特集論文

大規模日本語Webアーカイブの構築とその分析

田村孝之*
喜連川　優**

Building a Large-scale Japanese Web Archive for Societal Analysis

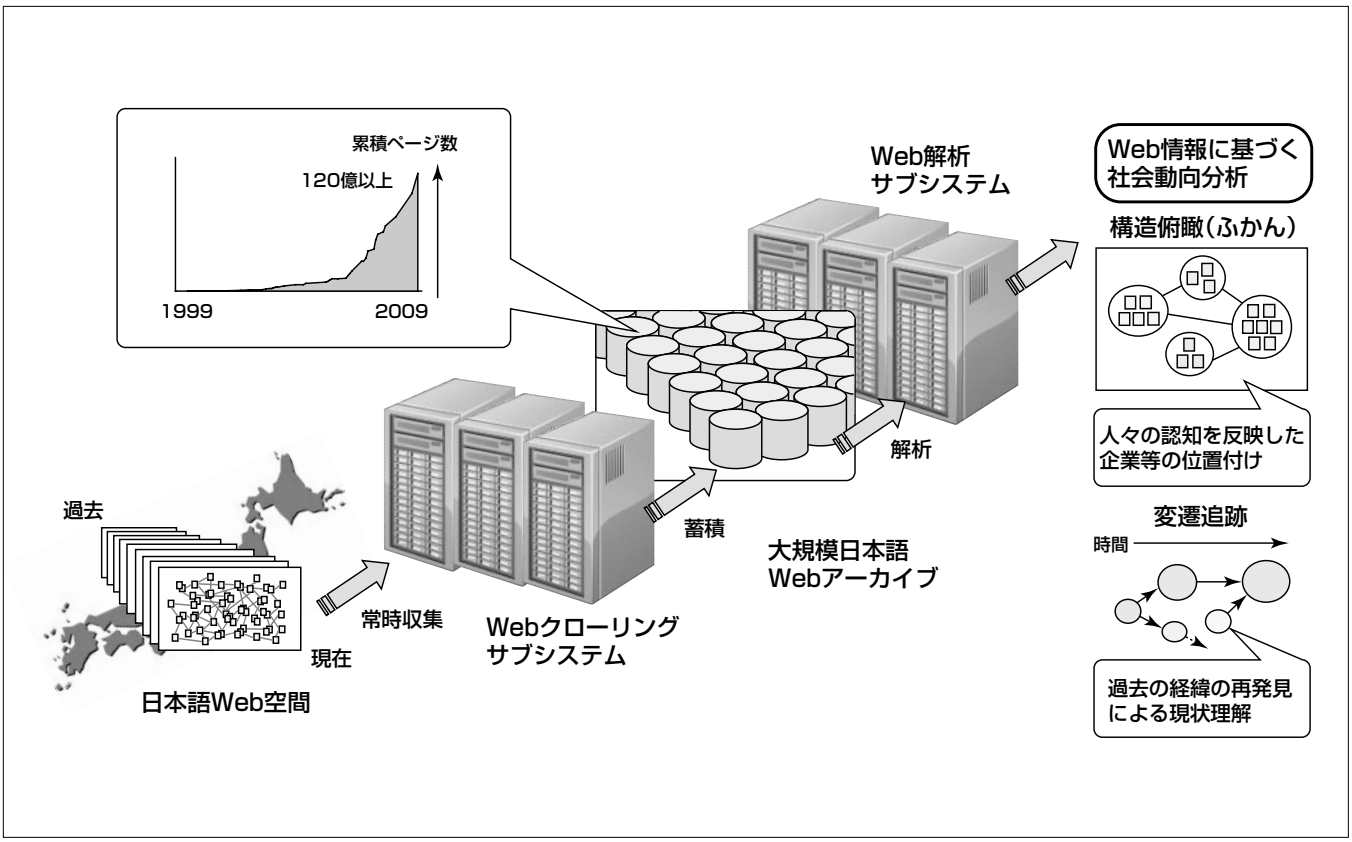
Takayuki Tamura, Masaru Kitsuregawa

要　旨

三菱電機では，東京大学との産学連携研究を通じWebの情報構造や時間変化の解析に基づく社会動向分析の実現に取り組んでおり，その基盤として，日々変化するWeb情報を蓄積した大規模日本語Webアーカイブの構築に携わってきた。Webは実世界の様々な事象を即時に把握するには不可欠のメディアであるが，現行の検索エンジンは数十億のWebページから選択した数十件をリストアップするにすぎない。複数のWebページを内容の関連性に基づいてグループ化し，グループ相互のつながり（情報構造）を視覚化するとともに，過去から現在に至る情報の発展・伝播（でんぱ）過程を提示することで，Web情報のより高度な活用が可能になる。

