様式 C-19

# 科学研究費補助金研究成果報告書

平成 22 年 5 月 25 日現在

研究種目：基盤研究（B）
研究期間：2007～2009
課題番号：19300025
研究課題名（和文）　社会科学の新しい研究方法論としての総合型ウェブマイニング環境の開発研究

研究課題名（英文）　Research on Developing an Integrated Web Mining Environment as a Novel Research Methodology of Social Sciences

研究代表者
増永　良文（MASUNAGA YOSHIFUMI）
青山学院大学・社会情報学部・教授
研究者番号：70006261

研究成果の概要（和文）：検索エンジン結果ページ（SERP，Search Engine Results Page）に現れるウェブページやその表示順位は実世界の事象や動きを表しているとの考え方から，SERPWatcher と名付けた統合型ウェブマイニング環境を設計・開発した． SERPWatcher は社会変革を発見したい社会科学の研究者に対して，これまでのアンケート調査やインタビュー調査に代わる全く新しい研究方法論となるであろう．プロトタイプが稼働しており有効性を確認している．

研究成果の概要（英文）：This research investigates the design and implementation of an integrated web mining environment named SERPWatcher based on the observation that the search engine results page (SERP) itself and the ranking order change with time reflecting the changes in society. It could be a novel social survey method in that it totally differs from the traditional methods such as questionnaires and interviews. A research prototype of SERPWatcher is currently under operation, and its validation test shows that it is working as intended.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2007 年度 | 5,400,000 | 1,620,000 | 7,020,000 |
| 2008 年度 | 4,900,000 | 1,470,000 | 6,370,000 |
| 2009 年度 | 3,900,000 | 1,170,000 | 5,070,000 |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 14,200,000 | 4,260,000 | 18,460,000 |

研究分野：総合領域
科研費の分科・細目：メディア情報学・データベース
キーワード：ウェブマイニング，ジェンダー，社会科学，研究方法論，SERPWatcher

## １．研究開始当初の背景

ウェブ（World Wide 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．ウェブ 2.0 が提唱され，ウェブが有する潜在的可能性がますます顕在化されつつあるなか，我々は，ウェブコミ

ュニティの分析研究を通してウェブマイニングが社会科学の新しい研究方法論になりうる可能性を明らかにしてきた．また，その研究過程で，検索サイト Google の SERP（ search engine results 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があることを発見して，検索サイトの信用性（trustworthiness）に関する研究も発表してきた．この一連の研究で，我々が次に行わなければならないとは「ウェブマイニングは社会科学の新しい研究方法論」という，これまでの研究を通して得た発見を確たるものとすることである．我々は，このような研究を進めるには，単にデータベースエンジニアがウェブマイニングツールを構築して何かを検証しようとしても，ドメイン知識の欠落ゆえに， その真価を問えないことによりそれ以上研究が進捗しないことを認知し，特に社会学で活発に研究が行われているジェンダー分野に焦点を当てて，ジェンダーに関するドメイン知識を豊富に有する者を研究チームの主要メンバーとして擁して研究を遂行することにより，理工学の域を超えた研究成果を得ることに成功してきた．換言すれば，文理融合した研究体制を整えることにより，初めてウェブマイニングの分析結果や SERP Ranking の信頼性を的確に判断することが可能となり，研究が進展する一方，そこで得られた知見をエンジニアにフィードバックすることにより，真に有用な工学的進展が達成されたのである．

２．研究の目的

ウェブには実世界のさまざまな出来事が写し込まれている．実世界は時間の経過と共に時々刻々と変化しているので，その変化をウェブをマイニングすることにより摑まえることができるならば，それを手がかりとして，実世界で一体何が起こっているのかをタイムリーに知ることができるのではないかと考えられる．そこで我々は，実世界の出来事はそれに関連する検索キーワードによる検索エンジン結果ページ（search engine results page, SERP）のランキングに変動を与えるという知見に基づき，利用者（社会科学分野の研究者）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 SERP を定期的に収集し，SERP 順位の変動を分析し，その変容により警告を発して，利用者に調査を促すシステムを開発することとした．それが SERPWatcher である． これまで，社会科学分野ではアンケート調査，インタビュー調査，あるいは実地調査が主たる社会調査法として知られているが， SERPWatcher はこれらとはまったく異なる新しい社会調査法，つまり社会科学の新しい研究方法論となるであろう．

