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

石川 沙織　（指導教員：渡辺 知恵美）

1. はじめに

ウェブマイニングツールを使ったジェンダーコミュニティの分析研究を通し，ウェブマイニングはアンケート調査，などといった従来の研究方法論とは一線を画す社会科学の全く新しい研究方法論になり得るという知見を得ている．また，その研究過程で検索サイト Google の SERP（search engine result page，検索エンジン結果ページ）には，その表示順位に Google が公表している順位付けストラテジでは解明しがたい不可解さも見受けられ，検索サイトの信用性に関する研究も展開してきている[1]．よって，社会科学の研究手法として SERP を用いるのは確実な研究手法であるとはいいがたい．

本稿では，このような知見をさらに確実なものとすべく，SERP といった一意的な側面にとらわれない社会科学の研究推進に役立つと考えられる「統合型ウェブマイニング環境」の根幹となるシステムである，Comparator と称する機能を開発する．

2. Comparator

本研究では検索エンジンを用いて調査や分析を行う社会科学を対象に，多角的な視野からの検索結果を提供するシステムを考案し実装する．具体的には，調査したいキーワードに対し，複数の検索エンジンを用いて検索結果を取得し，それらの順位，ページランク，バックリンク数を共に比較し，また検索結果表示順位の遷移を時系列に沿って表示する機能を提供する．

2.1. 基本機能とデータベース

前節で述べた機能を実現するために，Comparator は 2 つにモジュール，SERP 収集部・問合せ処理部から成る．

図 1 に構成とモジュール間の処理の流れを示す．実装開発言語は perl，DBMS は MySQL を用いる．

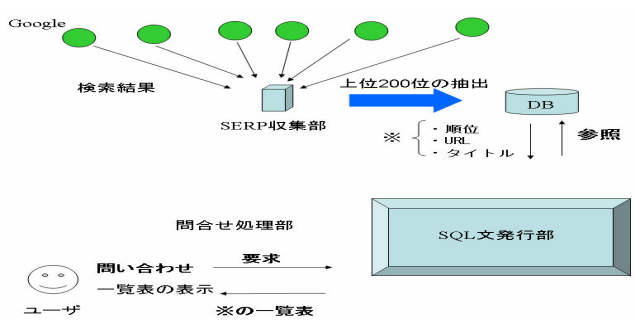

図 1:　Comparator の構成とモジュール間の処理の流れ

(1) SERP 収集部

検索結果収集プログラムを週に一回 unix の cron コマンドを使用し実行する．このプログラムは 5 検索エンジン: Google, Yahoo! Japan, MSN, goo, excite，における，社会科学の専門家からアーカイブを取得するように指示のあった，ジェンダー，ジェンダーフリー，男女共同参画，フェミニズム，セクシュアルマイノリティー，セクシャルハラスメント，夫婦別姓，児童虐待，ニート・フリーター，美容整形という，指定された検索キーワードの検索上位 200 位をウェブページの形で保存する．次に，収集したウェブページを読み込み，ランキング上位 200 位の，URL，タイトル，ページ内容要約を抽出して，データベースへ格納する．SERP 収集の対象世界を表す実体—関連モデルを図 2 に示す．なお，*keyword*, *url_id* が外部キーとなり，id が主キーとなる．

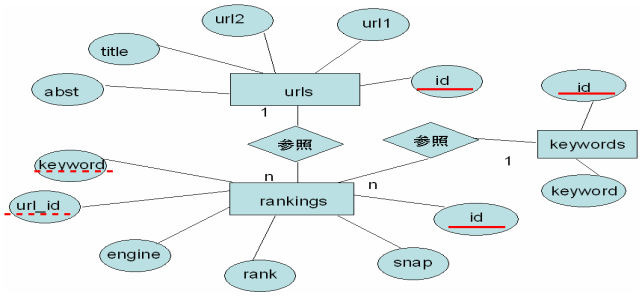

図 2:　SERP 収集の世界を表す実体—関連モデル

同一ホームページが幾度となく Comparator 収集部において収集されるため，URL はドメイン部とファイルのディレクトリパス部の 2 つに分割する．抽出にあたり，解析した Yahoo!Japan の HTML ファイルの一部を図 3 に示す．Yschttl クラスのアンカータグに囲まれたテキストはタイトル部分である．Yschabstr クラスの div クラスタグに囲まれたテキストは，ページ要旨である．Sinf クラスの div クラスに囲まれたテキストはページ URL にであり，更に<wbr>タグによってドメイン部とファイルのディレクトリパス部に分割される．他の検索エンジンにおいても同様にパターンの解析をし，それらを抽出するプログラムを実装する．

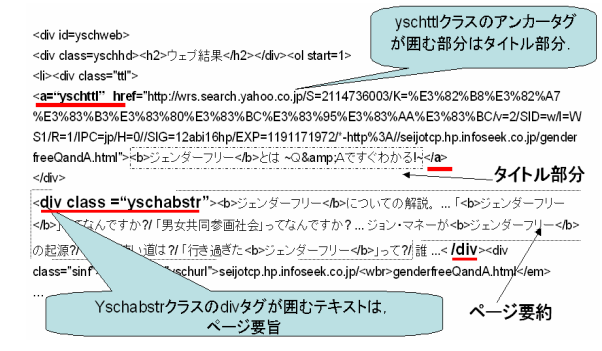

図 3:　Yahoo!Japan の HTML ファイルの一部

抽出データは，テーブル *rankings*，テーブル *urls*，テーブル *keywords* の 3 テーブルに格納される．テーブル名 *rankings* には，*id*，*url1*（URL のドメイン部），*url2*（URL のディレクトリパス部），*title*（サイト名），*abst*（要約内容）が格納される．テーブル名 *urls* には，*id*，*snap*（デー