膨大なWeb情報に様々な分析手法を適用する上で重要となるのが，分析対象データへの自在なアクセスを提供するWebアーカイブ基盤である。Webアーカイブ基盤はデータの格納・検索に止まらず，能動的なデータ収集をも司（つかさど）る。10億ページ規模の日本語Web空間の日々の変化をとらえるため，ページごとの過去の更新傾向を学習し，個別の周期で更新チェックを行う可変周期収集を実現した。10年間にわたる累積ページ数は120億を超えている。

Webアーカイブは，学術・文化資産としての価値も高いが，企業の観点からは個人ブログ等に記述された製品やサービスに対する感想や要望，生活における価値観を分析し，意思決定に活用することが期待される。これまで，複数の協力企業とともにマーケティングリサーチへの活用や，CSR（Corporate Social Responsibility）活動の一環としての社会風潮の把握，リアルタイムでの注目話題分析等の応用における有効性を確認してきており，三菱電機インフォメーションテクノロジー㈱（MDIT）による事業化を準備中である。



大規模日本語Webアーカイブに基づく社会動向分析

大規模日本語Webアーカイブは日本語Web空間の情報を絶えず収集し，過去10年にわたる120億ページ以上の情報を蓄積した巨大データベースである（東京大学喜連川研究室内に構築）。Webアーカイブから人々の意識や過去の経緯を読み取ることによって，実社会の動向に関する深い分析が可能になる。

*三菱電機㈱ 情報技術総合研究所（工博）　**東京大学（工博）

1. まえがき

三菱電機では東京大学との産学連携プロジェクト(注1)を通じ，Webを人間の社会活動を捕捉(ほそく)するセンサととらえ，その情報の解析による新たな価値創出を目指すSocio Senseシステム[1]の開発に携わってきた。Web上では企業や行政，著名人から一般人に至るまで様々な主体による情報発信が刻々と行われており，閲覧者の意思決定に重要な影響を及ぼすとともに，商品やサービスの購入が行われ，オークションやアフィリエイト広告が個人の収入源となるなど，直接的な経済活動も盛んである。今やネット上で過ごす時間は実生活の一部とみなす方が自然であろう。

Web上の情報を分析する取り組みは多いものの，広く実用に供されているツールは検索エンジンしかないのが実情である。検索エンジンはインデックス規模の拡大やランキングアルゴリズムの改良を通じ，通常必要とする情報にはほぼ1ホップで到達できるという安心感を与えるに至った。しかし，検索エンジンは本質的には数十億のWebページから検索結果の上位数十件程度をリストアップしているにすぎないため，特定のページに到達することではなく，ある分野について概観することが目的の場合，情報全体の直感的な把握を目指したシステムが望ましい。また，現時点の情報をありのままに提示されただけではそれ以上理解が深まらないこともある。このような場合，過去にさかのぼってどのような経緯で現在の姿に至ったかを示すことが糸口となることが多い。

このような観点から，Socio Senseシステムでは情報構造の大局的な俯瞰(ふかん)と時間変化の追跡を実現するための研究開発を行った。分析の基盤として不可欠となるのが，過去のWeb情報を網羅的に蓄積したWebアーカイブである。

2. 大規模日本語Webアーカイブの構築

2.1 構築実績

Web上の情報は絶えず発生・変化・消失を繰り返しており，過去の状態を復元し，時間変化の解析を可能にするには，Webデータを常時収集(クローリング)し，長期にわたって保存し続ける必要がある。Socio SenseシステムにおけるWebアーカイブは，1999年のWebスナップショット以降，10年間にわたる120億ページ分の情報を蓄積するに至った。**図1**はその累積ページ数の推移である。

世界最大のWebアーカイブ保有機関は米国Internet Archive(IA)であり，1996年以降Web全体を対象に累積1,500億ページを持つとされている。IAは自らを図書館と位置付け，文化資産としてのWeb情報の保全を主眼としているが，Socio Senseシステムでは情報の分析による価値創出を目指し，対象を日本周辺に限定する代わりに，分析手法の高度化，分析結果の検証に注力してきた。1999年に東大でクローリングを開始した際は.jpドメインのみを対象としたが，その後Webページで使用されている文字集合を手掛かりに動的に収集範囲を制御する技術を開発し，.com上の日本語ページまでも含めた収集を実現している。

