Socio Sense：過去9年に及ぶWebアーカイブから社会の動きを読む

喜連川 優 *1
田村 孝之 *2, *1
今村 誠 *2
藤原 聡子 *3
豊田 正史 *1
鍛治 伸裕 *1
高山 泰博 *2

*1 東京大学
*2 三菱電機
*3 三菱電機インフォメーションテクノロジー

はじめに

2008年7月のGoogle公式ブログによると，1兆個のURLが確認されている[1]．1991年に生まれたWebはきわめて急速に，非常に広く世の中に受け入れられた稀有なメディアと言えよう．コンピュータサイエンスの歴史の中で，50年後にもWeb技術のインパクトは明確に位置づけられることは確実である．2006年12月に，Time誌はperson of the yearを「You」とした[2]．特定の政治家，実業家，学者ではなく，ネットユーザ全般を示唆する「You」を記載した表紙は読者に新鮮な驚きを与えた．Web上でCGM, UGC☆1とも呼ばれる個々人の意見発露が可能となり，それが社会を動かし始め，製品改良に影響を与え，ユーザニーズに即したサービスを生み出すに至ると同時に，政党が政策アピールに活用するなど，政治手法にも影響を与えつつある．Time誌のメッセージは，非常に大きく社会に影響を与えることを可能とするメディアの出現によって従来にない新しい時代に突入したことを鮮明に表現している．

さて，本Socio Senseプロジェクトは2003年から開始されたが，Webが社会を大きく動かすメディアになるであろうとの想定の中で，その特性を十二分に活用し，検索エンジンとは異なる新たな価値創出を目指してきた．

☆1 CGM（Consumer Generated Media），UGC（User Generated Content）は消費者発の情報を集約・共有するWebサイト，およびそのコンテンツを指す．代表的なものに，ブログ，クチコミサイト，写真・動画共有サイトなど．

【Socio Sense】

現在，人々は2つの世界，すなわち，実世界とWebの世界で生活をしていると言える（**図-1**）．とりわけ，インターネットを介したWebの利用は広く浸透し，ページの検索や閲覧，商品の購買，ブログでの日々の行動の

