100億規模のWebページ収集・分析への挑戦

村岡　洋一 *1　山名　早人 *1
松井くにお *2　橋本三奈子 *3
赤羽　匡子 *3　萩原　純一 *4

*1 早稲田大学理工学術院
*2 米国富士通研究所
*3 富士通（株）
*4 アクセラテクノロジ（株）

● A ● B ● S ● T ● R ● A ● C ● T ●

Webページ数は，2006年11月時点で537億ページと推測されている[1]．我々は，2004年1月～2006年7月の間に，全世界の5,548万台のWebサーバからテキストのみを対象に収集を行い，ユニークなWebページ数として約144.5億ページを収集した．また，収集済みWebページに対して，トップレベルドメイン分布，記述言語分布，Webサーバの地理的位置の解析，バックリンク解析やPageRank計算を進め，Web空間の現状分析を行った．さらに，Webページの解析がビジネスに利用可能であることを示すために，企業のWebサイト上の活動を可視化するe企業調査プロトタイプを構築し，企業の特徴，戦略，評判などの抽出を行った．

分散収集型Webクローラ

Webページを収集するためのプログラムを一般的にWebクローラあるいはWebロボットと呼んでいる．インターネット上から高速にWebページを収集するためには，複数のPCを用いて並列に収集を行わなければならない．我々はこのWebクローラを合計70台のPCで動作させた．また，さまざまな種類のバックボーンを用いるため70台のPCを都内5拠点に分散設置した．

開発した分散収集型Webクローラでは，インターネット上に分散設置されたロボットセットと呼ばれるWebクローラをインテグレーションサーバが集中管理する（**図-1**）．インテグレーションサーバは，分散されたロボットセット間においてWebページを重複して収集しないための制御を行う．

インテグレーションサーバは収集開始時に投入された起点URLのリストを，収集を担当するロボットセットごとに振り分けて配信する．担当分の起点URLのリストを受け取ったロボットセット側では，起点URLからリンクを辿りながらWebページを収集する．リンク先が自身の担当する範囲内ならば自身で収集を行い，範囲外であればそのリンク情報をインテグレーションサーバに送る．これらのリンク情報は，インテグレーションサーバで再び担当するロボットセットごとに振り分けられ，