タ取得時間)， *rank*（順位），*engine*（検索エンジンの種類），*url_id*（テーブル *urls* の *id*）が格納される．テーブル名 *keywords* には，*id*，*keyword*（検索対象のキーワード）が格納される．

(2) 問い合わせ処理部

SERP アーカイブに問合わせを行うことで，検索結果表示順位のさまざまな比較表を生成可能である．たとえば，典型的に次の問合せを考えることができる．

［Q1］2007 年 10 月 20 日で Google の 1 位から 10 位の順位は Yahoo での対応関係を求める SQL は以下の通りである．

```
SELECT   r1.rank,r2.rank, r1.url_id
FROM     ranking r1, ranking r2
WHERE    r1.snap = '2007/10/20'
AND      r1.snap = r2.snap
AND      r1.engine = 'Google'
AND      r2.engine = 'Yahoo! Japan'
AND      r1.rank between 1 and 10
AND      r1.url_id = r2.url_id
ORDER BY      r1.rank;
```

2.2. Comparator の実装と GUI

SQL での問合せの結果は表なので視認性に優れているとはいえない．よって，さまざまな条件で SERP アーカイブを検索した結果である，順位基軸の指定や順位変動を色付けにより表示する可視化機能などを加えた．図 3 に，2007/11/18 基点での Google での「夫婦別姓」の検索順位の推移を示したかを表示する GUI を示す．検索エンジンの選択コマンドには 5 検索エンジンのプルダウンメニュー，キーワードの選択コマンドには 10 種のキーワードのプルダウンメニューがあり，各々を指定して組み合わせて検索結果を得ることができる．図 3 では 2007/11/18 基点での表示結果を基軸として左右に展開している．そして，別の日付をクリックして指定することにより，その日付で得た順位が基軸として再読み込める機能を盛り込んだインタフェースである．表示結果は，基点日での上位 50 位の閲覧が可能である．基点日時よりも順位が高くなると赤く，低くなると青くなるようにし，その順位差に濃淡を付けている．よって，徐々に上がっているもの，徐々に下がっているもの，急激に順位が上下しているものなどの判別が容易になる．

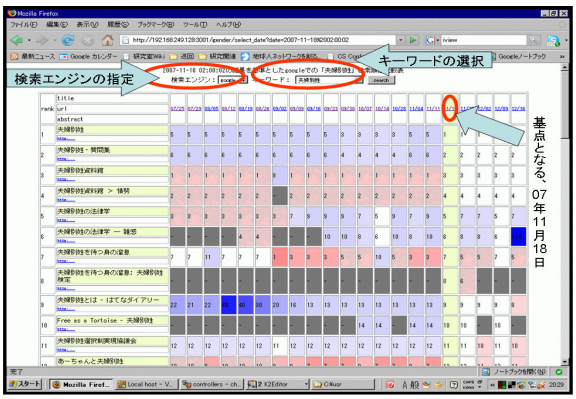

図 4: キーワード「夫婦別姓」での SERP 順位の比較一覧

濃淡設計の際，赤の濃淡には，(R, G, B) = (255, x, x) と変数を設定し，青の濃淡にには，(R, G, B) = (x, x, 255) と変数を設定した．これにより赤・青、共に基準順位と表示順位との差によって，濃淡が定まる．但し，順位差が 0 である場合は，一律に x = 0 となり，白色となる．また，基準順位と表示順位との差が 50 以上では最も濃い色彩での青・赤で統一した．この機能は視認性に優れており，社会科学の研究者にとって優しいユーザインタフェースであると同時に，強力な分析ツールとなる．なぜなら，順位の入れ替わりに際しては社会科学的に大変興味深い事象が関連していることが多いからである．

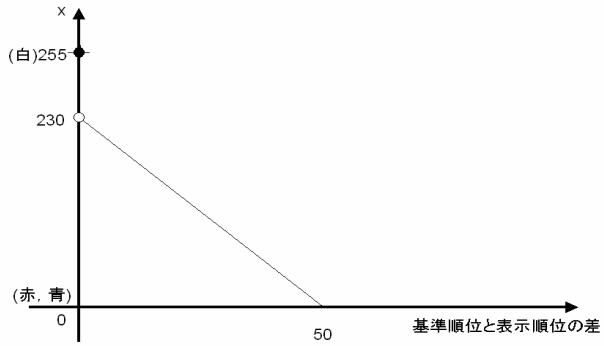

図 5: インタフェースの濃淡設計

3. まとめと今後の課題

ウェブマイニングが社会科学の新しい研究方法論になることを実証する目的で，ウェブ空間を巨視的かつ微視的に分析できるシステムの開発実現のための Comparator の構築を述べた．設計際しては，ユーザにとって優しいインタフェースであることはもちろん，社会科学の調査における多角的な研究調査内容を提供可能なツールの完成を目指した．本システムの有効性を社会科学者からも評価を得るに至った．これは，統合型ウェブマイニング環境実現のための基本機能であり，統合型ウェブマイニング環境の布石となった．

参考文献

[1] 小山直子，増永良文，舘かおる：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．

[2] 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，DEWS2008（電子情報通信学会 19 回データ工学ワークショップ，2008，（投稿中））