また，日本の国会図書館によるWARP(インターネット情報選択的蓄積事業)を含め，各国国立図書館を中心とするWebアーカイビングプロジェクトも進められている。これらのプロジェクトでは網羅性に比してコレクションの再現性やテーマ性が重視される傾向にあり，収集対象サイトの選定や収集結果のチェック等で人手を介在させることが多い。

2.2 Webページ可変周期収集技術

Webページの更新頻度は様々であり，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年に一度程度しか更新されないものも多い。実際に多数のWebページを観察した結果は**図2**のようになっており，更新間隔は幅広く分布していることが分かる。

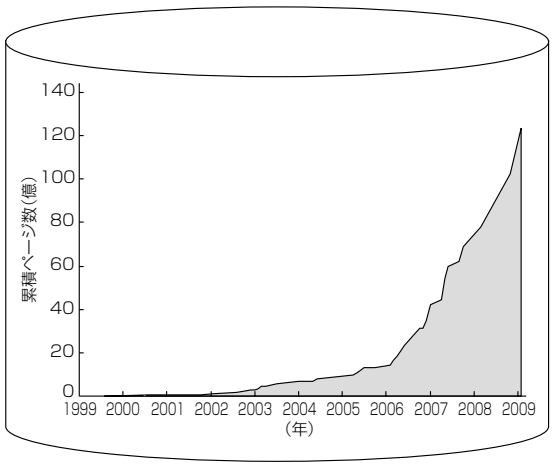


図1．日本語Webアーカイブにおける累積ページ数推移

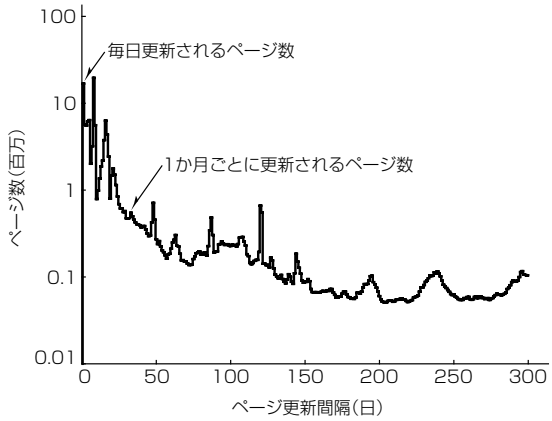


図2．Webページ更新間隔の分布

(注1) 文部科学省"e-Society基盤ソフトウェアの総合開発"(2003～2007年度)

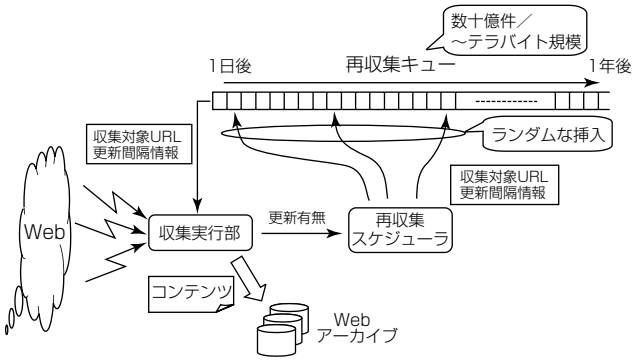


図3．Webページ可変周期収集

Webアーカイビングではweb全体の一括収集を一定期間ごとに繰り返すことが一般的であるが，数十億ページの収集を1か月未満の周期で繰り返すことはシステム性能上難しく，日々の情報の変化をとらえることはできない。そこで，観察した各ページの更新間隔に基づき，ページごとの収集間隔を個別に制御する可変周期収集技術を開発した[(2)]。Webページ可変周期収集では，**図3**に示すように収集対象のURL(Uniform Resource Locator)を収集期日順に保持する再収集キューが中心的な役割を果たす。収集実行部は再収集キューから期日に達したURLを取り出し，Webサーバからのコンテンツ取得とWebアーカイブへの格納を実行する。一方，再収集スケジューラは，取得したコンテンツが以前と異なっているかどうかに基づいて更新間隔を修正した上で，再収集キューの適切な位置に当該URLを再度投入する。