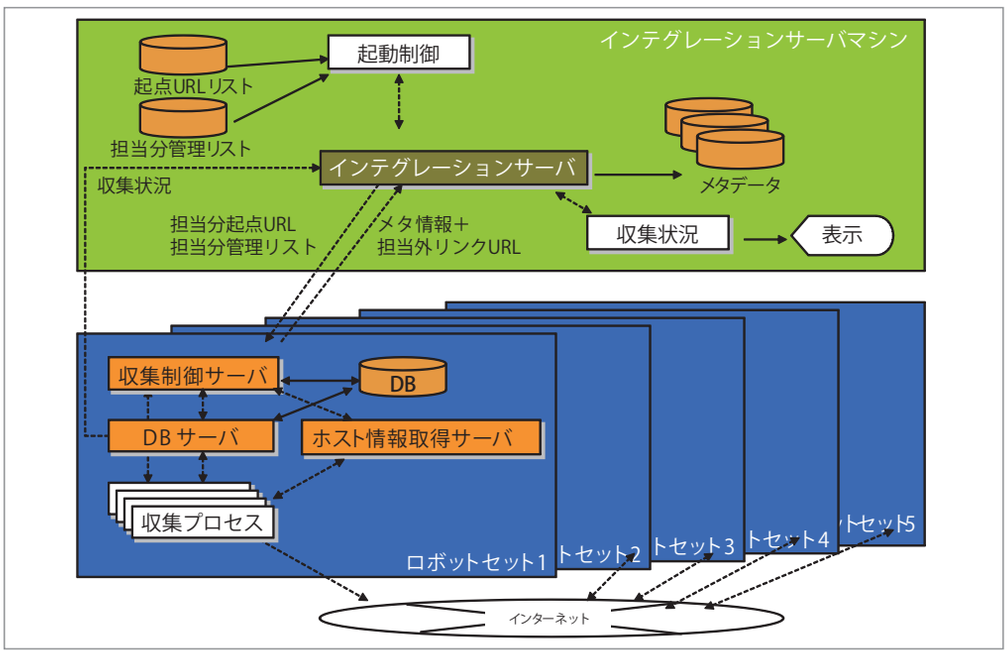

図-1　分散収集型Webクローラ

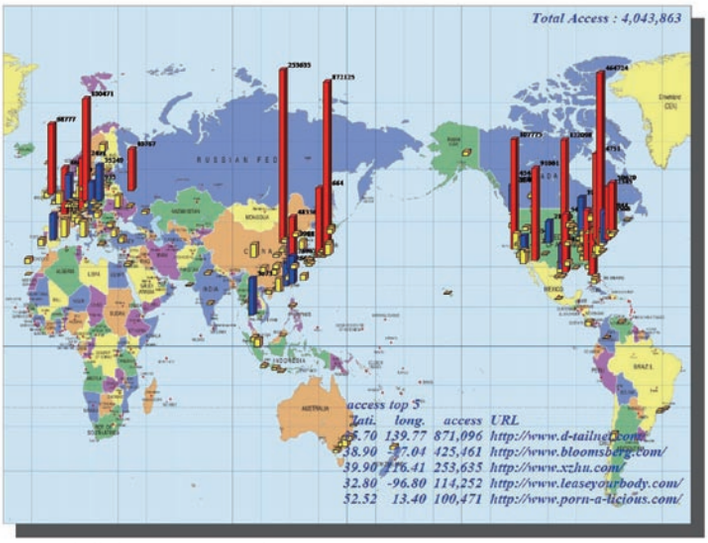

図 -2　収集先リアルタイム表示

機能	仕様
サスペンド機能	同一サイトに対して一定時間継続して収集した場合，一定時間収集をサスペンド.
Keep-Alive 機能	Keep-Alive が生きている間は連続して収集し，切られた場合は一定時間収集をサスペンド.
CGI ページからのホップ数制限機能	CGI ページからのリンクについては，収集ホップ数を指定値に制限.
IP アドレスによる収集制限機能	1 つの IP アドレスで複数のサブドメインを構築している Web サイトに対して，同時に複数の収集プロセスが集中してアクセスを行わないように収集プロセス数の上限を設定.
robots.txt の googlebot 互換機能	robots.txt のクローラ排除規約に対して googlebot が独自に拡張を行った正規表現による記述方法に対応.
収集間隔の動的算出機能	実際のネットワーク状況（回線速度および通信品質）から収集間隔を自動的に調整.

表 -1　負荷低減機能

それらを担当するロボットセットへ配信される.

【収集先 Web サーバのリアルタイム表示】

図 -2 は，本 Web クローラがサポートしている収集先リアルタイム表示機能による表示例であり，世界のどこに存在する Web サーバを対象に収集しているかが一目で理解できる．図中では，収集ページ数が多い順に棒グラフを赤，青，黄で表示している．なお，Web サーバの設置位置特定のために，IP2Location 社の IP —経度・緯度—国名・市名変換テーブル☆1 を用いて，IP アドレスから経度・緯度情報への変換を行っている.

【Web サーバに対する負荷軽減】

収集にあたっては，高速収集と相手 Web サーバへの負担軽減といった相反する事項の両立が必要であり，相手 Web サーバの能力に合わせた収集制御が重要となる．具体的には，同一の Web サーバに対して短時間に複数回アクセスしたり，長期間にわたって連続して収集を行うと収集先の Web サーバに負荷を与えることになる．開発した分散収集型 Web クローラでは，Web サーバに対する負荷を軽減するために**表 -1** に示す機能をインプリメントした．以下では特徴的な機能について詳述する.

CGI ページからのホップ数制限は，Web サーバへの負荷を軽減するために，動的に生成されるページを大量にアクセスしないように設けた制限である．一般に Web クローラでは，リンクを辿る回数を制限した上で収集を行う．これは，無限にリンクを辿り続けないようにするためであり，たとえば，起点 URL からリンクを辿る回数を 10 ホップ等に制限する．しかし，日記サイトのように動的に生成されるページが収集途中に出現した場合，日記中のカレンダーを辿り続けることで内容を持たない Web ページを指定制限ホップ数に達するまで収集してしまう．このような動的に生成される Web ページに対する収集は相手の Web サーバに与える負荷が大きいため，できるだけ無駄なアクセスを減らす必要がある．そこで，本 Web クローラでは CGI ページからのホップ数を起点からのホップ数とは別に制限する仕組みを導入した.

robots.txt の googlebot 互換機能は，Google 社が提唱する robots.txt の記述方法への対応である．robots.txt は，http://www.waseda.jp/robots.txt のように Web サーバの直下に配置するテキストファイルであり，指定フォーマットを用いて排除すべき Web クローラや収集除外ページを記述できる．Google 社では，従来の robots.txt の記述法にはない，正規表現を用いた Web クローラ排除記述を採用しており，多くの Web サーバで利用されデファクトとなっている．たとえば，「disallow: /*.asp$」のような正規表現を用いることができる.

収集間隔の動的算出は，回線速度の遅いサーバや通信エラーが発生するサーバに対して，収集間隔を自動的に長くすることで負荷を減らすことを目的としたものである．特に，CGI を多用して重い処理を行っているサーバへの負荷を抑えることを目指した．具体的には，以下の計算式を用いて収集間隔を算出（計算機パラメータは**表 -2**）している.