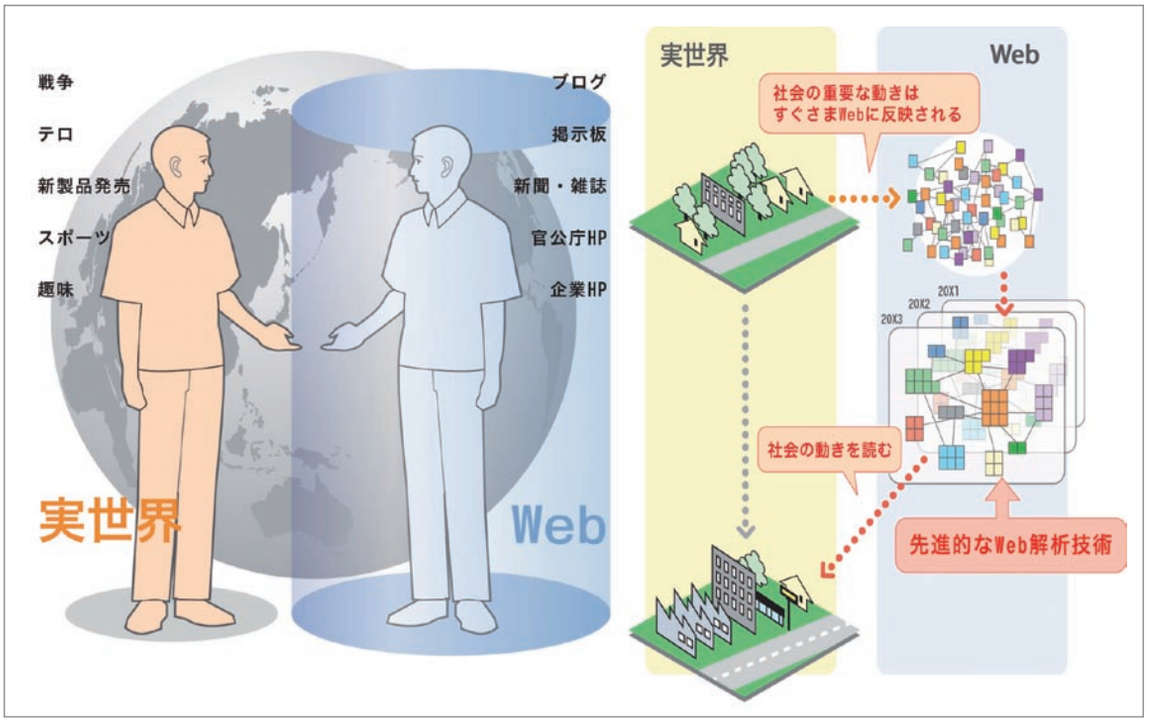

図-1 実世界とWebの転写構造

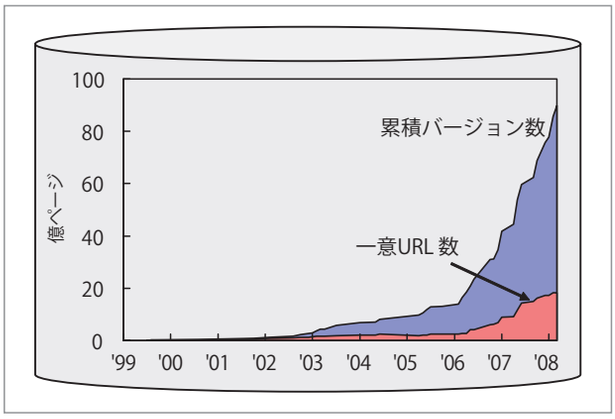

図-2　Webアーカイブの規模の推移

紹介など多くの時間をネット上で過ごすようになりつつある．それと並行し，近年では，重要なアナウンスはほとんどWeb上でもなされるようになり，最初の情報発信メディアとしてWebを取り上げることさえ進んでいる．すなわち，Webは実社会を映す鏡，あるいは，実社会の動きを把握するための「センサ」として利用できるのではないかと考え，社会分析のためのツールとしてSocio Senseなるシステムを開発してきた．

Web上の情報を分析すること自体は，広くなされているわけであるが，実際に利用可能なツールは現時点では検索エンジンしかない．もちろん検索エンジンに関しては，非常に多くの研究・改良がなされているものの，あくまでも，検索語と広告の紐付けによるビジネスモデルに依存していることから，分析という観点での利用しやすさは必ずしも十分とは言えない．第1に，検索エンジンは検索語を含むドキュメントを検索結果として戻すことを原則としているが，ある分野に関して分析をする場合，特定の検索語に強く影響されることなく，関連する多様な情報を俯瞰できる機能が望まれる．第2に，検索エンジンは現在のスナップショットに対しての検索を可能とするものであり，過去の情報は原則アクセスできないという限界がある．情報分析をする場合，現時点での状況だけではなく，過去からの現在に至る経緯，時系列的な発展過程を見ることが現象の理解に大きく貢献することは明白であり，Webアーカイブ基盤の構築は重要と言える．

このような問題認識から，現行の検索エンジンでは利用しにくい，あるいは，提供されていない分析機能を研究開発のターゲットとした．

【Webアーカイブ基盤】

過去からの経緯を解析するためには，当然のことながら，長期間にわたるWebデータを保存したWebアーカイブの構築が不可欠である．東大では1999年より，Webクローリングを開始し，現在まで10年にわたる収集量は約100億ページに達する．Web空間全体ではなく，.comを含め日本語で書かれたページのみを対象として収集してきた．これは，大学の一研究室が全世界のWebページを連続的に収集することはそもそも体力的に不可能であるとの判断による．**図-2**にWebアーカイブの規模の推移を示す．最大のWebアーカイブ保有機関は米国Internet Archive（IA）[3]であるが，そのWebサイトの情報によれば全世界850億ページを有するとされている．当該機関はコレクションに大きな興味があると想定され，当初よりALEXA社よりページデータの提供を受けてきた．現在は，IIPC（International Internet Preservation Consortium）の中核機関として，クロールツールを開発するとともにWebアーカイブの構築を開始しようとする諸外国に技術供与をし，自身でもクロールを行いつつある．そのほか，海外ではオーストラリア，スウェーデン，デンマーク，フランス，ノルウェー，イギリス等において，図書館が中心となり，Webアーカイブの構築に取り組んでいるが，その多くは，人手によって特定したサイトを収集する方式を採用しており，規模的には小さなものにとどまっている．我が国では2002年から国会図書館が「WARP（現インターネット情報選択的蓄積事業）」なるプロジェクトにおいて約2000サイトを対象にアーカイブを構築しつつある．

IAを始めとするWebアーカイビングプロジェクトは一括収集を固定周期で繰り返しており，その周期は最短でも1カ月程度である．Webページの更新頻度はさまざまであり，ニュースサイトのように日々更新されるものもあれば，1年に一度程度の更新しかないサイトも数多くあるため，これらを一様な時間間隔で収集することは適切ではない．これに対し，本Webアーカイブにおいては，ページごとに更新頻度を観察し，各ページの収集間隔をその更新間隔に適応させる制御を2005年以降導入している[4]．

