

社会調査支援の為の Web ページのランク変動特徴抽出

中部文子[†] 渡辺知恵美[†]
小山直子^{††} 舘かおる^{††} 増永良文^{†††}

商用検索エンジンを使って、あるキーワードを検索した際に現れる Web ページのランク変動には、実世界での社会的事象が反映されていると考えられる。そこで我々は、これまでに社会科学者向けに Web ページのランク変動から社会的事象を読み取することを目的としたランク変動収集・提示システム「SERPWatcher」を開発している。本論文では、SERPWatcher から、特に社会的事象を反映している動きを取り出すことを目的としている。ある事象が社会的に注目が集まることでそれに関連する Web ページのランクが上がって下がるまでの山に表れると考え、ランク変動の山を抽出し、そこからできる社会分析例を示した。

Analysis of transition of ranking in SERP Watcher for social science research

Fumiko Nakabe[†] Chiemi Watanabe[†]
Naoko Oyama^{††} Kaoru Tachi^{††} Yoshifumi Masunaga^{†††}

Recently, finding social status and activities from the enormous quantity of pages from Web is gaining an increasing interest. We focus on the Search Engine Result Page (SERP) because we consider that the rank transitions of web pages are reflected social events and trend of public opinions about the search keywords. Based on the consideration, we are developing SERPWatcher, which is SERP data archiving and visualization system. The system collects SERPs from seven search engines by using search keywords, and it provides OLAP style visualization tool for analyzing the rank transition of web pages. In this paper, we propose a method for extracting socially distinctive rank transitions from SERP archive data. We focus on "active period", which is the period from the rank is drastically raised to the rank comes down. We extract of the active periods from the rank transition of web pages, we analyze the distribution of start dates of active periods and length of the periods, and we found that extracted periods and web pages affect the trend of public opinions about the search keywords.

1. はじめに

Web には実世界のさまざまな出来事が映しこまれている。また、実世界の出来事は時間の経過とともに刻々と変化している。そこで我々は、時系列で変化する Web ページのランク変動をマイニングすることで実世界で起こっている事象を発見・分析できるのではないかと考え、Web ページのランク変動収集・提示システム「SERPWatcher」[1]を開発している。SERPWatcher は、利用者が検索キーワードを登録すると週に一度 7 つの検索エンジンによる検索結果のページとランク (1 位～500 位) を収集し、随時過去のランキングを確認することが可能なシステムである。これは、社会科学者が社会調査を行う際のアンケート調査、実地調査、インタビュー調査に代わる新しい調査方法となりうると考えられる。

本論文では、SERPWatcher をさらに発展させ、ランク変動のマイニングにより利用者がよりスムーズな分析・知識発見ができるようにすることを目標としている。具体的には、多くの Web ページのランク変動の中から特徴的な個所を自動抽出し、社会的事象が反映されていそうなページ群を発見する方法を考えた。その為、長期でのランク変動の特徴をとらえる為、各 Web ページのランク変動を区分最小二乗法により単純化し、はずれ値の除去を行った。

次に、社会的事象を反映していると考えられるランク変動の形をランクが上がって下がるまでの山だと考え、ランク変動から山の抽出を行った。

最後に、これらを使った社会分析を行った。山を構成する要素である「浮上開始時期」「浮上後ランク」「浮上期間」に注目して Web ページを絞っていくことで、実世界で注目の集まった事象とその時期、発端となった出来事を追うことができた。

以後は、第 2 節で SERPWatcher の紹介、第 3 節で想定している目標、第 4 節で社会調査の為の Web ページの抽出、第 5 節で社会分析の例を示し、第 6 節でまとめる。

2. SERP Watcher

2.1 SERPWatcher を利用した社会調査

Web には様々な主体が情報を発信し実世界での出来事や営為が写し込まれている。

[†]お茶の水女子大学大学院人間文化創成科学研究科
Ochanomizu University Graduate School of Humanities and Sciences

