

ウェブテクノロジーがもたらす社会科学の新しい研究方法論

New Research Methodology for the Social Sciences
through Web Technology

増永良文

Yoshifumi Masunaga

青山学院大学社会情報学部

School of Social Informatics, Aoyama Gakuin University

Abstract: As various individuals and organizations have disseminated information on their Web pages, it is considered that various social events and changes have been mapped onto the Web. It is said that billions of Web pages now exist on the Web. Therefore, it might be possible to accurately know the events and changes of society if we mine or search the Web. The authors have investigated Web community analysis, revealing that Web structure mining could be a new and useful social survey method. In addition, we have been developing the SERPWatcher, which could also be considered as a new social survey method from the standpoint that SERP (search engine results page) and its Web page ranking system might be able to intensively reflect phenomena of the real world. We believe that Web mining and the SERPWatcher are new social survey methods which are totally different from traditional methods such as questionnaires or interviews.

キーワード: 社会科学, 社会調査法, ウェブマイニング, 検索エンジン, SERPWatcher

Keywords: Social science, Social survey method, Web mining, Search engine, SERPWatcher

— 1 — はじめに

ウェブ (the Web) は正式には World Wide Web (WWW と略されることも多い) といい, 文字通り「世界中に張り巡らされた蜘蛛の巣」という意味である。インターネットが世界の隅々まで張り巡らされて, それを基盤とした地球規模のアプリケーションとしての情報共有システムが, 地球全体が蜘蛛の巣で覆われてしまったような状態で実現する様子を表すべく命名された [1]。歴史的には, ウェブは1989年, 当時スイスのジュネーブにある欧州原子核研究機構 (CERN, European Particle Physics Laboratory) に勤務して

いたバーナース=リー (Tim Berners-Lee) により, 地球規模の情報共有のためのインターネットを基盤としたハイパーテキスト (hypertext) による構想として発明された。ウェブを稼働させるために考案された URL (Uniform Resource Locator), HTTP (HyperText Transfer Protocol), HTML (HyperText Markup Language) の最初の設計は彼による。世界で最初のウェブサーバ (<http://info.cern.ch>) は勿論 CERN に置かれたが, ウェブに関するさまざまな技術, これを総称してウェブテクノロジーと本論文では呼ぶ, の発展により, ウェブはさまざまな主体により急速に受け入れられた。

ウェブの最大の特徴は, バーナース=リーの言葉を繰り返すまでもなく, それがハイパーテキストである

ということにある。つまり、HTMLにより記述されたウェブページ（以後単にページということも多い）は一般に個々のページが独立して存在するのではなく、あるページのアンカーテキスト（リンクラベルともいう）から他のページを参照するためのリンク、これをハイパーリンクという、により相互に指しつ指されつの関係で繋がりあい、依存しあっている非線形の構造にある。ウェブページは情報発信のためにさまざまな主体によってアップロードされるが、ウェブを構成する数十億とも数百億ともいわれているウェブページの集合はリンクしあっているから情報が有機的に結びつき意味のある空間、これをサイバースペースという、を構築している。換言すれば、この非線形の構造があるからこそ、ウェブは実世界の輻輳した関係性を（もちろん限界はあるが）表現することができる。

実世界を写し込んだウェブには社会的出来事（social event）や社会的変革（social change）が潜んでいるはずなので、人々はそれらをウェブテクノロジーを駆使して掘り起こし、覗き見たいという衝動に駆られてきた。一つはウェブマイニングであり、もう一つはウェブ検索である。もちろん、ウェブマイニングには幾つかの研究対象があり、ウェブ検索にもさまざまな手法が提案され実装されてきた。大変興味深いことは、ウェブマイニングのなかでもウェブ構造マイニングといわれている分野ではウェブコミュニティの計算にあたりウェブページ間に張られたハイパーリンクの構造が本質的な役割を演じ、一方、ウェブ検索のなかでも広く受け入れられているロボット型検索エンジンの検索アルゴリズムは、GoogleのPageRankにその典型を見るように、ハイパーリンクの存在があってはじめて計算可能である。

ここでもう一つ勘案しなければならない要素がある。それは、実世界は非可逆で止まることのない時間の経過と共に変貌し続けていくことである。つまり、ウェブから社会的出来事や社会的変革を読み解こうとするならば、ウェブのある一瞬だけを切り取って読み解きを図ろうとするだけでなく、時間と共に変容していくウェブの構造なり、検索エンジンが返してくる検索エンジン結果ページ（search engine results page, SERP）が示すウェブページの順位付けの遷移を観察していくことが本質的であるという点である。前者を共時的分析、後者を通時的分析ということにする。

さらに、ウェブではウェブページはハイパーリンクで結びついていて、その繋がりこそ意味があると考えるとき、上記のウェブ構造マイニングやウェブ検索に加えて、バックリンク解析を共時的、通時的に行なうことも意味あることと考えられる。その結果、ウェブページをアップロードした者の素性や関係性を社会的に追うことが可能となり、ウェブ構造分析やSERP分析からは読み解けなかった実世界の姿を知ることが可能となるであろう。

筆者らは、上述のような観点から、「Webコミュニティの動的分析手法を用いたジェンダー研究ポータルサイトの構築」（日本学術振興会平成15年度～平成17年度科学研究費補助金（基盤研究（B）（2））（課題番号15300031）研究代表者：増永良文）や「社会科学の新しい研究方法論としての統合型ウェブマイニング環境の開発研究」（日本学術振興会平成19年度～平成21年度科学研究費補助金（基盤研究（B））（課題番号19300025）研究代表者：増永良文）を推進してきた。その結果、ウェブ構造マイニングや通時的SERP順位分析が、アンケート調査やインタビュー調査といった社会科学でこれまで用いられてきた社会調査法とはまったく異なるその学問分野の新しい研究方法論になりうるのではないか、という知見を得るにいたっている。本論文は、上述の観点から、これまで得られてきた結果を再構成して、ウェブサイエンス[2]の確立に向けて少しく貢献することを目指すとともに、ウェブサイエンスと社会科学の狭間を埋めようとするものである。表1に本論文執筆の視座を示す。

手法 分析	共時的分析	通時的分析
ウェブ構造マイニング	ウェブコミュニティの分析	ウェブコミュニティの発展過程の分析
ウェブ検索	SERPの分析および多検索エンジン間のSERPの比較・分析	単一あるいは多検索エンジンのSERPの変動パターンの分析

表1. 本論文執筆の視座
Table 1. Standpoint of This Paper Writing

以下、第2章で社会科学の新しい研究方法論としてのウェブ構造マイニングを、第3章で社会科学の新しい研究方法論としてのSERPWatcherの開発を論じる。第4章はまとめである。

—2— 社会科学の新しい研究方法論 としてのウェブマイニング

2.1 ウェブマイニング

ウェブを分析する方法の一つがウェブマイニング (Web mining) である。ウェブマイニングには、大別すると、(a) ウェブ構造マイニング、(b) ウェブコンテンツマイニング、(c) ウェブログマイニング、の3タイプがある。ここに、(a) はウェブのページ間に張り巡らされているリンクを頼りにウェブの構造を分析して知見を得ようとする。(b) はウェブページの中身を (何らかの意味解析ツールなどを用いて) 分析して知見を得ようとし、(c) はウェブのアクセスログを分析して、何らかの知見を得ようとする。図1に実世界、ウェブ、そしてウェブマイナー (Web miner) の関係、すなわちウェブマイニングの概念を示す。ウェブマイナーは高度なドメイン知識 (domain knowledge, ある専門分野に特有の知識) を有することが大事で、その結果、ウェブを通して実世界での出来事や動きを「読み解く」ことが可能となる。

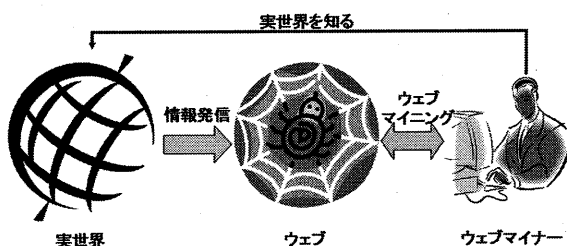


図1 ウェブマイニングの概念
Figure 1. Concept of Web Mining

筆者らは、(a) の典型例として、ウェブページ間のリンク構造を分析してウェブコミュニティ (Web community) を発見するウェブマイニングツール Companion- (コンパニオンマイナス) [3, 4] に着目し、それを用いて我が国のジェンダーコミュニティの共時的・通時的分析を行なった。ここで、ウェブコミュニティとは、基本的に同じトピック (topic) に関心を持つ人々や組織によって形成されたウェブページの集合体を指す。ウェブページ間のリンク構造に着目