また，2003年，IAではRecallなるサービスを行ったことがあるが，現在は停止しており，現時点で分析ツールは提供されていない．Socio Senseでは蓄積したコンテンツに対し，語に関する情報を集約したタームインデックス，リンクに関する情報を集約したグラフインデックス，URLごとの複数バージョン情報を集約した時系列インデックス，コミュニティ抽出（後述）やスパム判定の結果を格納した属性インデックスなど，多様なインデックスを付与し，大量のコンテンツをさまざまな角度から柔軟に分析するための基盤を形成している．

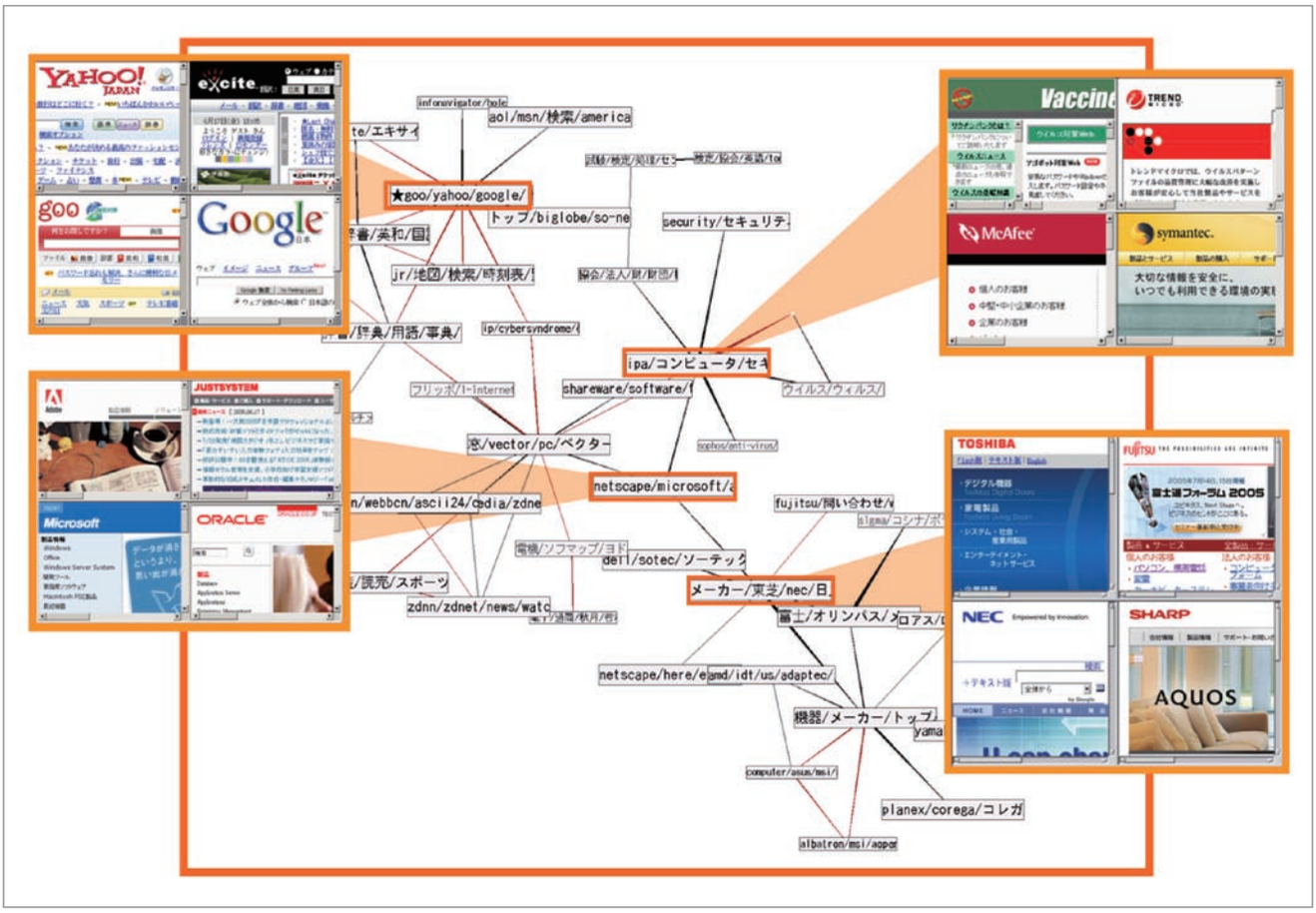

図 -3　Web 空間の構造俯瞰図

【 Web 空間における構造の俯瞰 】

情報分析において，対象となるトピックスに関連する情報空間の俯瞰機構は必須である．Socio Sense では，Web コミュニティ技術を利用し，俯瞰機構を実現した．互いに関連するページ群は稠密なグラフ構造を形成することから，Web コミュニティの抽出実験がなされ，いくつかの興味深い抽出結果が示されたものの [5]，コミュニティ間の関連については検討されていなかった．本プロジェクトでは独自の手法により，コミュニティを単位とする Web 空間の地図を作る手法，ならびに，コミュニティ間の関連を表すリンクをたどる等当該地図をインタラクティブに操作可能なシステムを開発した [6]．

図 -3 はコンピュータ関連の地図を描いたものであるが，IBM，HP，日立などのサーバベンダのコミュニティの周辺に，マイクロソフト，オラクルなどが含まれるソフトウェアベンダのコミュニティが接続されており，さらに，最近では，セキュリティ関連の製品が数多く開発されていることから，シマンテックやトレンドマイクロなどが含まれるセキュリティベンダのコミュニティへと辿ってゆけることが分かる．このように，コンピュータ関連の産業連関図のようなものを自動的に抽出することができる．対象とする分野に対して，関連あるいは競合企業を把握するなど，分野の概要を一目で捉えることができる本俯瞰機能は大変有用である．さらに，コミュニティ内のメンバを一覧表示することや，コミュニティ間の接続リンクの強さをグラフ描画のパラメタとし，強い接続関係のみを表示させるなど，柔軟なインタラクションを可能としている．

【 Web 空間の時系列分析 】

一般に，現時点での状況のみを見るのではなく，過去からの流れを見ることにより，対象への理解が大きく深まることは多々あることと言える．長期にわたる Web アーカイブ基盤の構築により，大局的な動きを把握することが可能となる．直感的には先に示した図 -3 を時系列的に複数枚用意し，それを自由自在に串刺しにして見ることが理想となるが，3 次元的な可視化機能は現時点では実装されておらず，今後の課題である．

手法の詳細は文献 7）にゆずるが，**図 -4** に，2001 年 9 月に発生した同時多発テロ直後の 10 月にクロールしたデータを元に，前後でのテロに関するコミュニティの変化を示す．これは左から右に年代ごとのコミュニティ形成の推移を示したものである．9.11 事件以前には，テロに対しての日本での意識はきわめて低く，わず