☆1 http://www.ip2location.com/
※連絡先：〒 169-8555 東京都新宿区大久保 3-4-1 早稲田大学理工学術院　電話番号：03-5286-3503　yamana@waseda.jp

項目	説明
最小間隔値	最少の収集間隔（秒）.
遅延最大値 1	回線速度による遅延最大値（秒）.
遅延最大値 2	通信品質による遅延最大値（秒）.
実効速度	実際の回線速度（Byte/s）. 直近 10 リクエストの平均値で算出.
目標速度	目標とする回線速度（Byte/s）.
エラー回数	直近 10 リクエスト中のタイムアウト，通信断，送受信エラーの回数.

表 -2　収集間隔設定のためのパラメータ

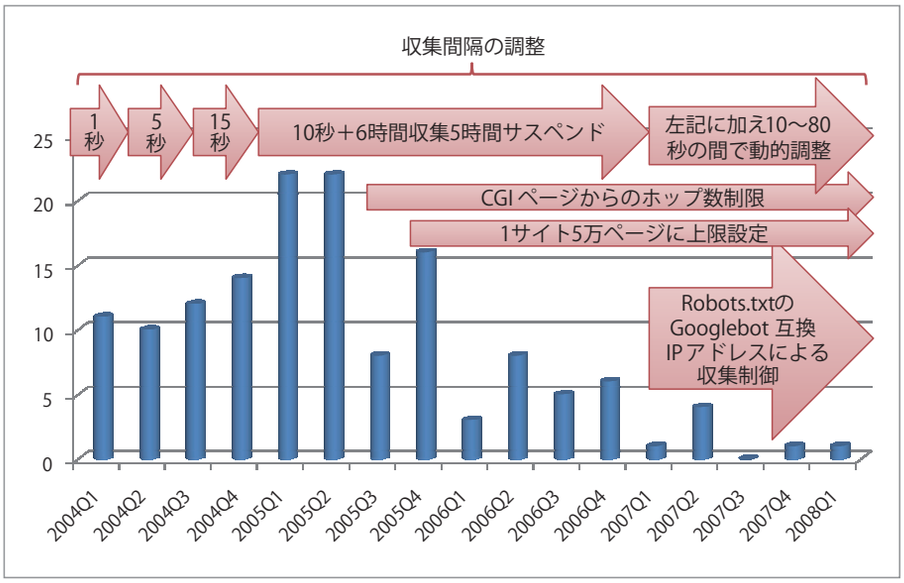

図 -3　苦情件数の推移

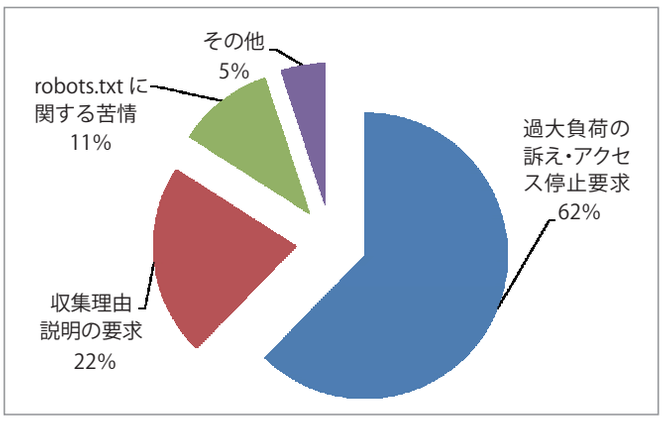

図 -4　苦情内容の分類

$$収集間隔 = 最小間隔値 + 遅延最大値1 \times \left(1.0 - \frac{実効速度}{目標速度}\right) + 遅延最大値2 \times \frac{エラー回数}{10}.$$

図 -3 は，我々が設けている苦情受付窓口に電子メールにより寄せられた苦情件数を 3 カ月ごとに集計したものである．2005Q1 ～ 2005Q2 にかけて苦情件数のピークが現れていることからも分かるように，単純に収集間隔を長くしただけでは苦情を減らすことができなかった．一方で，CGI ページからのホップ数制限や 1 サイトあたりの収集 Web ページ数の上限設定，収集間隔の動的算出は苦情件数の削減に一定の効果があることが分かる．

【 クロールに対する苦情 】

クロールに対する苦情は，**図 -4** に示すように大きく 4 つに分類できる．全体の 6 割以上を占めるのが「過大負荷の訴え・アクセス停止要求」である．2 番目に多いのは，Web クローラの動作の説明や収集理由の説明を求めるものである．これらの中には Web クローラのバグ報告も含んでいる．robots.txt に関する苦情では，先に述べたように「googlebot 互換の形式に対応していない」といった苦情や，「robots.txt を変更しても収集を継続している」といったものが多い．特に後者は，Web サーバ管理者側で Web クローラによる収集を排除しようとして robots.txt を書き換えたにもかかわらず収集が継続することに対する苦情である．Web ページを収集する直前に毎回 robots.txt を確認するのは，オーバヘッドが大きく，我々の Web クローラは，当初，24 時間ごとに再確認を行っていた．しかし，こうした苦情を教訓に現在では 6 時間ごとに再確認を行っている．