してウェブ上のコミュニティを発見するツールの開発は1990年代後半から急激に発展してきたが、代表的なアプローチが、いわゆるハブ (hub) とオーソリティ (authority) ページの関係性に着目した HITS 法 [5] である。筆者らが使用した Companion- はこの HITS 法の改良アルゴリズムというべきもので、東京大学生産技術研究所喜連川研究室の豊田正史氏らが開発した。Companion- が有する機能は大別すると2つあり、一つは抽出されたコミュニティを可視化してその共時的分析を可能とする「ブラウザ」である。もう一つは、コミュニティの通時的分析を可能とする「ビューア」である。Companion- が分析を行なえるためには、ウェブページが定期的に収集されている必要がある (このアーカイビングもウェブページがハイパーリンクで繋がっているから可能であることに注意する。ロボット型検索エンジンの探索ロボットがウェブページをアーカイビングできるのも同じ原理による)。喜連川研究室では1999年以来毎年 "jp" ドメインのウェブアーカイブを構築していたが、それが筆者らのジェンダーコミュニティの分析研究が必要とした時代のアーカイブと一致していた。

2.2 ジェンダーコミュニティの分析

ウェブには社会的出来事や社会的変革が写し込まれているのならば、社会的にセンシティブ性の高いキーワードを用いてウェブコミュニティを抽出してみることはウェブと実世界の関係性を検証するに都合が良いのではないかと筆者らは考えた。そこで、当時我が国で男女共同参画社会の実現に向けて社会で大きく取り上げられていたジェンダー (gender, 社会的・文化的に形成された性) [6] というキーワードに着目して研究を展開した。もちろん、ウェブマイナーとしてジェンダー学際を専門とする研究者が筆者らの研究チームの主要メンバーであったことももう一つの理由である。当時の歴史的背景に若干言及すれば、1999年6月に我が国では「男女共同参画社会基本法」が制定された。男女共同参画は英語では gender-equality と訳されているように、性別によって不利益をこうむることがない平等な社会の実現を目指すためのものである。この法の施行に伴い、多様な動きが全国各地で起こったが、このような表面に出やすく我々がすぐに認

知しやすい社会的現象とともに、一方では顕在化しにくいジェンダー問題も多々あると考えられ、筆者らはウェブをマイニングすることによってそれらを把握することが可能ではないかと考えたわけである。換言すれば、もしその試みが成功すれば、ジェンダーのような社会的にホットなキーワードに対しては、ウェブマイニングは社会の動きを確実に捉えることのできる社会調査の一手法として新たに認知されて然るべきものといえる。

2.3 ジェンダーコミュニティの発展過程

筆者らが使用したウェブマイニングツール Companion-のビューア機能により、時間の経過と共にウェブ

コミュニティがどのように変遷してきたかを分析することができる。図2は、ビューアにキーワード「ジェンダー」を日本語で入力して、2003年3月のジェンダー関連コミュニティ21個を起点に、各コミュニティが1999年から2003年までの5年間どのように変遷してきたかを大筋のところて捉えたもので、ビューアの「Main History」モードの画面を切り取ったものである。コミュニティの発展過程は、共通するURLの数を基軸にして、起点（この場合2003年）から年次ごと段々に過去（左方向）にさかのぼって、脈絡を持たせていくことで横方向に繋がれて表示されている。縦方向の上からのコミュニティの表示順は、コミュニティが保持するキーワードの集合体において、該当キーワードの頻度が高いほど上位に表示されるよ

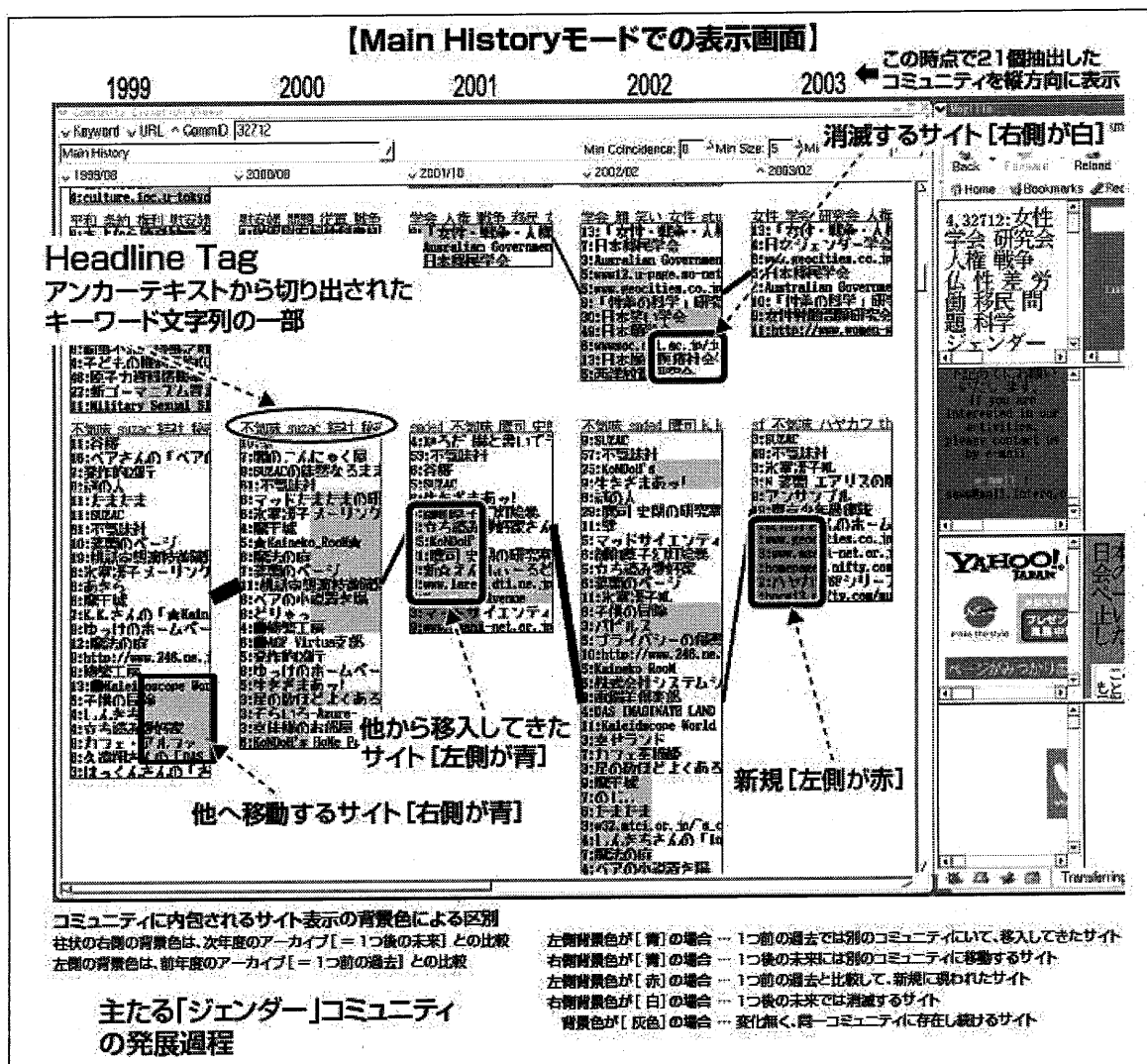


図2. 主たる「ジェンダー」コミュニティの発展過程
Figure 2. Growing Process of a Major Gender Community

うになっている。コミュニティを特徴付けるキーワードの集合体はそのコミュニティを構成するいずれかのウェブページを指すリンク元（＝バックリンク（backlink）先）のウェブページのアンカーテキストから切り出されたもので構成されている。柱状に表されたコミュニティは左右に分かれ、新規〔赤〕、消滅〔白〕、移動〔青〕によって色分けされており、過去から未来に向けたサイトの変遷が見やすくなるように工夫されている。Companion-は2003年に我が国にジェンダーコミュニティが21存在すると表示し、筆者らはその21のコミュニティを詳細に分析して、その妥当性を確認した。これら21のコミュニティはリンク構造から計算されたものであって、実世界を調査して得られたものではない。すなわち、ウェブマイニングツール Companion-はこの観点からも社会調査の機能を果たしている。