^{††}お茶の水女子大学ジェンダー研究センター
Institute for Gender Studies, Ochanomizu University

^{†††}青山学院大学社会情報学部
School of Social Informatics, Aoyama Gakuin University

我々は、ウェブコミュニティの分析研究[2～11]を通して、ウェブマイニングが社会科学でこれまで行われてきた、アンケート調査、インタビュー調査、あるいは実地調査に加えて、有望な研究方法論になりうるのではないかと、という知見を得てきた。その中で我々は、商用検索エンジンによる検索結果ランキングを観察することで社会的事象に関する発見することや、検索エンジンごとのランキングの違いを比較・観察する必要性を感じ、SERPWatcherの開発を行っている。SERPWatcherは、継続的に検索エンジンによる検索結果ランキングを取得してどのようなWebページがどのようにランク変動しているかを蓄積・提示することで利用者である社会学者は、ランキングから社会的動きを読み取ると同時にこの研究方法論が使えるものであるかを検証することができる。

2.2 SERPWatcherの機能

SERPWatcherは、利用者が検索キーワードを登録すると、毎週指定した曜日に7つの検索エンジンによって検索結果のページを1位から500位まで収集し、データベースに格納する。使用している検索エンジンは、Google, Yahoo!Japan, msn, infoseek, baidu, excite, gooである。501位以下のランクはすべて501位として考えている。そして、利用者は、どの検索エンジン・検索キーワードで検索した結果がいつどんなランキングだったのか、ランクインしたWebページがどのようなものであったかを随時確認できるようになっている。図1に結果提示画面の一例を示す。



図1 SERPWatcherのランキング結果提示画面

この画面では、Googleで“貧困”と検索したときの2010年5月20日(基準日)に収集されたWebページとそのランクを表示している。ランク値の背景色は、基準日のランクは緑色、前後のランクについては、基準日と比較して同じなら白、上位なら赤、下位なら青でその度合いが強いほど濃い。例えば、2010年5月20日(基準日)に1位であった「貧困-Wikipedia-」が過去もその後も変わらず1位であったことがわかる。また、11位の「脱！子どもの貧困 子ど…」は2カ月程前はもっと低いランクであったが上昇し基準日以降にまた下降している様子がわかる。それから、各日付でのページアーカイブを確認ボタンがあり、ニュースなど同じすぐ期限が切れてしまうページでも当時の紙面を確認することができるようになっている。さらに、各Webページの当時のバックリンク数を確認することができることや、ランクの上昇・下降が閾値を超えたとき利用者に通知するしくみもある。

2.3 社会分析におけるSERPWatcherの課題

SERPWatcherには、社会学者が特徴的ランク変動を発見することに関して課題がある。現在、社会科学として興味深い動きを発見する為には、利用者が図1を眺めて色やランクの数字を追っていく必要がある。これでは、興味深い動きがあっても見落と

す可能性があることやページをまたいだ分析が難しい。そこで、これからはシステムが社会的事象の動きを反映していると思われる特徴的な部分だけを自動抽出して提示する必要がある。

3. 想定するシナリオ

本節では、目指す社会分析がどのようなものであるかを述べる。大きな目標は、実世界でのある事象に対する議論がランク変動に反映されている様子を取り出すことである。その例として図2では、ある社会的事象（例えば、ジェンダーフリー、夫婦別姓など）について賛成意見をもつページ群 (pageC,D) がランク上昇した同時期に、反対意見を持つページ群 (pageA,B) がランクを落とす様子を示している。もしこのような例が抽出できれば、ランク変動から社会的事象に関する議論を読み取ることができ、興味深い社会科学の研究材料となる。その為に、我々は大きくランクを上昇・下降させたページを取り出し、グループ化し、内容を観察する必要がある。

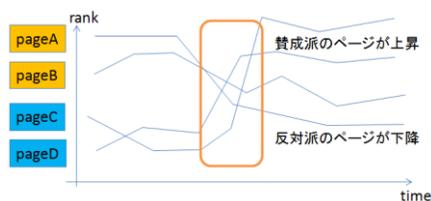


図2：想定するシナリオ

しかし、ランク変動の原因には社会的事象以外に検索エンジンのアルゴリズムによるものやデータの更新などさまざまな原因が考えられる。そこでさらに、上で述べた例が本当に社会的事象が反映されたものであるかを考える為、検索エンジンごとで結果を比較する必要がある。図3のように、検索エンジンAで図2で示した例のようなランク変動があり他の検索エンジンBでも似た動きをしていれば、このランク変動の原因は検索エンジンのアルゴリズムによらずに起こっている結果であり、社会的事象が反映されている可能性が高いと考えられる。一方、検索エンジンCのように反対派 (pageA,B) が上昇し賛成派 (pageC,D) が下降するという検索エンジンA,Bと異なる動きをする例があれば、社会的事象が反映されている可能性は低くなると考えられる。