図中のその他には「業務上のデータが誤って収集された形跡があるので削除してほしい」といった要望や，「Web クローラのエージェント名を詐称して我々の Web クローラ名を名乗る第三者の Web クローラがもたらす問題への対応」を含んでいる．特に後者の問題に対しては，Web クローラを動作させている PC の IP アドレスを公開し，苦情を受け付ける前に当該 Web クローラの IP アドレスを確認してもらうことが重要である．

100 億超の Web ページ収集

2004 年 1 月～ 2006 年 7 月に合計で 14,456,201,906 ページの収集を行った．収集対象はテキストのみである．Web ページの収集にあたっては，2004 年 1 月の収集開始時に，筆者らが保有していた com，org，edu，net，uk，jp，us，ca，at の 9 つのトップレベルドメイン（TLD）

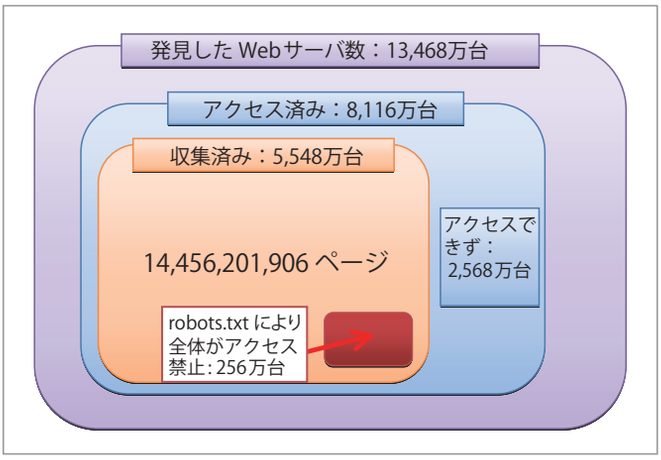

図 -5　発見した Web サーバの内訳

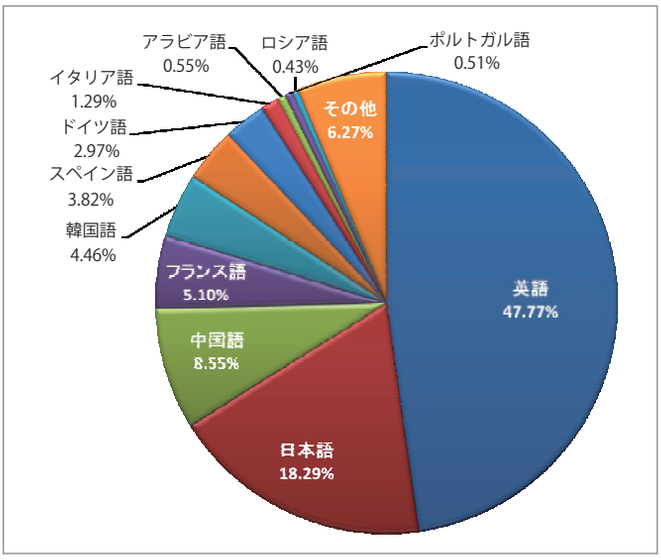

図 -6　com ドメインの記述言語分布

から合計約 600 万の Web サーバリストを起点として設定した．起点からは最大 15 ホップ先までを収集し，新規に発見した Web サーバは起点として随時追加を行った．

2006 年 7 月時点において，世界中で発見した Web サーバ数は約 13,468 万台であり，内 8,116 万台の収集を完了した．なお，8,116 万台のうち，実際に収集できたのは 5,548 万台であり，2,568 万台（収集済みサーバの約 32%）はすでに IP アドレスが存在しない等の理由からアクセスができなかった☆2．また，256 万台（収集済み Web サーバの 3%）については，/robots.txt に

☆2 IP アドレスが存在しない理由は，Web サーバがなくなった以外にアンカータグ内に記述された Web サーバのホスト名に誤りがある等の理由が考えられる．

☆3 ディスク故障等により一部のデータが利用できなくなり，解析にあたっては 107 億ページを利用している．

TLD	国名	取得ページ数	割合	平均ページ／サーバ
.com	-	4,070,092,124	38.05%	339
.net	-	890,604,259	8.33%	169
.de	ドイツ	878,838,449	8.22%	459
.org	-	745,984,032	6.97%	(※)
.jp	日本	543,654,556	5.08%	609
.ru	ロシア	407,169,769	3.81%	(※) 185 (.com, .net, .jp, .de 以外の平均)
.pl	ポーランド	321,209,334	3.00%	
.uk	イギリス	240,244,507	2.25%	
.edu	-	232,132,978	2.17%	
.nl	オランダ	215,722,380	2.02%	
.cn	中国	185,907,711	1.74%	
.it	イタリア	156,657,707	1.46%	
.kr	韓国	151,025,640	1.41%	
.us	アメリカ	143,135,686	1.34%	
.fr	フランス	129,326,495	1.21%	
other	-	1,385,290,926	12.94%	