2.4 女性センターの発展過程

図3は、ビューアにキーワード「ジェンダー」を日本語で入れることで得られた2003年2月から過去5年にわたる「女性センター関連コミュニティの発展過程」を特に示したものであるが、これは、実世界のジェンダーコミュニティでよく知られている現状を、ウェブコミュニティはどのように映し出したかという典型的な例といえる。このウェブコミュニティが発展した様子は、「女性センター」という施設が、各都道府県ばかりでなく各市町村のレベルまで、相互視察によって横並び状態で次々と造られていき、連携を図ってきた現実とぴったりと重なった。1999年6月に施行された男女共同参画社会基本法の影響で、各自治体では複数の部局にまたがっていた女性施策部局が、実務部局としての「女性センター」に収斂される動きが加速度的に進んだ。より詳細にサイトの移動を描き出す「Detailed History」モードで履歴をたどると、このことが全国規模で起こったことが再確認できた。また、共通するURLでコミュニティを繋ぐこのビューアは、ページのタイトル表記の変化の認知に威力を発揮した。タイトルを通時的に辿ると、「婦人」をやめて「女性」に変えていく様子や、「女性センター」が「男女共同参画センター」に取って代わられていく様子も確認できた。そして、「女性センター」のコミュニティ

が、「ジェンダー」のキーワードから抽出されたコミュニティ21個（2003年3月時点）の中で、最下位に配置されている理由も現実の世界を反映したものとして読み解ける。「ジェンダー」概念は「男女共同参画社会基本法」の審議段階で基本理念となっていたが、最終的にはさまざまな理由で基本法から削られてしまう。「女性センター」のような設立母体が公的機関のウェブページにおいて、「ジェンダー」を使用しないという傾向はそのような政策に直結していたわけである。「ジェンダー」の出現頻度が低く、結果としてビューアで「女性センター」のコミュニティが最下位に配置されている理由は、まさしくこのような社会の動きをそのまま映し込んでいると解釈できる。ウェブマイニングツール Companion-はこの観点からも社会調査の機能を果たしている。

2.5 ツールが抽出した事実 —セクハラコミュニティの履歴と分裂—

図4は2003年時点での異なる「セクハラ」関連の2つのコミュニティの発展過程を上下に並べたものである。この2つはコミュニティの履歴が重複している、つまり過去において一つのコミュニティを構成していたウェブページ群が、2003年2月時点で分裂して2つになったということになる。より詳細にページの移動を描き出す「Detailed History」モードで履歴を辿ると、よりはっきりしたことがわかる。2003年2月時点にはセクハラ関連のページの中で大学関係のセクハラ関連ページだけが、いくつかの大学関係ページを他のコミュニティから取り込んで、「大学関係のキャンパス・セクハラ・コミュニティ」としてまとまっている。一方、それに加わらなかった個人的ウェブページ群は新たにセクハラ問題を扱う個人のウェブページを迎えて、別のセクハラコミュニティを作ったと解釈できる分裂である。実社会において2003年は各大学においてセクハラ相談室の設置が騒がれた時期で、大学間のネットワークが形成された時期にあたる。キャンパス・セクハラとして社会問題化する以前は、さまざまなセクハラ問題が同一ページで扱われていた。その頃はセクハラ問題自体が細分化された状態で認知されてはいなかったということになる。そのことを Companion-は見事に描き出している。これは実

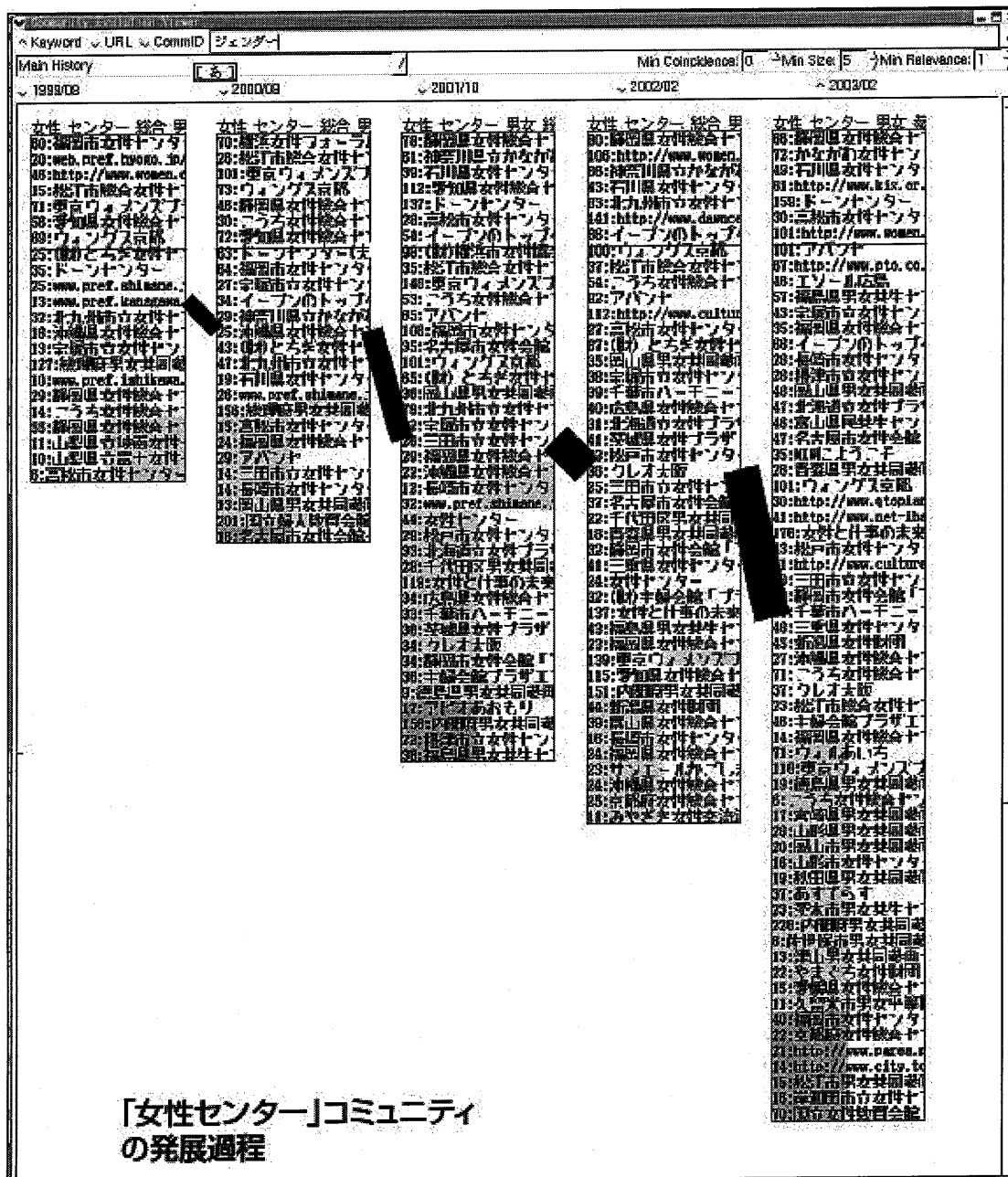


図3. 「女性センター」コミュニティの発展過程
Figure 3. Growing Process of Women's Center Community

世界を見ていただけでは、はっきりと確認できなかった興味あるジェンダー関連コミュニティの動向であった。ウェブマイニングツールの有用性が専門家により認められた好例と考えられる。ウェブマイニングツール Companion-はこの観点からも社会調査の機能を果たしている。

2.6 ウェブマイニングの意義

本章で示した事柄は筆者らがこれまで行なってきた一連の研究結果 [7,8,9,10,11] のエッセンスであるが、その意義は大別すると2つある。一つは、ウェブマイニングという手法が社会科学の新しい研究方法論となり得ることを実証してみせた点である。これは発見であり、今後社会科学を中心に幾多の潜在的適用可能分野に適用されて、新しい知見を生み出していくで