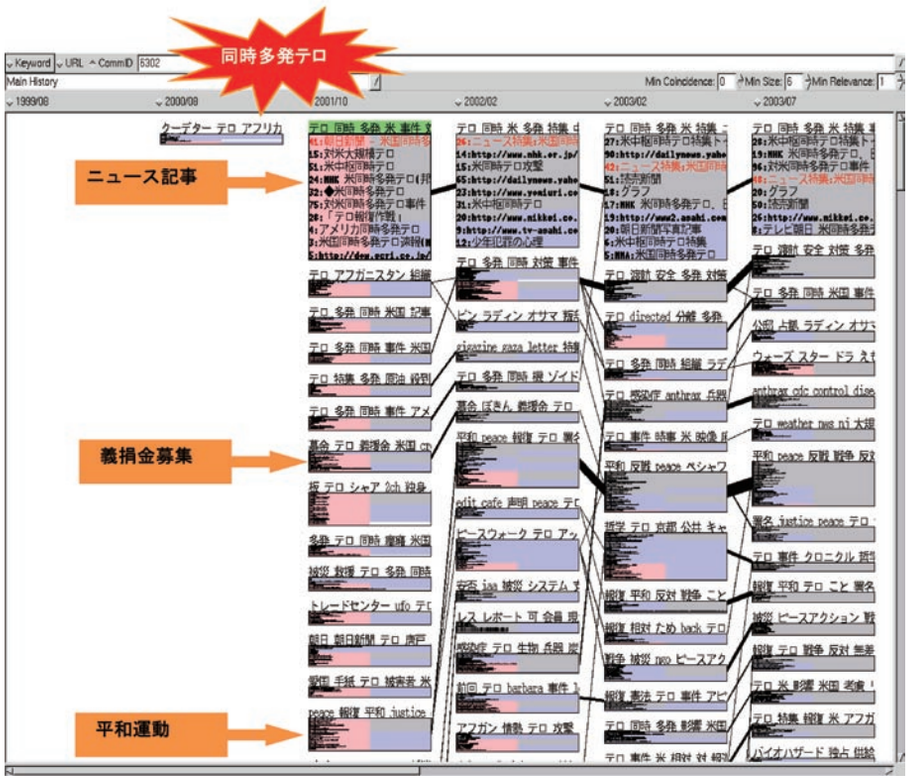

図 -4　Web の時系列変化分析（テロ）

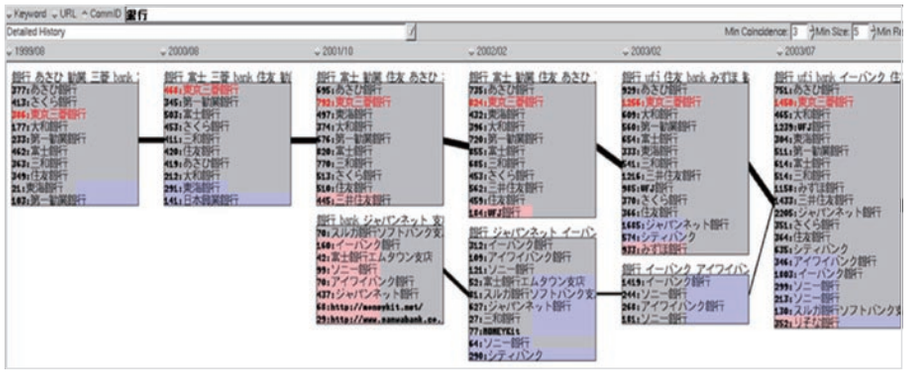

図 -5　Web の時系列変化分析（銀行業界）

かな数のコミュニティしか存在しなかったのに対し，当該事件以降，多く語られるに至り，義捐金を募るコミュニティ，報復攻撃に反対するコミュニティなど多様なコミュニティが発生していることが分かった．なお，図中の四角はコミュニティを示し，その中にコミュニティを形成する個々の URL を示しており，四角の横のつながりを示す線は年代ごとの対応するコミュニティを示す．

図 -5 には，銀行のコミュニティに関する変化を示す．最上段の 1 行は，一般の銀行コミュニティを示す．2 段目を見ると，2001 年に 1 つコミュニティが発生していることが分かる．この年に多数のインターネット銀行が生まれており，図で薄赤色の URL は新規に発生したものを示している．ソニー銀行やジャパンネット銀行等が含まれていることが読み取れよう．右に，すなわち，時間が経過するのを追うと，2003 年には 1 段目の銀行のコミュニティと一緒になってしまっている．これは，通常の銀行もインターネット銀行が提供する機能を提供するに至り，区別する特徴が少なくなり，ページ作成者のリンクの張り方に変化が起きたと推察される．コミュニティ抽出技術が社会の受け止め方を浮き彫りにしたとも見なせる．

