SERPAnalyzer：社会調査支援の為の SERP アーカイブからの特徴的ランク変動分析システム

中部 文子　(指導教員：渡辺知恵美)

1　はじめに

商用検索エンジンによる検索結果ランキングは世の中の動きにあわせて日々変化している．我々は，ランキングの変化を観察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，社会科学者向け Web ページのランク変動収集・提示システム「SERPWatcher」[1] とランク変動分析システム「SERPAnalyzer」を開発した．SERPWatcher は，利用者が指定する検索キーワードに対して，さまざまな検索エンジンの検索結果ランキングを定期的に収集し，ランキングヒストリを提示するシステムである．本稿では，SERPWatcher のランキングヒストリから特徴的なランク変動を自動抽出・提示するシステム SERPAnalyzer の実装と活用について述べる．SERPAnalyzer では，Web ページのランク変動からランクアップしてからランクダウンするまでの箇所を抽出し提示した．これにより，社会事象に注目が集まった時期や規模や定着度の推測が可能となった．

2　SERPWatcher

SERPWatcher は，過去のランキングヒストリを収集し提示し，ユーザに社会事象の発見・分析を促すシステムである．まず，ユーザが興味のある検索キーワードを指定するとシステムが週に一度検索エンジン Google, Yahoo!, Infoseek, baidu, goo, being, excite による検索結果を収集する．検索結果とは，1 位から 500 位までのランキングとランキングを構成する各 Web ページのアーカイブである．これらランキングヒストリ提示画面に反映する．図 1 は，ランキングヒストリ提示画面の一例である．Google で “貧困” と検索したときの 2010 年 5 月 20 日 (基準日) に収集されたランキングとそれらの Web ページの過去のランクを示している．SERPWatcher により，過去にどの検索エンジンでいつどのような Web ページが何位にランクインしていたかを確認することができる．

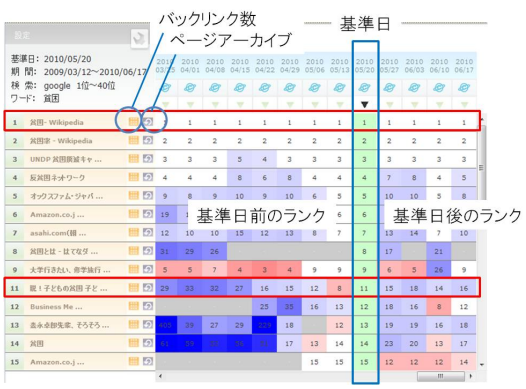

図 1: SERPWatcher ランキングヒストリ提示画面

3　SERPAnalyzer

SERPWatcher で長期間ランキングヒストリを蓄積した場合，データが膨大になりユーザが特徴的なランク変動を読み解くのは困難となる．そこで，ランキングヒストリから特徴的なランク変動の候補を抽出する SERPAnalyzer を提供する．

3.1　浮上期間抽出

社会事象を特徴づける特徴的なランク変動として，ランクアップしてからダウンするまでの動き「浮上期間」に着目した．図 2 に示すように，浮上期間の開始日からは事象に注目の集まった時期が，浮上ランクからは事象の注目された規模，浮上期間の長さからは事象の定着度が推測できる．

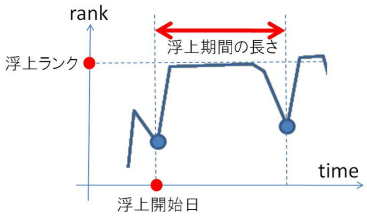

図 2: 浮上期間と社会事象を特徴づける 3 指標

浮上期間抽出は，ランク変動を区分最小二乗法 [2] とランクの補完によって単純化したランク変動に対して行った (図 3)．区分最小二乗法とは，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．ランクの補完とは，圏内 (500 位以内) のあるランクが圏外 (501 位以下) になりすぐもとのランクに戻る個所について，圏外のランクを前後のランクで補完することである．この単純化により，長期にわたるランク変動のおおまかな流れをとらえることができ，社会分析に適したランク変動を得ることができる．

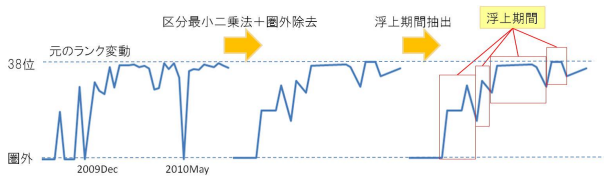

図 3: 浮上期間抽出手順

3.2　分析画面

SERPAnalyzer の分析画面をユーザの分析の流れに沿って図 4 中①から④に示す．

画面①でまず，分析対象 Web ページを検索エンジン・検索キーワード・サブキーワードによって絞る．サブキーワードを指定した際は，それをタイトルやスニペットに含む Web ページが対象となる．画面②で対象