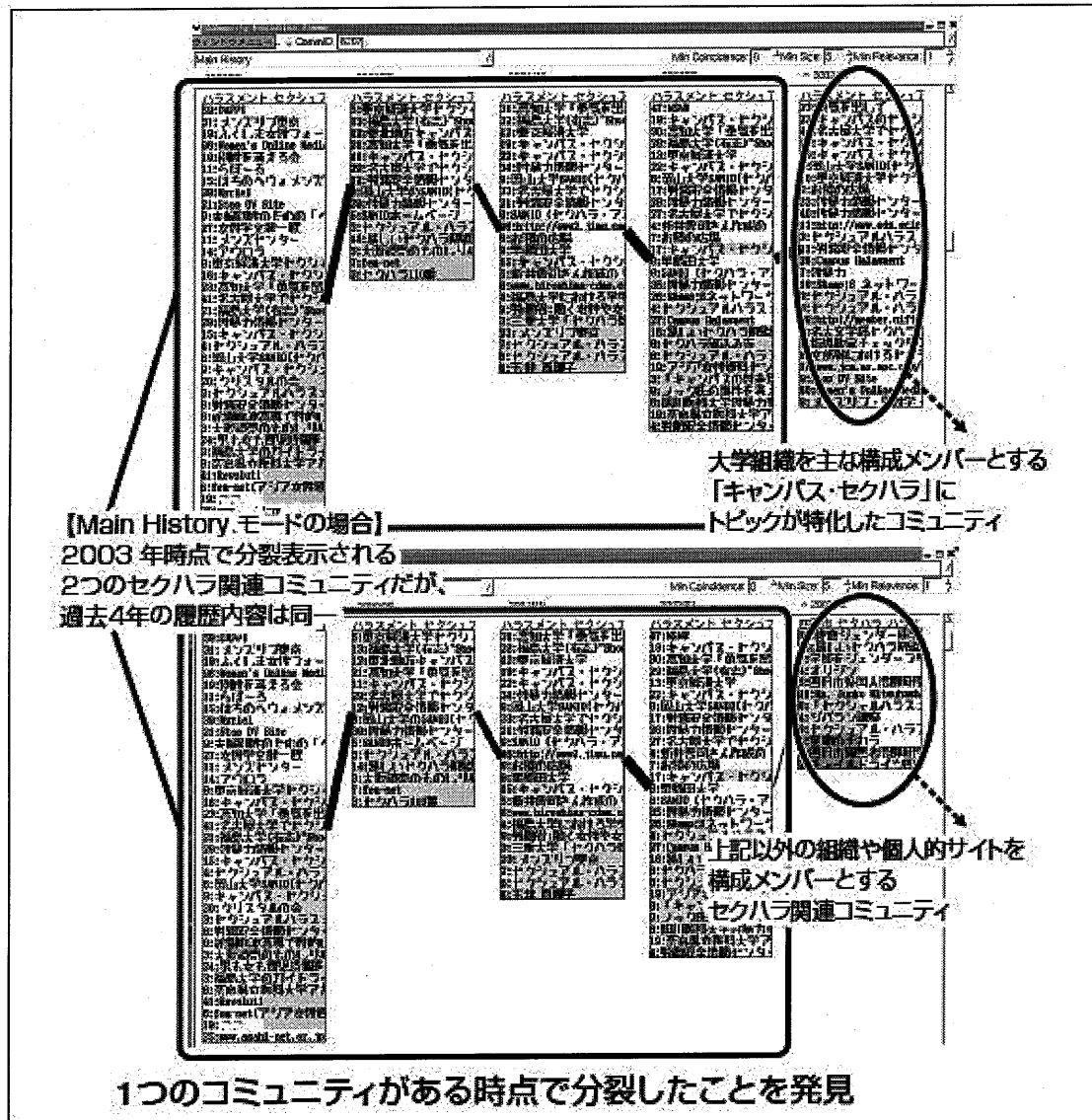


図4. 同根コミュニティ系列の分化の発見
Figure 4. Discovery of a Splintered Community Sequence

あろう。もう一つは、ウェブマイニングツールをドメイン知識が豊富にある「読み手」が鍛えるという視点である。すなわち、システム開発者はそのシステムが適用されるドメインの知識を持ち合わせていないのが通例であり、したがって開発したシステムが真に有効・有用なものであるかを知るすべがない。ドメイン知識を豊富に有する者がそのツールを使い心血を注いでマイニングをすることによってのみそれを知ることができ、ウェブマイニングツールの改良につながる適切な要求や助言をあたえることができる。

— 3 — 社会科学の新しい研究方法論 としての SERPWatcher

3.1 検索ポータルサイト

ウェブを構成している莫大なウェブページ群から利用者の検索要求に合ったウェブページ群を検索してくれるウェブサイトのことを検索ポータルサイト (search portal site) という。ポータルとは「玄関口」という意味であり、ウェブで情報検索をしたいなら、まずそのサイトを訪れなさいという意味である。検索

ポータルサイト（以後、単に検索サイトということも多い）には、大別するとロボット型とディレクトリ型の2種類がある。Google は前者の典型であり、Yahoo! は後者の典型である。ここでロボット型検索とディレクトリ型検索の特徴を簡単に述べておくと、ロボット型検索ポータルサイトでは、検索の対象はウェブページであり、検索手法はコンピュータアルゴリズムによる。Google は約80億ページと公表しているように、探索ロボットが収集した数十億のウェブページが検索の対象となり、利用者は検索キーワードを自由に入力できる。しかしながら、多数の検索結果が表示されるが必ずしも所望のページが上位に表示されているとは限らなかったり、本来重要度の高くないページがSEO（search engine optimization, 検索エンジン最適化）技術が駆使されるなどして高順位に検索されることがあるなど、検索精度上の問題が指摘されている。一方、ディレクトリ型検索ポータルサイトでは、検索の対象はウェブサイトであり、検索手法は人手による分類を反映したディレクトリ対応による。従って、予め精選されているので検索精度が高い反面、検索の対象となるサイトの数が限られている、あるいは検索は予め決められたカテゴリ階層を辿らなければならないなどの問題がある[12]。筆者らは、さまざまな主体がさまざまな観点からウェブに情報発信しているがゆえに、検索エンジン結果ページ（search engine results page, SERP）を読み解くことで社会の実情を把握できるとの見識から本章で述べる SERPWatcher を開発している。したがって、情報が恣意的選択の後に検索の対象となるディレクトリ型ではなく、自由に情報発信されたウェブページそのものがすべて検索の対象となるロボット型検索ポータルサイト（が提示する SERP）を研究対象とする。以降、検索サイトあるいは検索エンジンといえば特に断らない限り、ロボット型のそれを指すものとする。

SERP で最も重要なファクターはその順位付け（ranking）である。これに関しては「Google の黄金の三角形」として興味深い研究結果が知られている[13]：Google で SERP の上位10位を一画面に表示した場合、1位～3位に位置したウェブページは利用者の100%の視認率を得るが、7位で50%となり、10位となると視認率は20%に下がる。つまり、ウェブページは SERP 順位の7位までに入らないと半分以上の

利用者は見てくれないということである。Google は1998年に当時スタンフォード大学の院生であったページ（Larry Page）と布林（Sergey Brin）により設立された。彼らの学位論文が、ウェブ検索エンジンについてであり、検索キーワードの出現回数が多いウェブページを検索結果の上位に表示する従来の方法よりは、ウェブページが多数のバックリンクを有している場合や、数は少なくとも高いランクを持ったウェブページからバックリンクを有している場合（例えば、Yahoo！（のホームページ）からリンクが張られている場合）の方が重要なページであるという尺度、これを PageRank という、を使うほうがより良い検索結果を提示するであろうという発想に基づき起業されている。もちろん、現在の Google の検索エンジンは、ウェブページの重要度を計算するにあたり、バックリンクの数だけでなく、実にさまざまな要因を勘案・処理して検索結果を求めて表示していると公表している[14]が、Google の SERP 順位計算の基本は上述の通りである。