【 Web の時空間分析 】

コミュニティを対象とした時間変化は大きな挙動を見る場合には適しているものの，マーケティング等の業界解析には，より微視的な挙動分析を行いたい場合も多々発生する．**図 -6** は，i-mode の検索エンジンサイトに関して，Web グラフの時系列変化を表示したものである．ここで，図中の四角はコミュニティではなく，サイトを指す．また，本ツールでは，サイトの空間配置に絶対的な意味付けはないが，各年の図において，同じサイトは同じ座標位置にレイアウトされる．図において，赤いサイトは前年にはなく当該年に新たに発生したサイトであることを示す．1999 年には Oh-new など

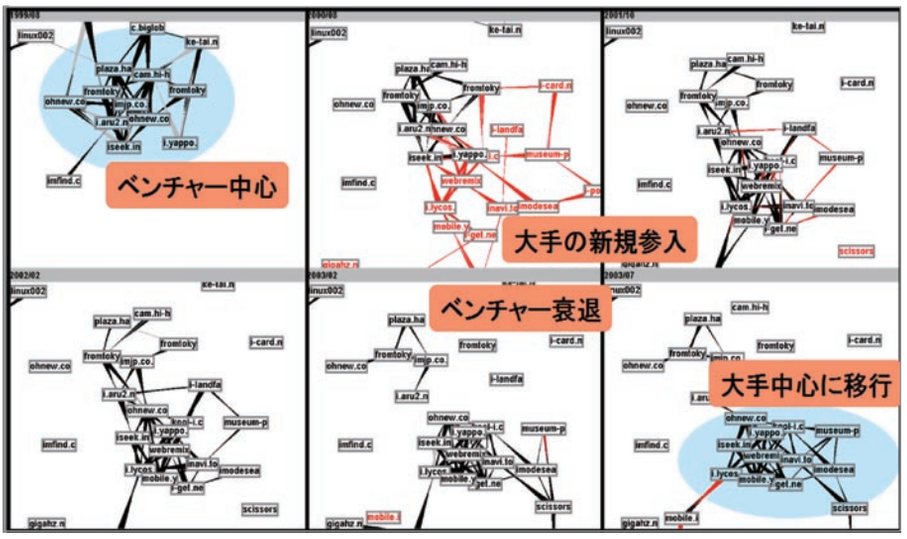

図 -6　Web の時空間分析（i-mode 用検索エンジンの変遷）

のサービスが有名であったが，翌年には Yahoo! や Lycos が参入し，さらに時間を経て，右下に目を転ずると，グラフの稠密個所が 1999 年には左上にあったものが 2003 年には右下に移動しており，業界のパワーシフトを容易に読み取ることができる．