３．研究の方法

(1) SERPWatcher の設計

SERP Watcher は次の機能を有しないといけない．

① SERP 収集
② SERP データクリーニング
③ 多次元データベースとしての SERP Archive 構築
④ SERP Archive の多次元分析と表示
⑤ ニュースなどの関連情報収集
⑥ アラートとマイニング
⑦ 検索キーワード登録
⑧ ユーザ登録

我々は，上記機能を有する SERPWatcher を設計・開発し，特にジェンダー学際分野の研究者により，SERPWatcher プロトタイプの社会学分野における新しい研究方法論としての有用性を検証する．図 1 に開発した SERPWatcher のシステム概念図を示す．

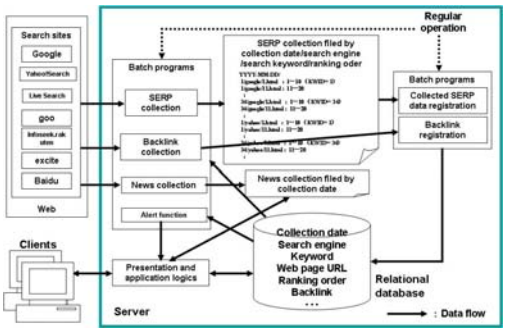


図 1　SERPWatcher のシステム概念図

(2) SERP 収集とデータクリーニング

ユーザが指定する検索キーワードについて，7 種類の検索エンジン(Google, Yahoo! 検索，Live Search, goo, Infoseek 楽天，excite, Baidu)からそれらが提供するウェブサービスを使い，1 週間に一度の割合で，上位 500 位までのウェブページに関するさまざまなデータを収集し，データクリーニングを行い，リレーショナル DBMS である MySQL に格納して，SQL による検索とアプリケーションプログラムからのアクセスを可能とした．現在，ジェンダー学際分野に関連する約 35 程度のキーワードが 登録されており，古く登録された検索キーワードのついては，2007 年夏頃からアーカイブデータが収集され研究に供されている．

(3)SERP Archive の多次元データベース構成

SERP Watcher が取り扱うデータを一元的に管理するためのデータベースは，検索キーワード，検索エンジン，(SERP の）収集日，(収集された)ウェブページデータであり，そ

れは図 2 に示すように多次元データとなる．

D_Keyword
　　Keyword
　　KeywordSpecifier
D_SearchEngine
　　SearchSiteName
　　SearchSiteURL
D_Date
　　Year
　　Month
　　Day
D_SERP
　　RankingOrder
　　Title
　　Snippet
　　URL
　　Backlinks
　　NumberOfBacklinks

図 2　SERP Archive の 4 次元 DB 構成

(4)　SERP Archive の多次元分析と表示

SERPWatcher の利用法を社会学を専門とする我々の研究分担者とともに分析をすると，分析は 3 つのタイプの観点から主として分析を行うことを突き止めた．それらは次のとおりである：

① {web page URL, collection date}→{ranking order of the web page}
② {web page URL, search engine}→{ranking order of the web page}
③ {search engine, collection date}→{ranking order of the web page}