一般にロボット型検索ポータルサイトがシステム全体としてどのように機能しているかを図5に示す。クローラ（crawler）とか蜘蛛（spider）とか称されるウェブページ探索ロボット（＝ウェブページ巡回プログラム）をウェブ上に多数放ち、ウェブページ間のリンクを辿って、定期的に世界中のウェブページを収集してくる。Google は現在80億ページに上るウェブページを収集し、主要なページについては連日更新が行なわれているという。収集されたウェブページはデータベース化され、ページに現れる重要語をキーワードとして索引（index）が付与され、リンク関係の情報も保持される。利用者は検索ポータルサイトのトップページの検索窓（window）に検索キーワードを入力して検索を要求すると、検索エンジンがコンピュータアルゴリズムにより SERP を計算して、それを利用者に表示する。

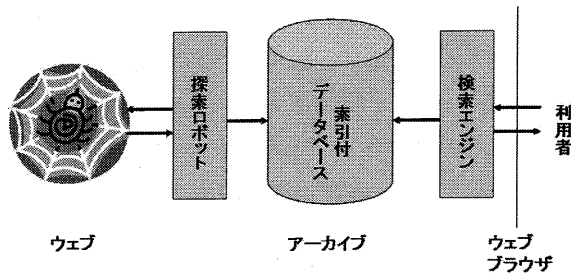


図5. ロボット型検索ポータルサイトの仕組み
Figure 5. Scheme of a Robot-type Search Portal Site

3.2 SERP の順位変動と社会の動き

検索ポータルサイト Google の検索アルゴリズムはウェブページ間に張られたハイパーリンクの構造が本質的な役割を演じ、PageRank の考えのもとに SERP を計算していることは前述した。また、SERP の順位は単に PageRank の値のみならず、さまざまな要因を勘案して決定されると Google 自身が表明していることも前述した。そのような Google の検索アルゴリズムの裏をかくて、作成したウェブページを SERP の上位に位置させるための検索エンジン最適化 (SEO) 技術に関する話題にも事欠かない。

さて、筆者らは、そのように複雑なアルゴリズムや背景のもとで SERP が計算されるとしても、検索の対象となったウェブには実社会が写し込まれているのであるから、SERP の共時的・通時的分析を通して、社会の変化が読み取れるはずであるとの見識のもと

SERP 分析ツールの開発を行なってきた。図6は Google で、検索キーワードを「ジェンダーフリー」とし、2007年8月5日から1週間間隔で連続して16週間収集した SERP の中から、収集し始めた時点での SERP で19位と23位のジェンダー擁護派のウェブページ、SERP で14位と17位のジェンダーバッシング派のウェブページのその後の順位変動を表示したものである。興味深いことは、2007年9月12日に内閣総理大臣を辞任した安倍晋三はジェンダー嫌いとして知られ、一方、同年9月26日に内閣総理大臣に就任した福田康夫は内閣官房長官時代からジェンダー擁護派として知られているが、この福田康夫が次期内閣総理大臣候補として有力となった丁度その頃の同年9月16日に収集した Google の SERP にみられるように、ジェンダー擁護派のウェブページの順位が一様に上昇し、ジェンダーバッシング派のウェブページの順位が一様に下降している。(特に、同年8月5日の時点で17位だったジェンダーバッシング派のウェブページは同日一気に20位ほど順位が落ち、そのまま定位した。) これらは、筆者らをして SERP の順位変動には社会の動きが写し込まれると確信させるに十分な現象であった。

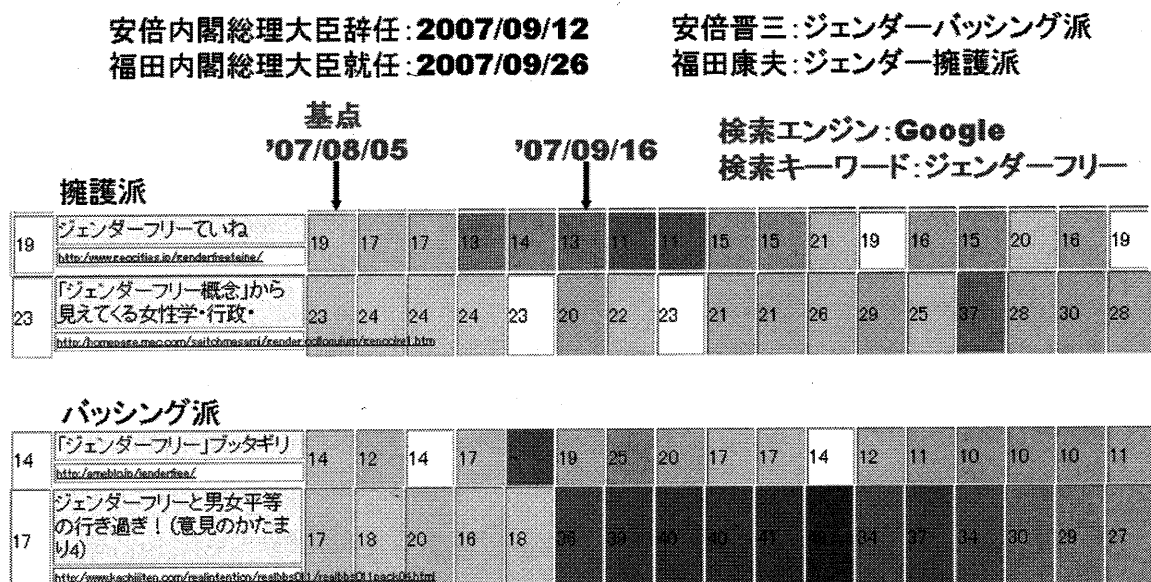


図6. SERP 順位に写し込まれる社会変化
Figure 6. Social Changes mapped into SERP Ranking

3.3 SERPWatcher

3.3.1 SERPWatcher とは

SERPWatcher は、正式には Search Engine Results Page Ranking Change Watcher と称するが、以下に示す事項を考慮して実現される SERP 順位変動監視ツールである。

- (a) 同じ検索キーワードを入力しても、検索エンジンが異なると SERP 自体が一般には異なる。これは、探索ロボットや SERP の計算アルゴリズムが検索エンジンにより異なるためである。したがって、SERPWatcher 構築の我々の目的が「新しい社会調査法」の提供にあるならば、一つの検索エンジンに固執するのではなく、さまざまな検索エンジンについては皆、その SERP に注目しなければ、ある特定の検索エンジンの SERP に踊らされたり、あるいは社会の変化を見逃してしまうことにつながりかねず、原則として全検索エンジンの SERP を監視の対象としない。
- (b) 現在、Google や Yahoo!Search を始め、数多くのロボット型検索ポータルサイトが運営されている。しかしながら、それら全ての検索サイトが独自の探索ロボットや検索エンジンを有しているわけではない。図 7 に 2008 年 12 月時点のロボット型検索エンジンの相関図を示す [15]。つまり、独立した検索エンジンを有する検索サイトは、Google, Yahoo!Search, Live Search

(旧 msn), Baidu などである。ちなみに Livedoor 検索サイトを訪れると、検索窓の横に「powered by Google」と記述があり、Livedoor は Google の提携検索サイトであり、Livedoor が表示する SERP は Google が提供するものであることを謳っている。したがって、提携検索サイトは SERPWatcher の監視対象としなくてもよいことになると考えられる。しかしながら、筆者らがこれまで行った検索サイトの信用性 (trustworthiness) に関する研究によると、Google がその提携サイトに提供した SERP とは異なる SERP を表示し、その信用性に著しい問題があった事例を発見し報告しており [16,17,18]、独自の検索エンジンを有する検索サイトの SERP だけを監視の対象としているわけにはいかず、SERPWatcher は Google, Yahoo! Search, Live Search, Baidu に加えて Google 提携サイトの goo と Yahoo!Search 提携サイトの excite と Infoseek 楽天の SERP も収集している。(Google と Yahoo!Search による提携検索サイトの争奪合戦は熾烈を極めており、2008 年 7 月時点では Infoseek 楽天は Google の提携検索サイトであった [19].)