【 ブログフィードの時空間解析 】

テクノラティによれば，日本語で記載されたブログの量は英語のそれを上回るとされており，我が国におけるブログメディアの普及は著しく，その解析は有用と言える．キーワードごとの日ごとのブログ量によるトレンド紹介は広くなされているところであるが，構造解析は筆者らの知る限り見当たらない．**図 -7** は「生協の白石さん」のブログフィードの発展過程を追跡したものである．(a) を見ると左の中心にあるサイトが人気の高いインフルエンサであり，ほとんどのブロガーは当該サイトを引用して記事を書いていることが分かる．時間を巻き戻し，約 1 カ月前に遡った時点での状況が (b) であり，この時点で当該サイトが生まれたことが分かる．さらに遡ったのが (c) であり，実は，図中の右にあるサイトが火つけ役となった起点であることが判明する．このように，時空間解析をすることにより，ブログの数だけではなく，より深い情報を得ることが可能となる．なお，図では 3 つのスナップショットを

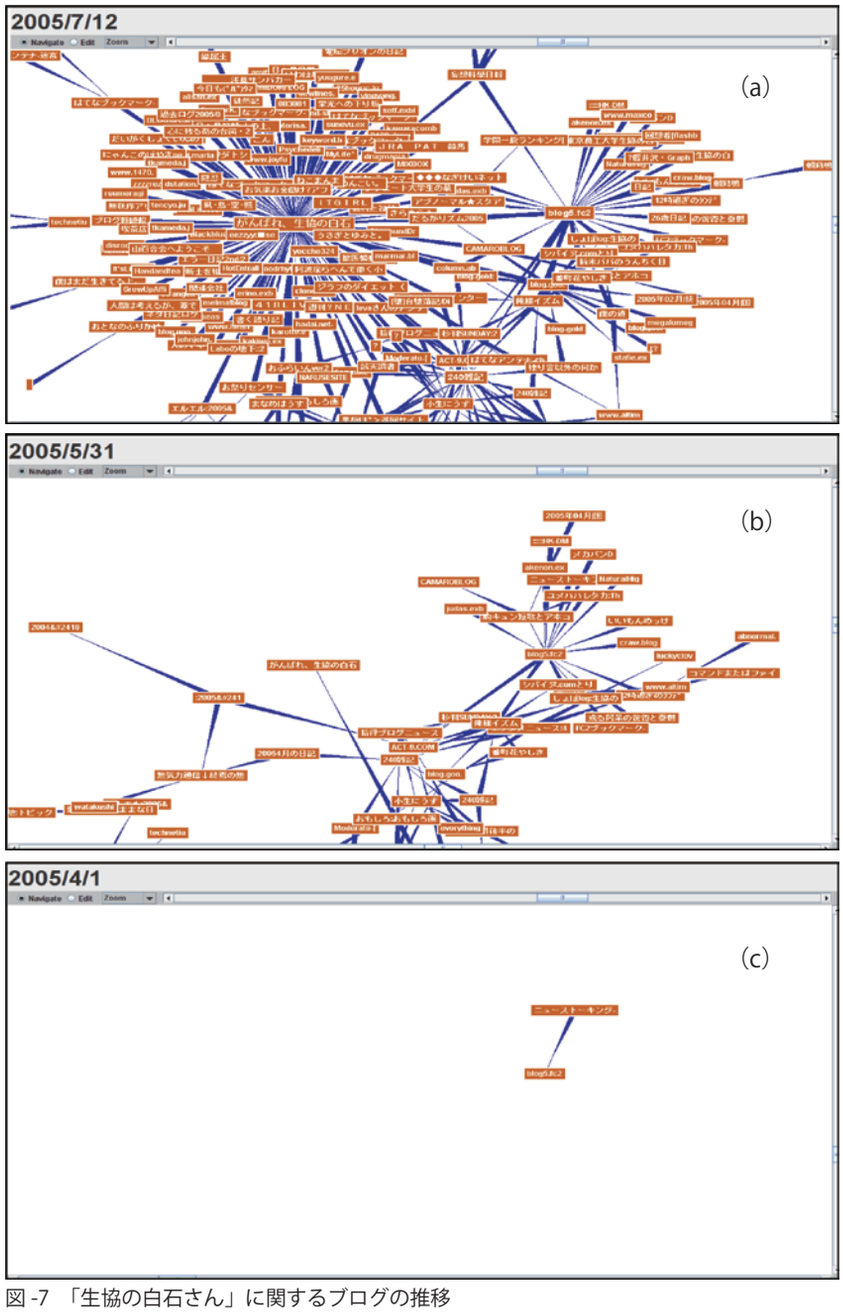

図 -7 「生協の白石さん」に関するブログの推移

図 -8　大規模壁面ディスプレイによる高精細可視化システム

示したが，ツールは連続的に変化を表示することが可能であり，Web 空間の時間的構造変化をタイムマシンのように把握することができる．

【大型高精細可視化システム】

いくつかの Web 空間分析ツールを開発したが，通常サイズのディスプレイでは結果を表示・把握することが極度に困難であることから，**図 -8** に示すように，15 個の個別ディスプレイを結合した大規模壁面ディスプレイ上に高精細可視化システムを構築した．当該システムにおいては，千個以上のコミュニティを同時に俯瞰し，インタラクティブに閲覧が可能であり，また，コミュニティの時系列変化についても同時に閲覧可能となっている．

【その他の適用事例】

電通，新井教授（専修大）との共同研究により，Socio Sense のマーケティング分野への利用に関し，消費財に適用し，テレビなどへのマス広告とブログ反応の関係を解析し，両者には強い相関があることを明らかにするとともに，ブロガーがどのような側面に強く反応するかを解析し，日本広告学会に発表した[8]．ブログのキーワード抽出に関しては本年 3 月より nikkansports.com の社会面において継続的に開発したツールが利用されている．