表 -3　Web ページの TLD 分布

よりクローラによるアクセスが禁止されていた．残り約 5,000 万台については未アクセスである（**図 -5**）．

2006 年 7 月時点において Netcraft 社（http://www.netcraft.co.uk/）が発表している統計情報によれば，全世界の Web サーバ数は約 8,761 万台であり，我々のクローラは，Netcraft 社が把握している Web サーバ数の約 1.5 倍の Web サーバを発見することに成功している．

なお，2007 年 9 月からは，日本語 Web ページを 1 ページ以上含む Web サーバ約 150 万サーバを対象に 1 カ月ごとの更新収集を行っている．

収集された Web ページの解析

解析では，2006 年 7 月までに収集された 144.5 億 Web ページの中から 10,696,996,553 ページ（Web サーバ台数は 47,674,832 台）を対象として解析を行った☆3, [2]．**表 -3** にトップレベルドメイン（TLD）ごとの Web ページ数の分布と Web サーバあたりの平均 Web ページ数を示す．表 -3 に示されるように，ドメインによって Web サーバあたりの平均 Web ページ数が大きく異なることが分かる．従来，1 台あたりの平均 Web ページ数は 200 ページ前後と考えられていたが [3, 4]，com，de，jp ドメインでは平均 Web ページ数が大きくなっていることが確認された．

【 記述言語の分布 】

図 -6 に com ドメインの Web ページの記述言語の分

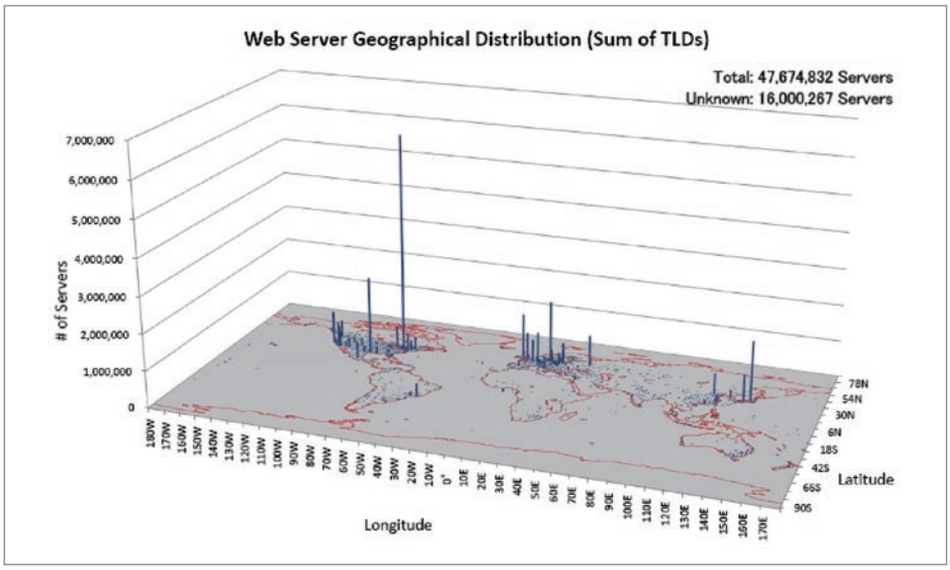

図 -7　Web サーバの地理的位置の分布

TLD	ドメイン名の説明	日本語ページの割合	当該国内での設置率
.jp	日本	90.3%	94.1%
.to	トンガ王国（南太平洋の島）	59.2%	0.1% 未満
.st	サントメ・プリンシペ民主共和国（西アフリカの島）	33.1%	0.1% 未満
.gs	サウスジョージア・サウスサンドウィッチ諸島（南大西洋の諸島）	29.0%	0.1% 未満
.bz	ベリーズ（中央アメリカ）	29.0%	0.3%

表 -4　日本語ページを多く持つ ccTLD

布を示す☆4．記述言語の判定は，ベイシス社の言語判定システム☆5 を利用しており，判定対象言語は，英語，日本語，中国語，フランス語，韓国語，スペイン語，ドイツ語，イタリア語，ロシア語，ポルトガル語，アラビア語の 11 言語である．11 言語に判別できなかった言語は，その他として分類される．図 -6 に示すように，Web ページ記述言語では英語が圧倒的シェアを占めており 47.77% となっている．一方，日本語は第二言語となっており 18.29% を占める．