その結果，SERPWatcher の分析のための Archive 表示法は，上記①，②，③に対応して，次の 3 つであることが明らかになった：

①検索エンジン固定ビュー
②収集日固定ビュー
③ウェブページ固定ビュー

図 3 にそれらを示す．

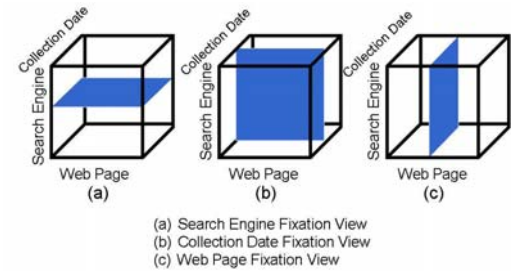


図 3　SERP Archive の多次元分析

４．研究成果

SERPWatcher の実装と社会学者によるそのプロトタイプの評価について述べる．

(1) SERPWatcher の主機能の実装

プロトタイプシステムはクライアント - サーバ方式で実装されている．サーバには Red Hat Enterprise Linux 5.1 (x86/x86_64) が使われ，開発用のプログラミング言語として Ruby 1.8.5 (2006-08-25) [i386-linux] と Ruby on Rails 2.1 が使われた．データベース管理システムは MySQL 5.0.45 である．前章で述べた 3 つのビューは図 4，図 5 のようである．

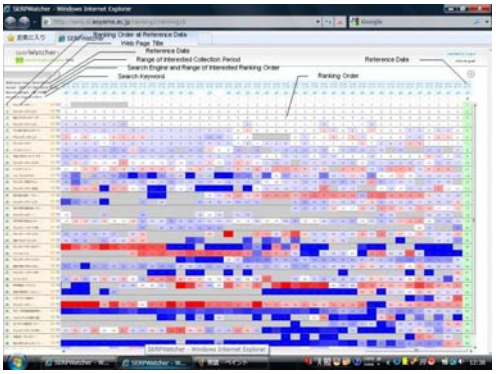

図 4　検索エンジン固定ビューの画面例

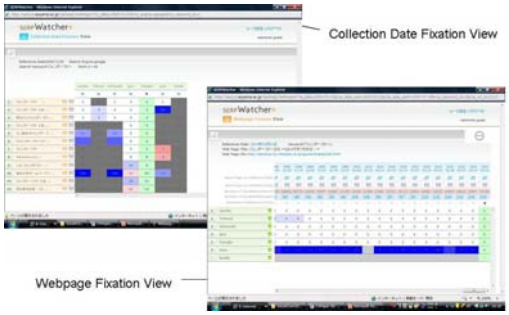


図 5　収集日固定ビューとウェブページ固定ビューの画面例

なお，ウェブページの順位変動を順位変動の大きさに比例して着色して示すために，図 6 に示す色付け関数を定義し，実装している．

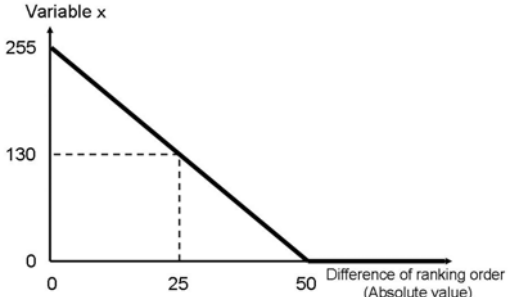


図 6　SERPWatcher の 8 ビットカラーシェイディングアルゴリズム

(2) SERPWatcher のアラート機能の実装

指定された検索キーワードで定期的に SERP を収集していくとき，SERP 順位の時系列的変化にあらかじめ決めていた以上の変化が生じたときにユーザに変化が起きたことを知らせる機能を「アラート機能」と呼んでいる．図 7 にユーザへのアラート e-mail の一例を示す．