- (c) SERPWatcher は利用者（ここで想定しているのは、社会の動きを鋭敏にキャッチしたいと考えている社会科学の研究者）が指定した検索キーワードについて、複数検索エンジンの SERP を定期的に収集し分析でき、SERP の遷移に何らかの異

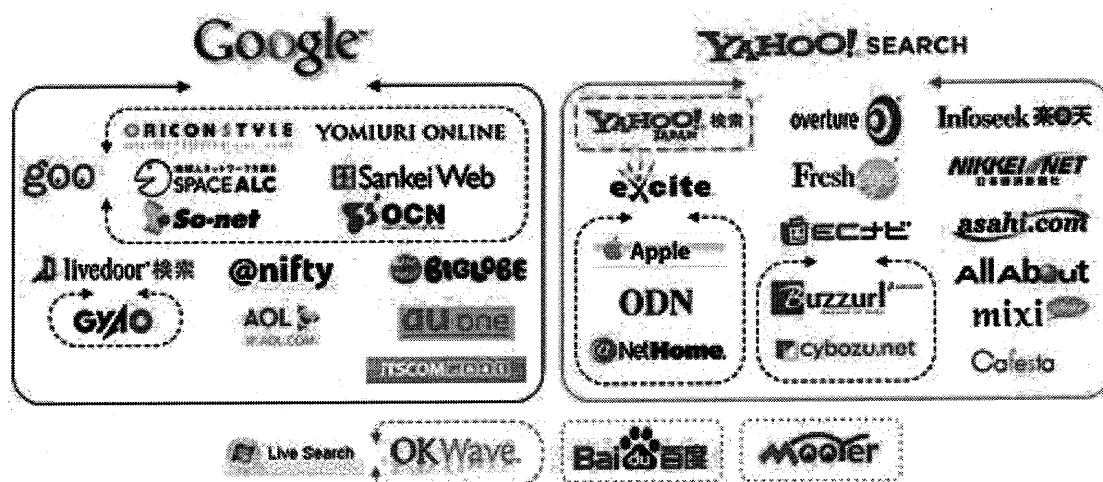


図 7. ロボット型検索エンジンの相関図
Figure 7. Correlation Diagram of Robot-type Search Engines

変を検出したならば直ちにその異変を利用者に知らせるアラーム機能を持たねばならない。

3.3.2 SERPWatcher のシステム概念

上述の議論から, SERPWatcher は次に掲げる機能を有していないといけない。ここに SERP データとは, SERP それ自体のデータ (SERP を構成する各ウェブページの順位, タイトル, スニペット (snippet, タイトルの下に記述されたウェブページの説明文), URL) ならびにそのウェブページに関連するデータ (バックリンク (の総数と個々の URL)) をまとめていう。他に, SERP 収集日の Google ニュースを採取している。

1. SERP データ収集・管理・表示機能
2. SERP データ多次元分析機能
3. アラーム機能

図8に SERPWatcher のシステム概念図を示す。

SERPWatcher は, 利用者により指定された検索キーワードの SERP データ (上位500位まで) を全ての検索エンジンについて一定間隔 (default は1週間) で収集している。また収集されたデータを一元的に管理するためにリレーショナルデータベースを構築して

いる。管理しなければならないデータ群は次に示すとおりである。SERP 収集日の Google ニュースはファイルに別途格納している。これは利用者が社会科学的分析を行う際に必要なデータである。

1. 利用者
2. 検索キーワード
3. 検索エンジン
4. SERP 収集日
5. SERP データ

3.3.3 SERP データ多次元分析機能

SERPデータを中核にして構築されるSERPWatcherデータベースには次の特徴がある。

SERPWatcher データベースは (利用者, 検索キーワード, 検索エンジン, SERP 収集日, SERP データ) で表される5項組を基本として構築され, そのデータは定期的に収集・蓄積され, 更新されることはない。

したがって, SERPWatcher データベースは多次元データベースを構成することになり, OLAP (Online Analytical Processing) [20] を行うのに適している。それにより, SERP データの多次元的分析が可能とな

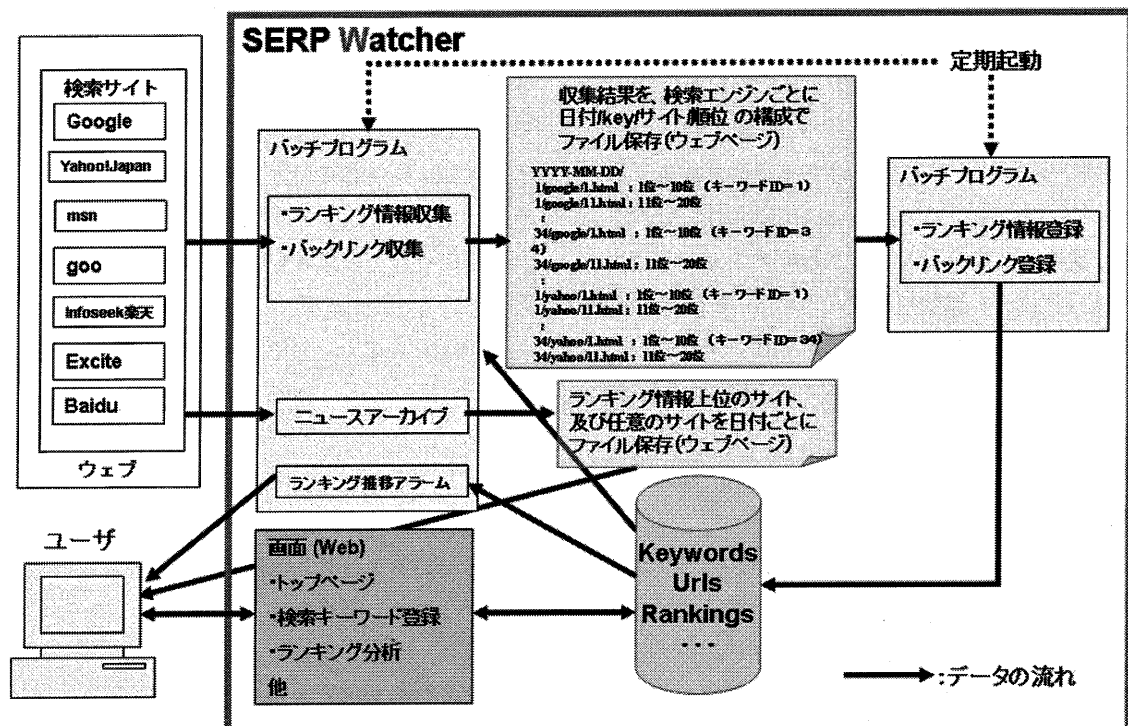


図8. SERPWatcher のシステム概念図
Figure 8. Conceptual Diagram of SERPWatcher

る。OLAP には ROLAP と MOLAP という2つの実現方式があるが、SERPWatcher はデータをリレーショナルデータベースとして組織化しているので、リレーショナルデータベースに格納されたデータを直接検索・集計し、結果をクライアント側で多次元データとして構成し視覚化する ROLAP 方式で OLAP を実現することができる。

3.4節で述べるように、SERPWatcher β版が現在提供している利用者のための標準インタフェースでは、図9に示されるように、SERP データは利用者を固定すれば4次元のデータベースであり、したがって、検索エンジンと検索キーワードが指定されると、縦軸がウェブページを、横軸が SERP データ収集日を表し、交点 (= 値) が順位を表す2次元の表として表示される。つまり、(ウェブページ, 収集日) → 順位, という関係性を表示する。しかしながら、分析を行っていく過程で、たとえば収集日と検索キーワードを指定して (ウェブページ, 検索エンジン) → 順位, という関係性を一覧したいとか、あるいはウェブページと検索キーワードを指定して (検索エンジン, 収集日) → 順位, という関係性を一覧しつつさらに分析を続けたいという要求が起こるはずで、これに 대응するのが SERP データ多次元分析機能である。

3.3.4 SERPWatcher のアラーム機能

利用者が指定した検索キーワードにより定期的に SERP を収集していくとき、SERP 順位の時系列的 (= 通時的) 変動に一定基準を超える変動が観察された場合、それを SERPWatcher が自動的に検出して、利用者に実世界での調査を促すためのアラームを鳴らす機能を「アラーム機能」と呼んでいる。SERP 順位の変動がどのような条件を満たしたときにアラームを鳴らすべきか研究し、利用者がドメイン特有な観点からそれを自由に指定できるための「アラーム記述言語」を設計し実装する必要がある。その際、次の点を考慮すべきである。