また，日本語で記述された Web ページのうち，53.6% は com ドメインに存在し，jp ドメインには 35.4%，その他のドメインに 11.0% が存在していることが分かった．すなわち，日本語 Web ページを収集するために jp ドメインのみを対象としても日本語 Web ページの 1/3 程度しか収集できないことが分かる．

☆4 TLD ごとの収集済み Web サーバ比率は一定ではないため，全収集 Web ページを対象とした言語分布を求めた場合，偏りが発生すると考えられる．このため，特定の地域や言語に依存しない com ドメインの言語分布を調査した．

☆5 http://www.basistech.co.jp/language-identification/

【 Web サーバの地理的位置の分布 】

IP2Location 社の IP ―経度・緯度変換テーブルを利用して，Web サーバの設置位置の分布を求めた結果を**図 -7** に示す．図中の「Unknown」は，設置場所を特定できない Web サーバ数であり，全体の約 34% を占める．設置場所が特定できない理由は，ドメインから IP アドレスの解決ができなかった，もしくは IP ―経度・緯度変換テーブルでのエントリがなかったことによる．図に示されるように Web サーバは北米，欧州，日本を中心に設置されている．

表 -4 は，同一 TLD 内の Web サーバから発信される Web ページに対して，日本語ページの割合が多い国コードトップレベルドメイン（country code Top Level Domain : ccTLD）のトップ 5 である．表には，当該国・地域内に設置されている Web サーバの割合も示す．表 -4 に示されるように，jp ドメインを除いた ccTLD では，Web サーバの 99% 以上が当該国以外に設置されている．すなわち，日本語ページの割合が大きくても，当該国・地域内から Web ページが発信されているわけではなく，日本語ページの割合が多いから親日国であるということにはならない．このように，言語分布と共に地理的位置分布を用いることにより，詳細な解析が可能となる．

【 PageRank の分布 】

図 -8 に Web サーバ単位で PageRank[5] を計算し，最大値を 10 として正規化した PageRank の分布を示す．x 軸に PageRank 値，y 軸に x 軸で示される範囲の PageRank 値を持つ Web サーバ数を示す．PageRank 計算時の dumping factor（リンクを辿る確率）は，文

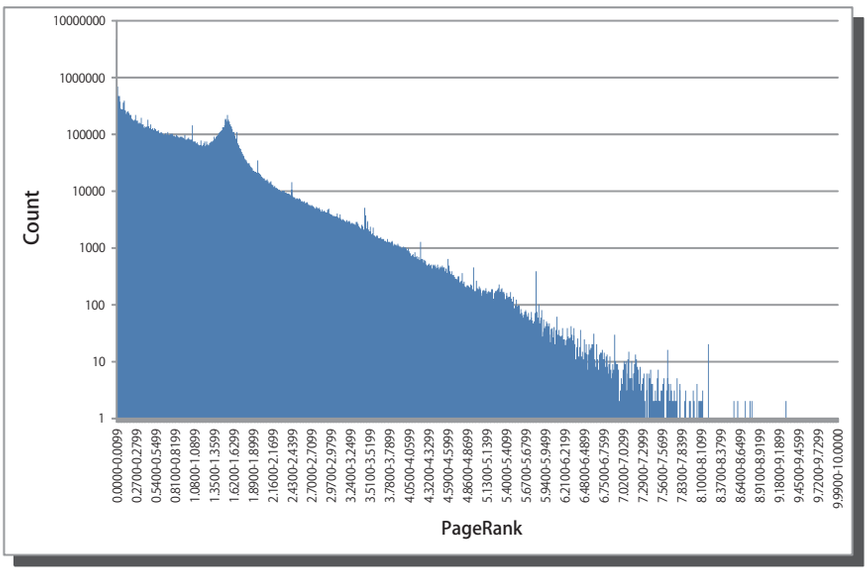

図 -8　PageRank の分布

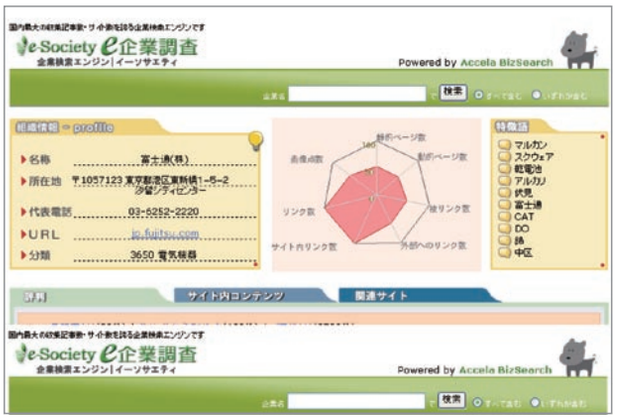

図 -9　e 企業調査プロトタイプ

献 5）と同様 0.85 を用いている．PageRank は，Web ページ間のリンク構造をもとに計算した Web ページの重要度を測る値であり，値が大きいほど多くの有用な Web ページから支持されていることを示す．Google 社では，この値をランキングに用いている．