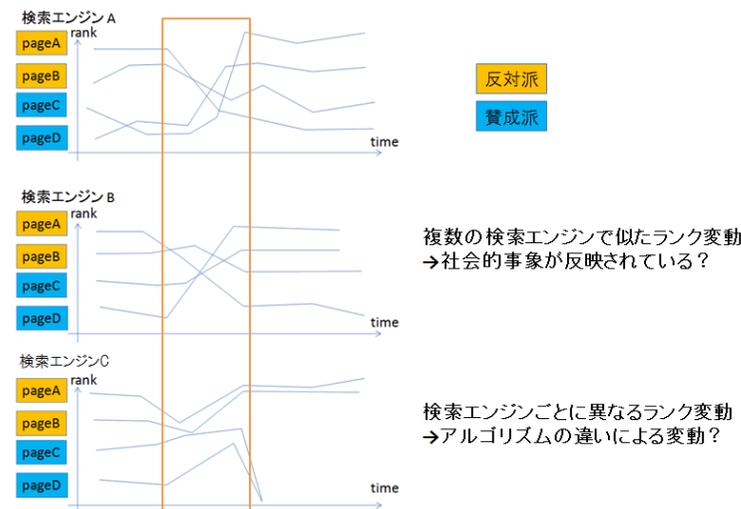


図3：ランク変動原因の予測

4. 社会調査の為の Web ページ抽出

SERPWatcher が持つランク変動のデータは膨大であり、社会分析に有益な特徴的変動とそうでないものが混在している。例えば、Wikipedia が常に上位にとどまっていることは検索エンジンの意向によるものと考えられ、我々が発見したいタイプのランク変動ではない。また、Wikipedia のように作成元の企業・団体の規模が大きいページや公式ページといったページは上位で固定的である傾向があることが以前の我々の研究[12, 13]でわかっている。一方、下位のページ (50 位以下のページなど) では上位より激しくページの入れ替わりが起きていることが SERPWatcher の観察から見てとれる。その中には一般的には知名度が低い、一部の人たちの間で起こっている議論や、まだ話題になっていないがこれから話題に上ってくる可能性のある動きが反映されていると考えられる。本節では、このような社会調査にとって興味深いランク変動をする Web ページの抽出方法について述べる。以下の (1), (2) でランク変動の形に着目して Web ページを絞り込み、(3) でランク変動から分析の為の指標を取り出す手順を述べる。

(1) ランク変動の単純化

まず、長期のランク変動の大きな流れをとらえる為、区分最小二乗法[14] を用いて、ランク変動を単純化した。区分最小二乗法は、2次元平面上の点列をできるだけ少な

い本数の直線で近似するアルゴリズムである。各 Web ページにおいて、ランク変動の情報を、座標が(収集日, ランク値)である二次元平面の点列と考え、それを直線近似する。図4にランク変動に区分最小二乗法を施した例を示す。左が Web ページ「反貧困ネット北海道」の2009年3月から2010年6月までの Google におけるランク変動、右がそれを単純化したものである。細かい変動が一本の線分に近似されていることがわかる。このように、小さな変化を無視することで大きく変動した部分が際立ち、大きなランク変動の流れをわかりやすくすることができる。

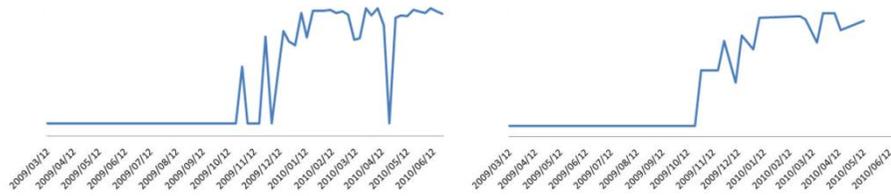


図4：元のランク変動(左)と区分最小二乗法・はずれ値除去後のランク変動(右)

