbanglabs.com/
☆15 http://dmoz.org/
☆16 http://about.com/