PageRank は，本来，Web ページ単位で計算すべきであるが，ここでは計算量を抑えるため，Web サーバ間のリンクのみを用いて Web サーバ単位での PageRank を計算した．図に示すように PageRank 値の分布はべき乗則（power-law）に従っている．なお，PageRank が 1.5 付近にピークが出ているのは，大量のスパム Web サーバによるものである．

e 企業調査プロトタイプの構築

Web ページの解析が企業の特徴，戦略，評判などの抽出に役立ち，ビジネスに利用可能であることを示すために，収集済みの Web ページのうち，日本語で記述されたページを用い，企業の Web サイト上の活動を可視化する「e 企業調査プロトタイプシステム」を試作した（**図 -9**）．

本システムでは，企業 Web サイトに対する基本的な情報を提供する「組織情報」，Web サイト上の活動を概観するための「レーダチャート」，当該企業の特徴を表現する「特徴語」，評判情報を表示するための「評判タブ」，当該サイト内のコンテンツを分類して表示する「サイト内コンテンツタブ」，リンク関係をもとに抽出した関連サイトを表示する「関連サイトタブ」を機能として持つ．

【 競争企業に対する比較調査 】

競争関係にあると思われる有名企業数社をピックアップし，差異の有無の調査を主に「レーダチャート」を比較分析した．

レーダチャートは，**図 -10** に示すように，7 角形で表される．内周から外周に向けて 0 ～ 100 点の点数付けを行い，50 点を登録された全企業の平均値とした．平均値に対して，少ないものはマイナス（0 点方向へ），多いものはプラス（100 点方向へ）とした．

「静的ページ数」は調査対象 Web サイト内の静的なページ数を示す．「動的ページ数」は調査対象 Web サイト内で動的に生成されるページ数を示す．「被リンク数」は，調査対象 Web サイトに対して張ってあるリンク数を示す．「外部へのリンク数」は，調査対象 Web サイトから，調査対象 Web サイト以外の他の企業やドメインの Web サイトに張ってあるリンク数を示す．「サイト内リンク数」は調査対象 Web サイト内に閉じたリンク数を示す．「リンク数」は調査対象 Web サイトが持つすべてのリンク数であり，「外部へのリンク数」と「サイト内リンク数」の合計である．「画像点数」は，調査対象 Web サイト内に含まれる画像の数である．

図 -10 に示す A 社と B 社のレーダチャートは，電気系の同業 2 社のものである．図に示されるように，電気系の同業 2 社間においても Web ページの作りが大きく異なることが分かる．実際に当該 Web ページにアクセスしたところ，A 社はユーザビリティを高めるためにサイト内リンクが多いのに対し，B 社は静的なコンテンツ中心の従来型の Web ページであることが分かった．このように，企業の Web ページのリンク数や関連性の分析により，企業活動をある程度推測できる．