また、はずれ値の除去もあわせて行った。図2左の2010年1月から2010年6月までのように全体を見ると比較的ランクが上位にとどまっているように見える途中で2010年4月下旬にランク圏外になって次にはすぐ戻る例がある。これは、実際に検索エンジンによる検索結果のランクが圏外になったこと以外に、検索エンジンがアルゴリズムの変更やデータ更新を行ったことによる変化である可能性がある。また、Googleでは、ブラウザで検索した際の結果とAPIで検索した際の結果が異なることが我々のこれまでの観察からわかっており、我々はAPIを利用せずURL指定によるスクレイピングで収集を行っている。ただし、このブラウザからの検索で得られる結果は問い合わせの際に異なるサーバにアクセスしている可能性があり、それぞれのサーバが持つデータが異なることによる変化が起こる可能性がある。そして、それが突然のランク圏外への変化に現れると予想している。そこで我々は、圏内のランクから圏外になり、すぐ元の圏内のランクに戻る時、圏外であったランクを前後のランクで補完し連続値とすることにした。図2右ではランク補完によりはずれ値が除去されていることがわかる。

(2) ランク変動の山の抽出

社会的事象の変化がどのようにランク変動に反映されるかについて、我々は、ランクが上昇し下がるという形に現れると考えた。なぜなら、ある出来事に注目が集まるとランクが上昇し、話題が別に移り注目されなくなるとランクが下がると考えられるからだ。このランクが上がって下がるという一連の流れ(以降、山と呼ぶ)を取り出

し観察することで実世界の話題の内容や発生・消滅の流れを取り出せるのではないかと考え、各 Web ページから山の抽出を行った。図5が先ほどの区分最小二乗法を施した Web ページから山の抽出を行った様子である。ランクが上がりを始めた時を山の開始点とし、下がった後を山の終了点とした。

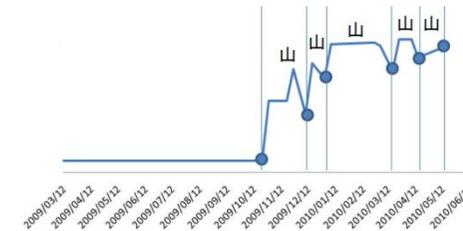


図5：山の抽出

(3) 社会的事象を特徴づける3つの指標

(2)で抽出したランク変動の山から、社会的事象を特徴づける為の3つの指標を取り出した。一つ目は、山の開始日「浮上開始日」である。これはそのページに関連する話題に注目が集まった時期を表わし、実世界の出来事に照らし合わせるのに重要である。二つ目は、上昇している期間「浮上期間」である。これは注目を集め続けた期間であり、その話題の定着度がわかる。三つ目は、最も上昇した時のランク「浮上順位」である。浮上順位はからその話題のメジャー度が予測できる。例えば、浮上開始日と浮上期間が同じでも浮上順位が5位と50位のものでは、前者の方がよりメジャー度は高い。

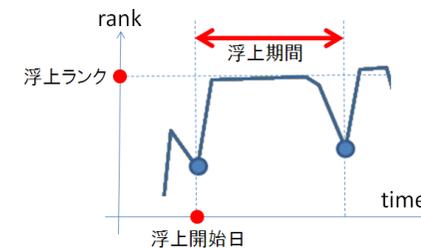


図6：ランク変動の山からわかる、社会的事象を特徴づける3つの指標

5. 社会分析の例

4 節で定義した浮上開始日、浮上期間、浮上順位に注目して、ランク変動と社会的事象の動きとの関係を追った。ここでは、主に Google を使って「貧困」と検索した際の結果を使用して分析を行った。その結果、ランク変動から実世界の社会的事象を読み解くことができた。

・浮上開始日の集計

図 6 は Google で貧困と検索した際に現れる Web ページについて浮上開始日の頻度を月ごとに集計した結果である。

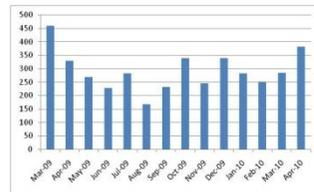


図 7 : Google で貧困と検索した際の Web ページの浮上開始日の頻度

ここからランク上昇の多い時期がわかる。2009 年 5 月頃にランクを上昇させる Web ページが多かったことがわかる。逆にその後 3 カ月は頻度が少なくなることから、ランク変動が安定していったとわかる。さらに、Web ページの内容の観察から、これらの Web ページで取り上げられている話題が主に 3 種類あることがわかった。低収入の労働者の為の NGO 団体である反貧困ネットワーク、貧困層をターゲットにした悪質な貧困ビジネス、全国民の所得の中央値の半分に満たない国民の割合である“貧困率”である。そこで、「貧困率」「反貧困」「貧困ビジネス」というキーワードをタイトルまたはスニペットに含むという条件で Web ページを分類し、それぞれで再び浮上開始日を集計した。その結果が図 8～10 である。