アーカイブを利用することによる新語の形成解析も進めた．ブログ等では辞書にはない砕けた表現が多用されることから，Web の解析においては新語の獲得は重要な課題と言えるが，同時に，アーカイブを利用することにより，その語が社会に受け入れられる様子を把握することが可能となる．詳細は省略するが，機械学習を利用することにより，「ハラシマる」「バモる」「ファブる」「モフる」等多くの新語を獲得した．たとえば，「ハラシマる」は，原稿という単語の「稿」の漢字を「縞」と間違え，「原稿を書く」を「原縞を書く」と書いたことが発端で，とりわけ同人誌の執筆者の間で，必死に書きものをする際にハラシマ（原縞）ると表現するようになったようである．その他有名なものとしては，Google を利用することを「ググる」と記載することが多い．また，最近は花粉症に悩む人々が増え「ファブる」という言葉が生まれている．これはファブリーズという商品からきている．これらの新語の利用頻度（百万文当たりの出現頻度）を Web ページ数で正規化したものを**図 -9**，**10** に示す．一般的な語，マズい等の語では，利用頻度は変化がないことも分かる．言語学者との共同研究も少しずつ始めつつある[9]．

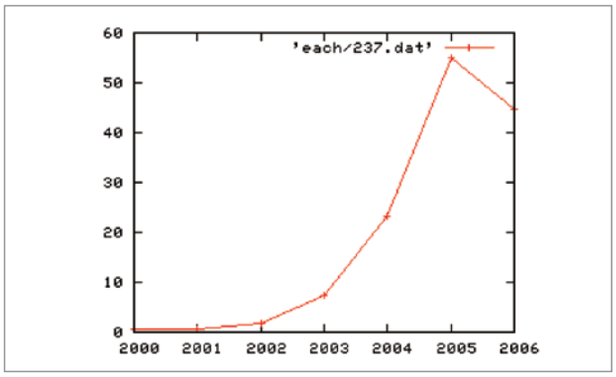

図 -9　新造語「ググる」の時系列頻度

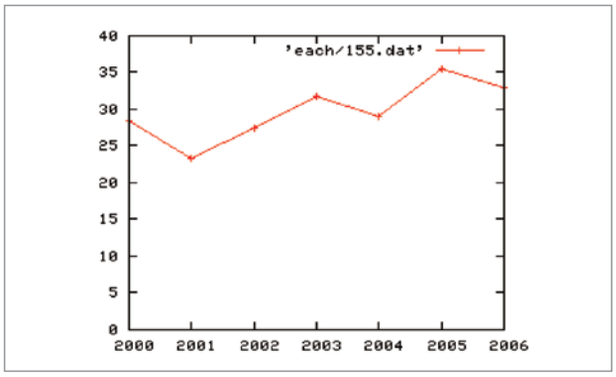

図 -10　既存語「マズい」の時系列頻度

おわりに

リーディングプロジェクト e-Society において開発を進めてきた Socio Sense なるシステムについてその概要を紹介した．構築してきた 10 年にわたる 100 億ページ Web アーカイブは世界的に見ても Internet Archive に次いで大きな存在であり，日本における Web 文化の資産と見なすこともできよう．また，当該アーカイブ基盤上で，多様な社会分析ツールを開発するとともに，その利用形態ならびに有効性を紹介した．分析対象も，ま