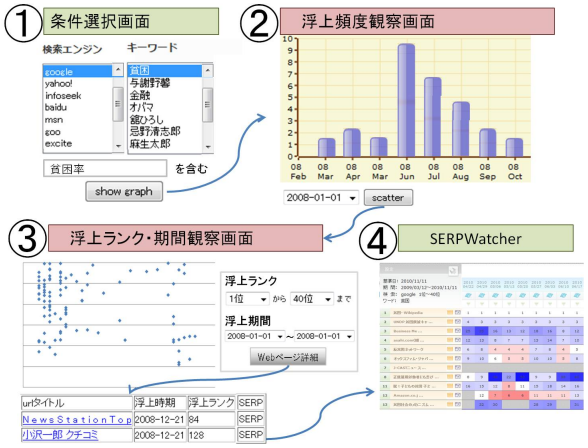

図 4: SERPAnalyzer による分析の流れ

Web ページの浮上期間を抽出し，浮上開始日の頻度を月ごとに集計したヒストグラムを示す．ここから，注目の集まった時期の推測ができる．ユーザが興味のある月を選択すると③でその月に浮上した Web ページの浮上ランクと浮上期間の長さを示す散布図が表示される．各 Web ページ群を選択して，それらのタイトルを確認したりそのページにリンクすることも可能である．これにより，どのような Web ページが何位までどのくらいの期間浮上していたかを知ることができる．さらに，SERPWatcher ランキングヒストリ提示画面④に移動し詳細なランク変動を確認することもできる．

4 SERPAnalyzer を使った分析例

SERPAnalyzer を使い，有効な分析ができるか検証を行った．検索エンジン：Google，検索キーワード：貧困，サブキーワード：貧困率，を条件に絞り込んだ Web ページを対象に分析した例を示す．

まず，浮上開始日の月ごと頻度を観察した (図 5)．

図 5: 浮上開始日の月ごと頻度

図 5 を見ると，2009 年の 10 月に多くの Web ページが浮上しており，この時期に貧困率に関する話題に注目が集まったことが推測できる．そこで，2009 年 10 月に浮上開始した Web ページの浮上ランク・浮上期間について観察する．

図 6 の Web ページの分布から，主に浮上期間の短い Web ページのグループ，浮上期間が長いグループのうち浮上ランクが上位であるグループ，下位であるグループの 3 グループがあることがわかる．まず，浮上期間の短いグループはニュース記事やブログであった (浮上期間が短い上位 27 件中ニュース 7 件ブログ 20 件)．内容はどれも元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

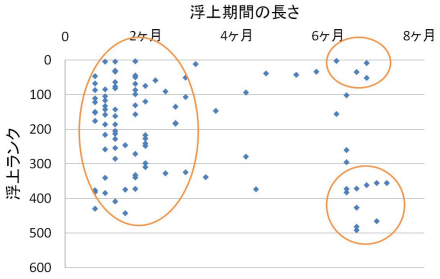

図 6: 浮上ランク・浮上期間 (2009 年 10 月浮上)

伝えるニュースとそれに言及するものであった．次に，浮上期間の長いグループのうち浮上ランクが上位であるグループの内容を調査した (表 1)．相対的貧困率を算出したことを伝える内容の厚生労働省公式ページが含まれていた．これは多くの人が話題にしていた事象の発端となるページである．多くの人が参照したことにより上位に長期浮上したと考えられる．また，浮上

表 1: 長期・上位に浮上していた Web ページ

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

期間の長いグループのうち浮上ランクが下位であるグループは，掲示板が含まれることが特徴的であった (表 2)．掲示板は少数の人々が更新を繰り返す性質から低いランクで長期浮上したと考えられる．

表 2: 長期・下位に浮上していた Web ページ

タイトル	種類
[鳩山首相、貧困率を見て「大変ひどい数字。なんでこんな日本にしたの ...	掲示板
日本の貧困率15．7%	掲示板
日本の「貧困率15.7%」格差是正と経済全体の成長、どちらを優先す ...	掲示板
日本の「貧困率」15・7%、OECD中4位 - BIGLOBEニュース	ニュース
日本の貧困率は15．7% 厚労省が初公表：イザ！	ニュース
タカマサのきまぐれ時評2 ＜貧困率＞政府として調査する方針固める ...	ブログ
「貧困率」に関する若干の問題点（1）- 松尾光太郎 de 海馬と玄関 ...	ブログ
＜貧困率＞日本15．7% 先進国で際立つ。子ども手当は所得制限の ...	ブログ
日本の「貧困率」15・7%という数字？ -- メルペーユ森通信	ブログ

5 まとめと今後の課題

本論文では，商用検索エンジンの検索結果ランキングのヒストリから社会事象を反映すると思われるランキングの変化を抽出し，提示するシステム SERPAnalyzer を提案した．ランクが浮上する期間を抽出することで，注目された社会事象の詳細を知るきっかけを提示することができた．今後，さまざまな検索エンジン・検索キーワードの組み合わせによる分析を行い，より有効な社会分析を行えるよう改良を加えていきたい．

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, 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 — SERP Watcher の設計—” DEIM 2009 D7-5, 2009 年 3 月.

[2] JonKleinberg,EvaTardos(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.