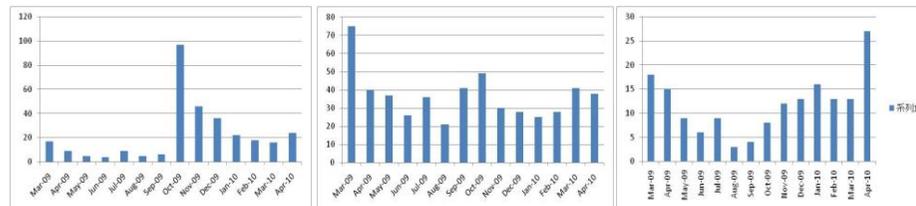


図 8 : 浮上開始日頻度
「貧困率」

図 9 : 浮上開始日頻度
「反貧困」

図 10 : 浮上開始日頻度
「貧困ビジネス」

それぞれの話題が注目を集めた時期がわかり、それぞれ異なることがわかる。この中から最も浮上開始日の頻度の増え方が激しかったのが「貧困率」の 2009 年 10 月である。この結果から、2009 年 10 月周辺で貧困率にまつわる何らかの出来事が注目を集めたと推測できる。そこで、以降ではこの月の貧困率関連を話題についてさらに詳しく調べてみた。

・定着度・メジャー度の集計

図 8 の 2009 年 10 月に集計された Web ページ群がどのようなものであるかを調べた。まず、それらが上昇後どれだけ上昇を続けたかの指標浮上期間と最上昇ランク浮上順位を集計した。図 11 は、横軸を浮上期間、縦軸を浮上順位として図 8 での 2009 年 10 月に上昇した Web ページを散布図にしたものである。この図から、多くのページが比較的上位まで上昇し、すぐランクを落とす一方、いくつかのページだけ上昇をしばらく続けたことがわかる。

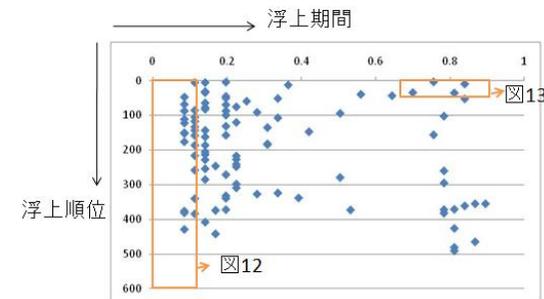


図 11 貧困率に関するページの浮上順位・浮上期間(2009 年 10 月に上昇したページ)

まず、短期間に集中的に注目を集めたページとして浮上期間が短い順に上位 23 件のページのタイトルを図 12 に示す。

タイトル	種類
<貧困率>政府として調査する方針固める長妻厚生労働相(毎日新聞...	ニュース
子どもの貧困率調査へ 国内 Reuters	ニュース
母子、父子家庭の貧困率 OECD中ワースト1(エコノミックニュース...	ニュース
「貧困率」って？	ブログ
「相対的貧困率」について色々と考えてみる.....(4)「経済格差」は「世代...	ブログ
10/21(水) 貧困率ワースト4位。酔語酔吟夢がたりウェブブログ	ブログ
首相が相対的貧困率15.7%に「ひどい」-政治ニュース..	ブログ
虹とモンスーン【反貧困ネットワークから】貧困率測定についての声明	ブログ
JavaScript+かも日記【雑談】相対的貧困率って何？	ブログ
甚之介の剣道雑記帳:疑問だらけの貧困率	ブログ
【調査】「日本は7人に1人以上が貧困」日本の貧困率、15.7...	ブログ
村野瀬玲奈の秘書課広報室 日本の貧困率がやっと明らかに	ブログ
時事ドットコム:貧困率「ひどい数字」=鳩山首相	ブログ
厚生相、子どもの貧困率調査へ不況で問題深刻化受け-47NEWS(よんな..	ニュース
<貧困率>日本15.7%先進国で際立つ高水準(毎日新聞社..	ニュース
貧困率:政府として調査する方針固める長妻厚生労働相-毎日.jp..	ブログ
7人に1人も...日本の貧困率が世界4位-政治・社会-ZAKZAK	ニュース
えっ日本は世界5位の貧困国「貧困率」が示すものってなに R25	ブログ
山陰中央新報-日本の貧困率は15.7% 07年、98年以降で最悪	ブログ
日本の貧困率は15.7%/07年、98年以降で最悪?四国新聞社	ニュース
えかわ珈琲店のブログ:貧困率	ブログ
千葉発 帯張経由-社労士事務所のblog:日本の「貧困率」15.7..	ブログ
大脇道場 NO.1398 新政府の「貧困率調査」への着手を歓迎する。調査..	ブログ