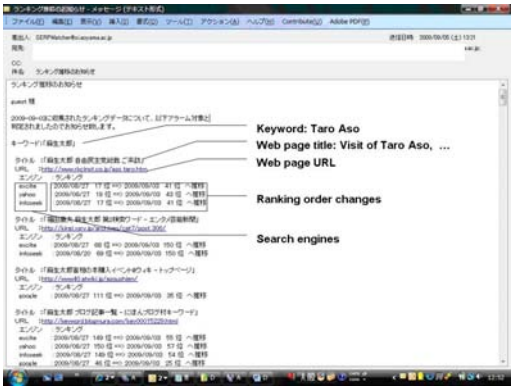


図 7　アラート e-mail の一例

(3) SERPWatcher 機能の検証

本研究で開発した SERPWatcher のプロトタイプを使用して，システムの有用性を検証した．まず，さまざまな分析を行った結果として，検索エンジン固定ビュー，収集日固定ビュー，ウェブページ固定ビューという 3 種類のビューを定義・実装し，それら 3 つのビューの間を分析の視点に応じて自由に行き来できるように実装した．約 1 年半にわたり 37 個のジェンダー学際関連の検索キーワードでその機能の有用性を検証した結果，当初の目的通り，SERPWatcher は社会科学の分野で，これまでのアンケート調査やインタビュー調査に代わる全く新しい研究方法論となるであろうという確信を得ることができた．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）


① Yoshifumi Masunaga, Naoko Oyama, Chiemi Watanabe, Kazunari Ito, Kaoru Tachi, and Yoichi Miyama, SERPWatcher: A SERP Mining Tool as a Novel Social Survey Method in Sociology, Database Systems for Advanced Applications, 査読有, DASFAA2010 Proceedings, Part II, Springer LNCS 5982, 2010, 412-415

② 増永良文，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，青山社会情報研究，査読有，1 巻，2009，1-18

③ Naoko Oyama, Yoshifumi Masunaga, On the Trustworthiness and Transparency of a Web Search Site examined using “Gender-equal” as a Search Keyword, Progress in WWW Research and Development, 査読有, Proceedings of APWeb2008, Springer LNCS 4976, 2008, 625-630


〔学会発表〕（計 4 件）


① 中部 文子，渡辺 知恵美，小山 直子，舘 かおる，増永 良文，社会調査支援の為の SERPWatcher からのランク変動特徴抽出，DEIM2010 論文集，査読無， 2010，A2-5

② Yoshifumi Masunaga, Naoko Oyama, Chiemi Watanabe, Kazunari Ito, Kaoru Tachi, and Yoichi Miyama, SERPWatcher: A SERP Mining Tool as a Novel Social Survey Method in Sociology -- Demonstration --, 第 15 回先進応用のためのデータベースシステムに関する国際会議（DASFAA2010） デモストレーション，査読有，April 2, 2010， 筑波大学

③ 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－SERP Watcher β版の開発－，DEIM2009 論文集，査読無， 2009，D7-5

④ 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 論文集，査読有，2008，A1-5


〔図書〕（計 1 件）


① 増永良文，サイエンス社，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，2008，245 ページ


〔その他〕

学会発表②は第 15 回先進応用のためのデータベースシステムに関する国際会議（DASFAA2010）のデモストレーション部門で Excellent Demonstration Awardを受賞した．22 件のデモンストレーションから 1 件の Best Demonstration Award と 2 件の Excellent Demonstration Award が表彰された．受賞は，本研究で開発してきた SERPWatcher のシステムの完成度やそれが社会科学の新しい研究方法論となるであろうという先進性がとても高く評価された結果である．より詳細な記述が青山学院大学のホームページに掲載されている（http://www.aoyama.ac.jp/news/439.html）．

６．研究組織

(1)研究代表者
増永　良文（MASUNAGA YOSHIFUMI）
青山学院大学・社会情報学部・教授
研究者番号：70006261

(2)研究分担者
舘　かおる（TACHI KAORU）
お茶の水女子大学・大学院人間文化創成科学研究科・教授
研究者番号：50155082

小山　直子（OYAMA NAOKO）
お茶の水女子大学・ジェンダー研究センター・客員研究員
研究者番号：00194639

渡辺　知恵美（WATANABE CHIEMI）
お茶の水女子大学・大学院人間文化創成科学研究科・講師
研究者番号：20362832

伊藤　一成（ITO KAZUNARI）
青山学院大学・社会情報学部・助教
研究者番号：20406812

喜連川　優（KITSUREGAWA MASARU）
東京大学・生産技術研究所・教授
研究者番号：40161509
（H20→H21 連携研究者）

(3)連携研究者
竹内　純人（TAKEUCHI SUMITO）
青山学院大学・情報科学研究センター・助手
研究者番号：60464799