SERPWatcher は SERP を一般に複数の検索エンジンで収集しているので、単一の検索エンジンの SERP 順位の変動だけに注目してアラームを鳴らすのか、そうではなくいくつかの検索エンジンで同時にアラーム条件が「真」となった場合にアラームを鳴らすのか、いずれの場合もその条件

を利用者が自由に指定できる記述能力を有すること。

3.4 SERPWatcher のプロトタイピング

筆者らは、現在 SERPWatcher β版 (ソフトウェアの性能や機能、使い勝手などを評価する段階の版) を開発中である。システムはクライアント・サーバ方式で開発されている。クライアントに要求される機能は Internet Explorer や Firefox などのウェブブラウザだけなので、リモート表示方式の実現法となっている。サーバのオペレーティングシステムは Red Hat Enterprise Linux 5.1 (x86/x86_64)、開発言語は Ruby 1.8.5 (2006-08-25) [i386-linux] と Ruby on Rails 2.1. で、リレーショナルデータベース管理システムは MySQL 5.0.45 を使用している。

図9に SERPWatcher の SERP データ表示画面のスナップショットを示す。これは、利用者が検索キーワードを「ジェンダーフリー」、検索サイトを「Google」、ランクを「1位から40位まで」、日付を「2009/02/07から2009/06/14まで」と指定した場合の SERP 順位の変動を表わしている (SERP データは2007/07/29より1週間間隔で自動的に収集されている)。画面をスクロールすることにより、隠れた部分を閲覧することができる。利用者の視認性を向上させる目的で、ウェブサイトの順位枠は着色されている。この例では、基準日として指定した直近の2009/06/14での順位と比較して、同一であれば白抜きで、それより上位であれば赤色系に、下位であれば青色系にそれぞれ深みを増して表示されるように実装されている。図9が示すように、検索エンジン窓のプルダウンメニューで Google, Yahoo!Japan, Live Search (旧 msn), Baidu, goo, excite, Infoseek 楽天の選択が、キーワード窓のプルダウンメニューで検索キーワードの選択が (現在37個の検索キーワードが登録されているが、随時更新可能である)、ランク窓および日付窓でそれぞれ閲覧したい順位と収集日の範囲を指定できる。利用者は、SERPWatcher のこの標準インタフェースを通して、閲覧しているウェブページ群の順位変動を読み解き、社会の動きを捉えようとする。

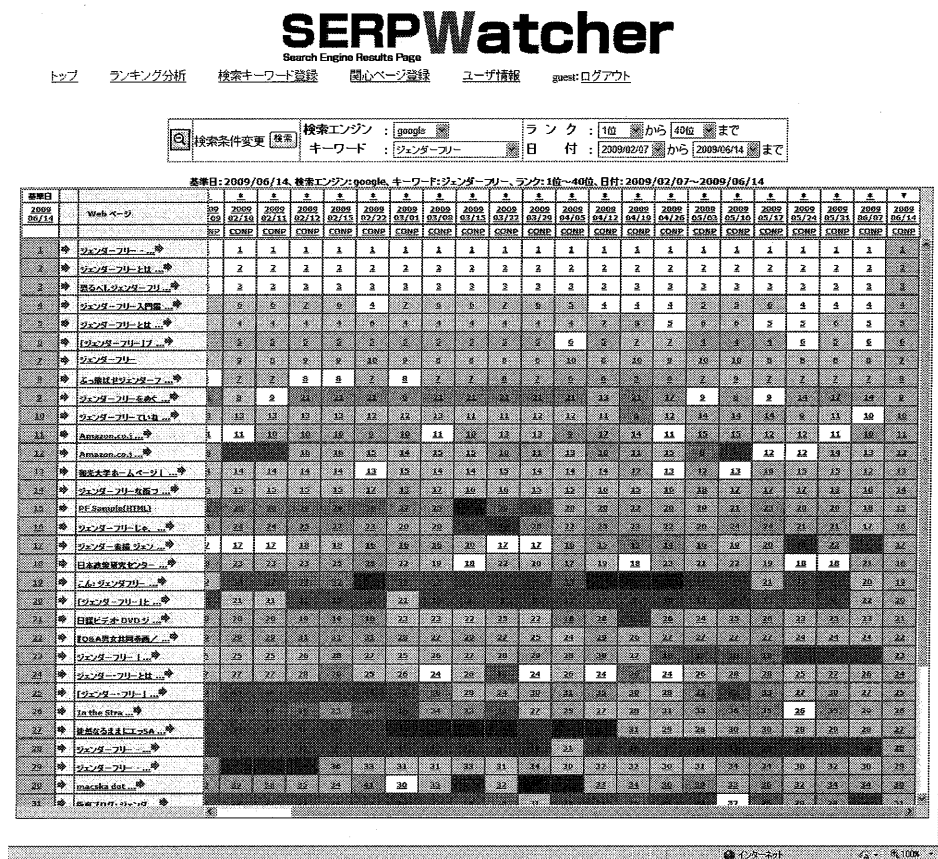


図9. SERPデータの表示例(縦軸: ウェブページ、横軸: 日付、対象: 順位)
Figure 9. A Snapshot of SERP Data Display (Vertical Axis: Web Pages, Horizontal Axis: Date, Subject: Rank)

3.5 SERPWatcher の意義

SERPWatcher は研究・開発の途上にあり、本章で示した結果は筆者らがこれまで行ってきた研究成果 [21,22] の概要である。今後の研究課題は次の3点に集約される。

1. SERP データ多次元分析機能の実現
2. SERPWatcher のアラート機能の実現
3. SERPWatcher を用いた社会科学的発見

1. は SERPWatcher の β 版ではほぼ実現されている。明らかに、2. と3. の問題は表裏一体の関係にある。SERP にどのような変動が現れたときにアラームを鳴らすべきかという問題は、逆に社会にどのような変化があったときに SERP にどのような変動が観察されるのかという問題と同値であるからである。

ここで SERPWatcher の意義を社会科学の研究方法論の立場からまとめる。いうまでもなく、社会科学 (social science) は社会現象を対象として研究する科学の総称であり、政治学・法律学・経済学・社会学・

歴史学・文化人類学およびその他の関係諸科学を含む(広辞苑)。社会からデータを取得する方法には、調査や観察などが知られているが、社会に住む人々の意識や行動などの実態をとらえるために社会調査 (social survey) がある。社会調査は大量のデータを取り社会の全体像を把握することを目的とする統計的社会調査と事例的社会調査の2つに大別される。前者を量的調査、後者を質的調査と呼ぶことも多い。これらの社会調査は、具体的には、アンケート調査やインタビュー調査などさまざまな調査法により行われる。

一方、本論文で示した SERPWatcher を用いた社会調査の手法は、さまざまな SERP を読み解くことで社会的出来事や社会的変革を見つけたり明らかにしようという考え方から編み出された手法であって、従来型の社会調査法とはまったく異なる新しい手法である。(これまで、インターネットを使った社会調査法としてウェブ調査が知られているが、これは調査票をウェブで公開して、回答を募る調査法であり、同じくインターネットを使っているが、SERPWatcher が主

張している手法とはまったく発想が異なることを付記する。) 従来型との比較という観点から SERPWatcher による社会調査の特徴を述べれば次のようになる。

- (a) 調査対象は直接に実世界の個人、世帯、グループ、組織体などではなく、それらがウェブに情報発信した結果としてのウェブページの集合体である。しかしながら、読み解きの結果、明らかになる事柄は、まさしく実社会での社会的出来事や社会的変革である。
- (b) 調査時点でアーカイブされているウェブページとそれら間のハイパーリンクの全体が調査の対象となるので、回収率は100%であり、その意味での回答の信頼性を問題にする必要がない。
- (c) ウェブ上の意見表明を利用して社会調査を行う場合、そのバイアスを補正する必要がある [23]。しかしながら、SERPWatcher の研究・開発の目標は、社会的出来事や社会的変革を SERP の順位変動を読み解くことで見つけることにある。筆者らがこれまで行ってきた事例研究によれば、ジェンダー問題にその典型をみたように、社会的にホットな話題に対して関係者や関係組織はこぞっていち早く敏感に反応して情報発信しており、その意味ではこれまでのバイアス補正問題の直接の対象とはなりづらい。