図 12：2009年10月に上位まで浮上し浮上期間の短いページ

これらは、内容はほとんどが2009年10月の朝日新聞に掲載された厚生労働大臣の長妻昭氏が日本の相対的貧困率が2007年に15.7%に達したことを伝えるニュースに言及するものであった。ページの種類は個人のブログやニュース記事が多くあった。次に、長期にわたり上昇を続けたページも観察した。図13は図11に示すWebページのうち、浮上順位が30位より上位、浮上期間が2カ月以上であるもののリストである。

タイトル	浮上開始日	浮上順位	浮上期間
厚生労働省:相対的貧困率の公表..	2009-10-15	9	210
asahi.com(朝日新聞社):日本の貧困率...	2009-10-15	3	189
池田信夫 blog:「貧困率」についての誤解	2009-10-29	12	91
貧困率 - Wikipedia	2009-11-29	2	175

図 13：浮上順位が上位で浮上期間が長い Web ページ

この4つは、相対的貧困率を算出したという内容の厚生労働省公式ページ、朝日新聞のニュースサイトが貧困率算出を伝えた記事、有名ブロガーが相対的貧困率の見方について考えを述べているブログ、Wikipediaの貧困率のページ、であった。中でも一番目のページはこの話題の発端となった厚生労働省の公式発表の文書を掲載しているページで、先ほど図12で示したページとは性質が異なり、話題の発端となったページである。また、Wikipediaの貧困率は、図14に示すように、11月以前は13位や圏外であったが11月頃から次第に安定的に3位にランクするようになっており、今回の出来事によって「貧困率」という言葉が一般的な言葉として扱われるようになった様子がわかる。

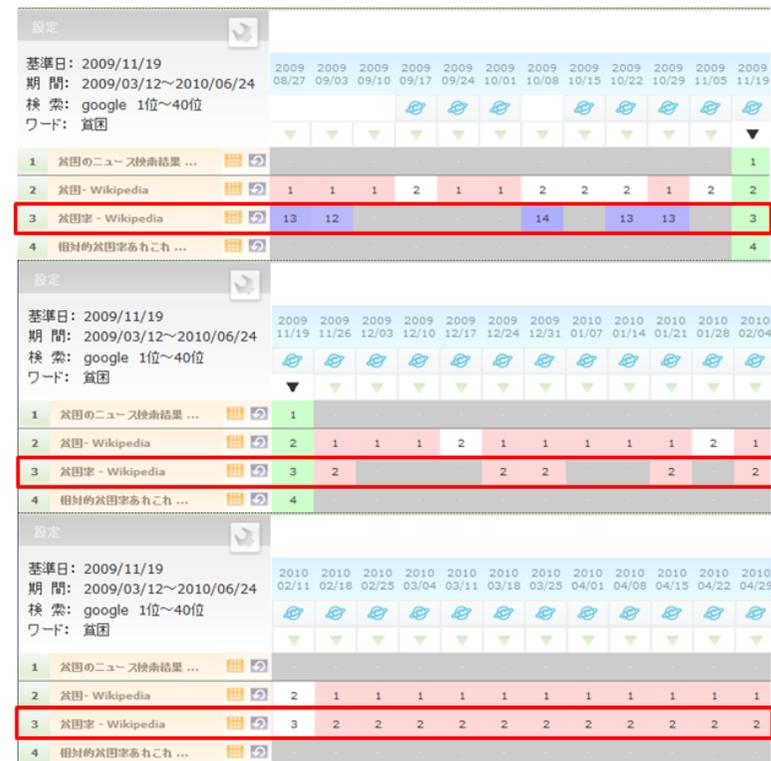


図 14 「貧困率-Wikipedia」の2009年8月27日～2010年4月29日までのランク変動

浮上期間、浮上順位とページ内容との関係を調べた結果、短期で上位まで上昇しすぐ落ちるページはある社会的現象が起こったときにそれに言及するニュース記事や個人のブログなどが多くそれらの観察により実世界での出来事を知ることができた。一方上昇を続けるページからはその話題の発端となったページや社会的に影響のある人物のページをなど、出来事の原因やそれによる社会的変化を知ることができた。