可変周期収集を実際に適用する上では，スケジューラによる更新間隔の推定アルゴリズムよりも，更新間隔情報を含むURLごとの制御情報の管理コストが問題となった。日本周辺の数十億ページを対象とすると，それぞれに対する制御情報の総量はテラバイト規模に達し，従来型のDBMS(Database Management System)では定常的に発生するランダムアクセスによる性能低下が著しかった。開発した日本語Webアーカイブでは，制御情報のデータベース自体をキュー構造とし，キー(URL)に基づくアドホックアクセスのコストと引換えに，定常動作時のアクセス性能を大幅に向上させることに成功した。

3. Webアーカイブに対する分析

3.1 業界連関図の抽出とその追跡

Webの情報構造俯瞰によって，指定したトピックにおける主要プレーヤーや関連トピックの発見が可能になる。Webページを単位とすると全体の把握が難しいため，ページ間のリンクの稠密(ちゅうみつ)性に基づいて関連ページ群をWebコミュニティとしてグループ化し，コミュニティを単位に相互の関連度に基づいて二次元上に配置して視覚化を行っている。

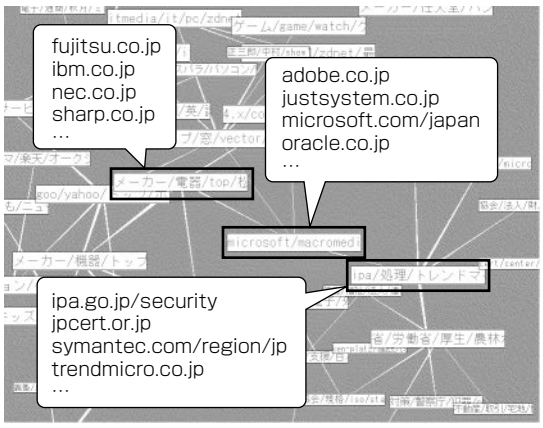


図4．コンピュータ業界における連関図抽出結果

図5．銀行業界の時間推移

トピックとして産業分野を指定すれば，当該分野における企業の連関図に相当する結果が得られる。**図4**はコンピュータ業界に関する連関図を抽出した例である。左上にハードウェアベンダーのサイト群からなるコミュニティが存在し，それと隣接してソフトウェアベンダーのサイト群からなるコミュニティが見つかり，さらにセキュリティ情報及びセキュリティソフトウェアベンダーのサイト群からなるコミュニティに接続されていることが分かる。セキュリティソフトウェアベンダーが一般的なソフトウェアベンダーのコミュニティではなくセキュリティ関連コミュニティに属しているのは，アルゴリズムが周辺ページ(個人ページ等も含む)からのリンクに基づいて導出した結果であり，企業自身の言明より社会的な認知を強く反映した形になっている点が重要である。

図5は銀行業界の中心的なコミュニティに着目し，異なる時点での抽出結果を時間順に左から右に並べたものである。異なる時点のコミュニティを接続しているのは，共通するメンバーサイトである。1段目の行は都市銀行のサイト群からなるコミュニティを表している。2段目を見ると，2001年に新たなコミュニティが発生したことが分かる。このコミュニティに属しているのはこの年に多数生まれたインターネット銀行のサイト群である。この時点では，都市銀行とは明らかに異なる存在として認知されていたと言える。しかし，時間の経過を追って右側を見ていくと，2003年には1段目の都市銀行コミュニティと一体化していることが分かる。このころには一般の都市銀行もインターネット銀行と同様のサービスを提供するようになり，リンクを