た，分析したい切り口も多様であることから，すべての要求に対応可能な万能ツールの開発は困難であると判断し，多様なツールを用意し，ユーザが目的に応じてそれらを駆使して分析を進めるという利用形態を描いた．巨大な Web データの収集から始め，膨大なデータの管理基盤を構築し，さらに，種々の分析ツールの構築を行うという一連の作業を行うには 5 年という期間は決して十分とは言えなかったものの，Web を介して実社会の動きをセンスするアプローチについては確かな手応えを得ることができた．Socio Sense はまだまだ未熟なシステムであり，さらなる開発が不可欠であるが，本稿により，わずかながらでもそのポテンシャルにご興味をいただければ幸甚である．

参考文献

1) Alpert, J. and Hajaj, N. : We Knew the Web was Big…, The Official Google Blog (July 2008), http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html
2) Grossman, L. : Time's Person of the Year : You. Time Magazine (Dec. 2006), http://www.time.com/time/magazine/article/0,9171,1569514,00.html
3) Internet Archive, http://www.archive.org/
4) 田村孝之，喜連川優：大規模 Web アーカイブ更新クローラにおけるスケジューリング手法の評価，電子情報通信学会論文誌，Vol.J91-D, No.03, pp.551-559 (Mar. 2008)
5) Rajagopalan, S., Kumar, R., Raghavan, P. and Tomkins, A. : Trawling the Web for Emerging Cyber-communities, In *Proceedings of the 8th World-Wide Web Conference* (1999).
6) Toyoda, M. and Kitsuregawa, M. : Creating a Web Community Chart for Navigating Related Communities, In *Conference Proceedings of Hypertext 2001*, pp.103-112 (2001).
7) Toyoda, M. and Kitsuregawa, M. : Extracting Evolution of Web Communities from a Series of Web Archives, In *Proceedings of the Fourteenth Conference on Hypertext and Hypermedia (Hypertext 03)*, pp.28-37 (Aug. 2003).
8) 馬渡一浩，富田英裕，新井範子，豊田正史，鍛治伸裕，喜連川優：ブログからレピュテーション分析の可能性を探る，日本広報学会第 12 回研究発表大会(2006)．
9) 福島健一，鍛治伸裕，喜連川優：機械学習を用いたカタカナ用言の獲得，言語処理学会第 13 回年次大会(2007)．

(平成 20 年 10 月 7 日受付)

喜連川　優(正会員)　kitsure@tkl.iis.u-tokyo.ac.jp
第 2 部「9. ストレージフュージョン：ストレージシステムとデータベース管理システムの融合」を参照.

豊田　正史(正会員)　toyoda@tkl.iis.u-tokyo.ac.jp
東京大学生産技術研究所准教授．Web マイニング，情報可視化の研究に従事．ソフトウェア科学会，ACM，IEEE Computer 各会員.

田村　孝之(正会員)
Tamura.Takayuki@eb.MitsubishiElectric.co.jp
三菱電機（株）情報技術総合研究所専任，東京大学生産技術研究所研究員．博士(工学)．Web アーカイブの研究開発に従事.

鍛治　伸裕(正会員)　kaji@tkl.iis.u-tokyo.ac.jp
東京大学生産技術研究所特任助教．博士（情報理工学）．自然言語処理の研究に従事.

今村　　誠(正会員)
Imamura.Makoto@bx.MitsubishiElectric.co.jp
三菱電機（株）情報技術総合研究所専任．博士（情報科学）．構造化文書処理，自然言語処理の研究開発に従事.

高山　泰博(正会員)
Takayama.Yasuhiro@ea.MitsubisiElectric.co.jp
三菱電機（株）情報技術総合研究所専任．自然言語処理，文書処理の研究開発に従事.

藤原　聡子　sfujihara@mdit.co.jp
三菱電機インフォメーションテクノロジー（株）第一事業本部データセントリックソリューション第二部次長．データセントリックにかかわるソリューションシステムの開発・販売に従事.