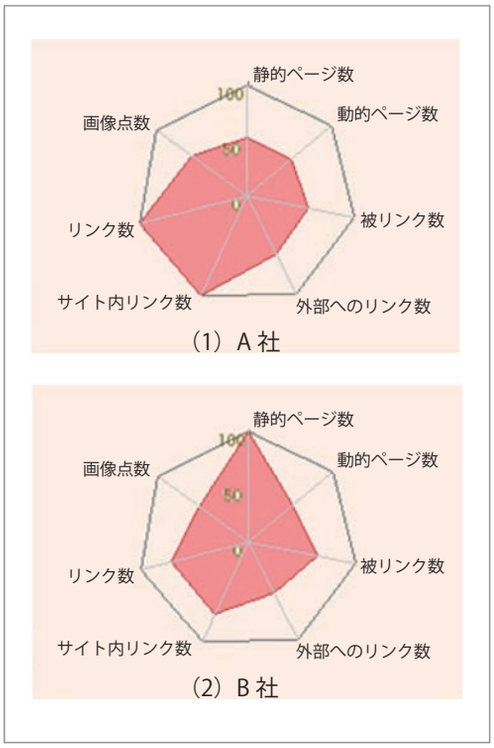

図 -10　A 社と B 社の Web ページ解析

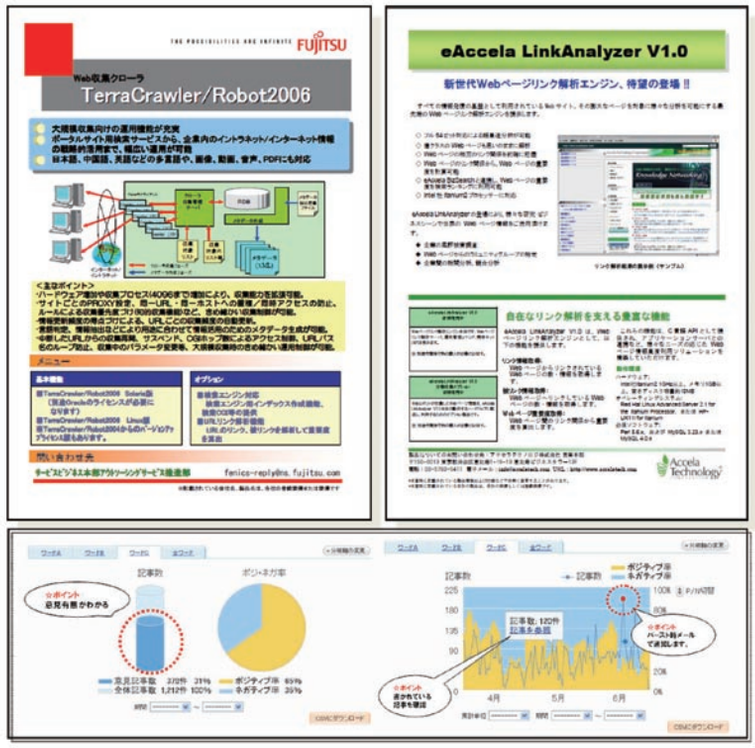

図 -11　技術移転

技術移転

本研究の成果は**図 -11** に示されるように商用クローラシステム（図左上）やリンク解析ツール（図右上）に技術移転されている．また，e 企業調査フレームワークは，Web サービスを用いて Web アプリケーション（図下）に組み込まれ利用が進んでいる．

なお，本研究によって得られた解析データ（言語分布，Web サーバ地理的位置分布，Web ページ間最短経路探索，フォワードリンク・バックリンク探索）は，http://www.yama.info.waseda.ac.jp/e-society/ において公開されている．

参考文献

1）Hirate, Y. and Yamana, H. : Web Structure in 2005, Proc. of the 4th Workshop on Algorithms and Models for the Web-Graph, Springer-Verlag, LNCS 4936, pp.36-46 (2008).
2）童　　芳，平手勇宇，山名早人：全世界の Web サイトの TLD・言語分布・地理的設置位置の特定，日本データベース学会論文誌，Vol.7, No.1, pp.31-36 (2008).
3）Lawrence, S. and Giles, C. L. : Searching the World Wide Web, Science, Vol.280, No.5360, pp.98-100 (1998).
4）Lawrence, S. and Giles, C. L. : Accessibility of Information on the Web, Nature, Vol.400, pp.107-109 (1999).
5）Page, L., Brin, S., Motwani, R. and Winograd, T. : The PageRank Citation Ranking : Bringing Order to the Web, Proc. of the 7th WWW Conf., pp.161-172 (1998).

（平成 20 年 7 月 30 日受付）

村岡　洋一（正会員） muraoka@waseda.jp
1971 年イリノイ大学電子計算機学科博士課程修了．Ph.D. 日本電信電話公社電気通信研究所を経て 1985 年早稲田大学理工学部教授．1995 年同大理工学術院教授．同大副総長，本会副会長など歴任，本会フェロー．

山名　早人（正会員） yamana@waseda.jp
1993 年早稲田大学大学院・理工学研究科博士課程修了．博士（工学）．1993 ～ 2000 年電総研．2000 年早稲田大学・理工学部助教授．2005 年同大理工学術院教授，現在に至る．データマイニング，情報検索，並列分散処理の研究に従事．

松井くにお（正会員） kunio.matsui@us.fujitsu.com
1980 年静岡大学工学部情報工学科卒業．同年（株）富士通研究所入社．2003 年東京工業大学大学院情報理工学研究科後期課程修了．博士（工学）．2007 年より米国富士通研究所勤務．自然言語処理，情報検索の研究開発に従事．本会理事を歴任．

橋本三奈子（正会員） hasimoto.minako@jp.fujitsu.com
1984 年東京女子大学文理学部日本文学科卒業．同年富士通（株）入社．1984 ～ 97 年情報処理振興事業協会（現，情報処理推進機構）出向．富士通復帰後，情報検索システム，インターネット収集クローラ，検索支援辞書開発に従事．

赤羽　匡子　akabane.kyoko@jp.fujitsu.com
1988 年宇都宮大学農学部農業経済学科卒業．同年富士通（株）入社．1995 ～ 99 年科学技術振興事業団（現，科学技術振興機構）客員．富士通復帰後，大規模検索支援電子辞書開発，情報検索システム開発に従事．

萩原純一（正会員） jhagiw@accelatech.com
1993 年早稲田大学大学院・理工学研究科修士課程修了．同年（株）富士通研究所入社．1995 ～ 2001 年富士通（株）．2001 年アクセラテクノロジ（株）設立に携わり，現在に至る．並列化コンパイラ・情報検索の研究開発に従事．