21世紀はコンピュータとインターネットに裏打ちされたウェブ社会であり、社会調査法そのものも劇的な変貌を遂げて然るべきときであろう。

— 4 —

本論文のまとめ

本論文では、ウェブテクノロジーがもたらす社会科学の新しい研究方法論について筆者らが行ってきたこれまでの研究の要点を論じた。

ひとつは、第2章で論じたウェブコミュニティの動的分析手法とその結果である。そこでは、ドメイン知識のあるウェブマイナーがウェブコミュニティを特に通時的に分析することにより、実世界を単に観察しているだけでははっきりと認識できなかったコミュニティの変遷を的確に捉えることができた事例が紹介されている。研究の対象となった社会科学のドメインは

ジェンダー学際であり、ウェブコミュニティの抽出ツール Companion- を使用して、男女共同参画社会基本法のもとで我が国の「女性センター」が時間の経過と共にそのように発展していったかが、如実に捉えられている。従来のアンケート調査やインタビュー調査ではここまで正確にかつ網羅的にその発展過程を捉え切れなかったのではないかと考えられる。加えて、セクシャルハラスメントのコミュニティが分化していく様子も見事に捉えられ、ジェンダー学際での新しい発見がウェブマイニングによってなされたといえる。

もう一つは、第3章で論じた SERPWatcher の研究・開発である。SERP とは検索キーワードの入力に対して Google などの検索エンジンが返す順位付のウェブページ群をいうが、SERP にはウェブコミュニティと同様、社会的出来事や社会的変革が写し込まれていると考えられる。したがって、SERP を分析することにより、それらを知ることができると考えられ、これも従来型のアンケート調査やインタビュー調査とはまったく異なる新しい社会調査法となると考えられる。SERPWatcher の開発は現在β版の段階にあり、それを使った社会科学分野の新しい発見を報告するまで研究が進んだ状況には至っていない。しかしながら、ジェンダー学際を対象にこれまで収集した1年半以上の SERP データアーカイブを用いた予備実験によれば、SERPWatcher の標準インタフェースを通して見る SERP 順位はダイナミックに変動しており、その変動パターンを的確に読み解くことにより、その裏に潜む社会の変化を的確に炙り出せる時はそんなに遠くない予感がする。

謝辞

本論文は増永の単著としたが、ここで論じた事項は以下に示す方々との共同研究によるところが多く、ここにそれらの方々の氏名を記して深く謝意を表す：お茶の水女子大学ジェンダー研究センター小山直子客員研究員、同館かおる教授、お茶の水女子大学理学部渡辺知恵美講師、東京大学生産技術研究所豊田正史准教授、同喜連川優教授、青山学院大学社会情報学部伊藤一成助教、青山学院大学情報科学研究センター竹内純人助手、青山学院大学大学院社会情報学研究科博士

後期課程深山鷹一君（肩書きは全員現在）他。

第2章で述べた結果の一部は日本学術振興会平成15年度～平成17年度科学研究費補助金（基盤研究（B）（2））（課題番号15300031）「Web コミュニティの動的分析手法を用いたジェンダー研究ポータルサイトの構築」（研究代表者：増永良文）の補助によって行

われた。また、第3章で述べた結果の一部は日本学術振興会平成19年度～平成21年度科学研究費補助金（基盤研究（B））（課題番号19300025）「社会科学の新しい研究方法論としての統合型ウェブマイニング環境の開発研究」（研究代表者：増永良文）の補助によって行われている。

文献

- [1] Tim Berners-Lee: "WEAVING THE WEB: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor," (book) 240p., HarperOne, September 1999.
- [2] Tim Berners-Lee: Hearing on the "Digital Future of the United States: Part I -- The Future of the World Wide Web," <http://dig.csail.mit.edu/2007/03/01-ushouse-future-of-the-web.html>
- [3] M. Toyoda and M. Kitsuregawa: "Creating a Web Community Chart for Navigating Related Communities," Proc. Hypertext 2001, pp.103-112, 2001.
- [4] 福地健太郎, 豊田正史, 喜連川優: "Web Community Browser: 大規模 Web コミュニティチャートの可視化," 第13回データ工学ワークショップ (DEWS2002) 論文集, 2002.3.
- [5] J. M. Kleinberg: "Authoritative Sources in a Hyper-linked Environment," Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.
- [6] 館かおる: "ジェンダー概念の検討," ジェンダー研究, 1号 (通巻18号), pp.81-95, お茶の水女子大学ジェンダー研究センター, 1998年3月.
- [7] 増永良文, 小山直子: "ジェンダー関連 Web サイトのコミュニティ分析とポータルサイト構築—Web コミュニティの関連性から見たグローバル化—," 「グローバル化とジェンダー規範」に関する研究報告書, お茶の水女子大学「グローバル化とジェンダー規範」に関する研究会 (編), pp.101-122, お茶の水女子大学, 2002年3月.
- [8] 小山直子, 増永良文: "Companion-を用いたジェンダー関連 Web コミュニティの詳細分析," 夏のデータベースワークショップ (DBWS) 2004会議録, 7A-3, 2004年7月.
- [9] 増永良文, 小山直子: "Web マイニングツールを用いたジェンダー関連 Web コミュニティの通時的分析," 日本データベース学会 Letters, Vol.3, No.3, pp.21-24, 2004年12月.
- [10] Naoko Oyama, Yoshifumi Masunaga, Kaoru Tachi: "A Diachronic Analysis of Gender-related Web Communities using a HITS-based Mining Tool," Frontiers of WWW Research and Development --APWeb2006, LNCS 3841, Springer, pp.355-366, January 2006.
- [11] 小山直子: "社会現象の分析手法としてのウェブマイニング," 日本のデータベース研究最前線第22回, 月刊 DB マガジン2006年6月号, pp.178-179, 翔泳社.
- [12] 増永良文: "コンピュータサイエンス入門—コンピュータ・ウェブ・社会— (本)," 第13章ウェブ, 第14章ウェブと社会, サイエンス社, 2008年1月.
- [13] Eyetools, Inc.: "Eyetools Research and Reports," http://www.eyetools.com/inpage/research_google_eyetracking_heatmap.htm
- [14] Google の人気の秘密: http://www.google.co.jp/intl/ja/why_use.html
- [15] α SEO: http://www.alphaseo.jp/seo-report/081210_143352.html
- [16] 小山直子, 増永良文, 館かおる: "ウェブ検索ポータルサイトの信用性と透過性—検索キーワード「ジェンダーフリー」を通して見るウェブの世界—," DEWS 2006 (電子情報通信学会17回データ工学ワークショップ/第4回日本データベース学会年次大

- 会) 会議録, ISSN 1347-4413, 1B-i7, 8p., 2006年3月.
- [17] 増永良文, 小山直子: キーワード “ジェンダーフリー” を通してみる検索サイト Google の信用性と透明性, 日本データベース学会 Letters, Vol.5, No.2, pp.105-108, 2006年9月.
- [18] Naoko Oyama and Yoshifumi Masunaga: “On the Trustworthiness and Transparency of a Web Search Site examined using “Gender-equal” as a Search Keyword,” “Proceedings of APWeb2008, LNCS 4976, Springer, pp.625-630, April 2008.
- [19] SEM チャンネル: <http://www.sem-ch.jp/library/000405.html#2>
- [20] E. F. Codd, S. B. Codd, and C. T. Salley: “Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate,” Codd & Date, Inc., 1993 <http://www.fpm.com/refer/codd.html>.
- [21] 石川 沙織, 渡辺 知恵美, 小山 直子, 舘 かおる, 増永 良文: “検索エンジン技術を用いた社会科学の多角的調査支援システムの開発,” DEWS 2008 (電子情報通信学会19回データ工学ワークショップ/第6回日本データベース学会年次大会) 会議録, A1-5, 8p., 2008年3月.
- [22] 増永良文, 渡辺知恵美, 伊藤一成, 小山直子, 竹内純人, 深山鷹一, 舘かおる: “新しい社会調査法としての検索エンジン結果ページ群の自動収集・分析装置の開発— SERPWatcher β 版の開発—,” データ工学と情報マネジメントに関するフォーラム (DEIM Forum) 2009会議録, D7-5, 8p., 2009年3月.
- [23] 星野崇宏: “ネット上の意見表明のバイアスとその補正について,” 人工知能学会誌, 23巻, 6号, pp.791-797, 2008年11月.