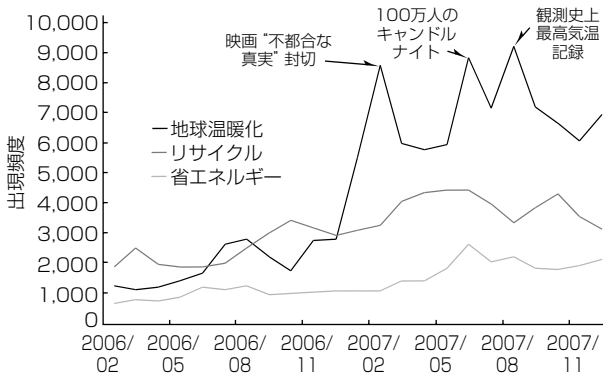


図6．環境問題関連キーワードの出現頻度推移

張る際に明確に区別されることが少なくなったためと考えられる。これは，Webが社会意識のセンサとして有用であることを示している。

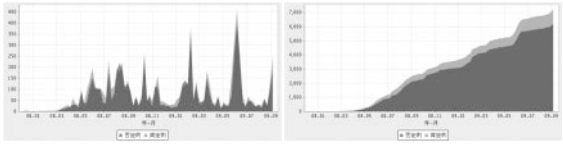
3.2　関心の時間推移分析

Web検索に時間軸を加味すると，特定のキーワードにヒットするページ数がどのように変化してきたかを把握することが可能になる。**図6**は環境問題に関するいくつかのキーワードに対するヒットページ数の推移を示したものである。これらの中で最も出現数が多いのは“地球温暖化”であるが，“リサイクル”を大きく上回るに至った契機は，映画“不都合な真実”の公開であったことが分かる。また，この言葉はこれ以降，夏至や冬至におけるキャンドルナイトのイベントや最高気温の記録更新などのニュースに伴って，最も想起されやすいキーワードとなっていることも読み取れる。また，Web上で人気を博し，実世界の書籍として出版されてベストセラーとなった“生協の白石さん”のケースについても同様に過去の経緯を解析したところ，最終的に多くのリンクを得ていたサイト自体は後発であったことなども分かった。Webアーカイブによって過去の事実を保存しておくことで，ブーム到来後に確認しようとしても難しい経緯の再発見が可能になったと言える。

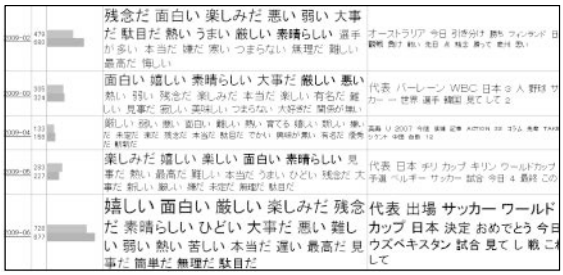
3.3　好不評表現の分析

企業が自社の製品やサービスの認知度を調査する際には，その関心が肯定的なものであるか否定的なものであるかが重要である。ブログ等を通じて現れた個人の主観は，ブログ筆者の嗜好(しこう)を示すだけに止まらず，そのブログを見た多数の閲覧者に影響し，その行動を左右する可能性もある。

Web上のテキスト情報全体から好評・不評表現を自動的に抽出するには，それらの言語表現(“素晴らしい”“今一つ”など)を登録した辞書が必要となる。精度の高い辞書の作成には人手がかかり，Web上の多様な言語表現に対応しきれないことが課題であったが，Webアーカイブ全体から評価表現を学習する手法を開発し，大規模な辞書の自動構築を実現した。**図7**は，構築した評価表現辞書を適用して実施した好不評表現の分析結果を示したものである。



好評(薄)・不評(濃)書き込み数の推移(右は累積値)



好評・不評書き込みの例

図7．好不評表現分析結果

好評表現及び不評表現の書き込み数，及び具体的な表現例を色分けして示している(実際には赤色と青色)。ここではサッカー日本代表に対する書き込みを例に挙げたが，企業では自社の製品やサービス，広報活動に対して定常的に好不評表現の監視を行うことで，不具合に対して早期に対応するとともに，新たなニーズの芽を発見することが可能になる。

4. む　す　び

大規模日本語Webアーカイブとその分析システムについて述べた。10年にわたって集積してきた120億ページ以上のWebアーカイブは，日本のWeb文化の記録であると同時に，社会現象や消費行動を読み解くための手掛かりを秘めた貴重な資産となっている。Webの情報は爆発的な増加を続けており，動画等の非テキストメディアの比重も増しつつある。今後もWebの姿を確実にとらえ，蓄積していくための収集技術及び格納・検索技術の高度化が必要である。

Webアーカイブ分析手法のうち，企業向けに有用なものについてはMDITで事業化の準備を進めており，食品や日用品等のメーカー，広告・マスコミ等の業界に向けた分析を試行してきた。一方Webの情報は多様であり，本稿で述べた分析手法のほかにも検証すべきアプローチは多い。今後，各分野の専門家やユーザーとの連携を通じ，更に分析を深めていきたい。

参 考 文 献

(1)　喜連川　優，ほか：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，情報処理，**49**，No.11，1290～1296（2008）

(2)　田村孝之，ほか：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，J91-D，No.3，551～559（2008）