・他検索エンジンとの比較

Googleで行ってきたこれまでの結果をYahooでの結果と比較したものを図15に示す。浮上開始日の頻度は、GoogleがYahooより全体的に少ないが、月ごとの上昇割合は類似していた。

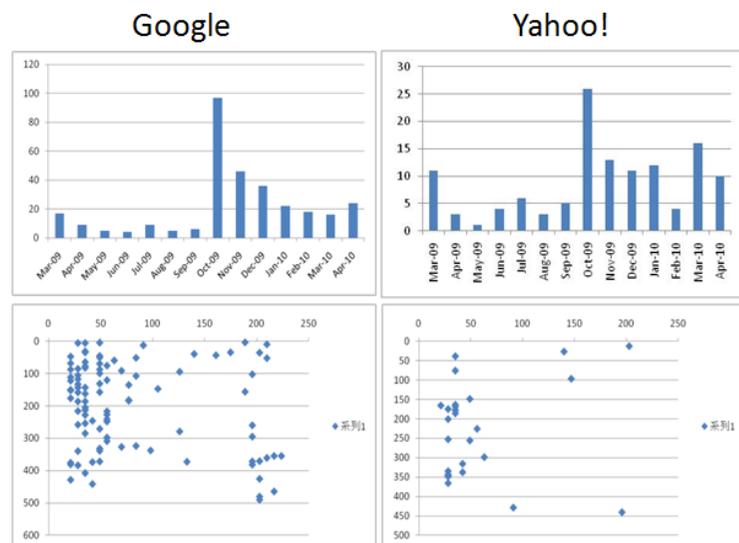


図15: Google, Yahooでの浮上開始日, 浮上順位, 浮上期間の集計

6. まとめと今後の課題

Webページのランク変動から社会分析する為の情報抽出として、ランク変動の単純化、山の抽出を行い、山の開始時期と上昇期間に着目した集計を行い、社会分析例を示した。今後は、それらをすべて自動で行い、利用者に直接、出来事の発生時期やその推移の分析結果を提示できるようにしたい。

謝辞 本研究を進めるにあたりご助言いただいた、青山学院大学社会学部情報学部の伊藤一成先生、青山学院大学情報科学研究センターの竹内純人先生に深く感謝いたします。

参考文献

- 1) 増永良文, 渡辺知恵美, 伊藤一成, 小山直子, 深山鷹一, 館かおる: "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—SERP Watcherの設計—" DEIM2009 D7-5, 2009年3月.
- 2) 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, pp. 101-122, お茶の水女子大学, 2002年3月.
- 3) 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 (DBWS2004) 会議録, 7A-3, 2004年7月.
- 4) 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.
- 5) Naoko Oyama, Yoshifumi Masunaga, Kaoru Tachi: ADiachronic Analysis of Gender-related Web Communities using a HITS-based Mining Tool, Frontiers of WWW Research and Development--APWeb2006, LNCS3841, Springer, pp. 355-366, January 2006.
- 6) 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—, DEWS2006 (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.
- 7) 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D Bマガジン 2006年6月号, 翔泳社.
- 8) 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.
- 9) 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14章 ウェブと社会, サイエンス社, 2008年1月.
- 10) 石川 沙織, 渡辺 知恵美, 小山 直子, 館 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, DEWS2008A1-5, 2008年3月.
- 11) 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, Springer, April 2008.
- 12) 中部文子, 渡辺知恵美, 小山直子, 館かおる, 増永良文: 社会調査支援の為の SERP Watcher からのランク変動特徴抽出, DEIM2010 A2-5, 2010年3月
- 13) Chiemi Watanabe, Fumiko Nakabe, Naoko Oyama, Kaoru Tachi, Yoshifumi Masunaga: Mining Socially Distinctive Transitions from Search Engine Result Pages, KJDB2010 May 2010
- 14) Jon Kleinberg, Eva Tardos (邦訳: 浅野孝夫, 浅野泰仁, 小野孝男, 平田富夫): "アル

ゴリズムデザイン" 共立出版. pp. 232-236. 2008.