# 科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１３　日現在

機関番号：　６２６１５
研究種目：　基盤研究(A)
研究期間：　2010～2013
課題番号：　２２２４０００７

研究課題名（和文）Ｗｅｂ情報構造と利用者行動の統合分析とその情報アクセス高度化への応用

研究課題名（英文）Integrated analysis of web information structure and users' behavior and its application to advanced information access

研究代表者
大山　敬三（OYAMA, Keizo）

国立情報学研究所・コンテンツ科学研究系・教授

研究者番号：９０１７７０２２

交付決定額（研究期間全体）：（直接経費）　　30,400,000円、（間接経費）　　9,120,000円

研究成果の概要（和文）：Webの構造や利用者の情報検索・閲覧行動を総合的に理解し，応用として展開するため，Web閲覧ログデータやマイクロブログデータ等，Web情報構造とWeb利用者行動に関連する様々なデータを収集・導入し，アンケート調査とも連動させることにより，これらを統合・分析した。
その結果，知りたい情報と知らせたい情報との間の乖離や，Webポータルサイトを利用することにより意図しない情報接触行動が生ずることなど，Web利用者に関する様々な知見が実証的に得られた。また，統合・分析により得られた情報に基づいて，情報推薦や情報検索等の情報アクセスを高度化するための様々な手法を研究・提案した。

研究成果の概要（英文）：For understanding Web structure and users' behavior of information retrieval and browsing in an integrated way, and for extending it to various applications, we collected and introduced various data reflecting Web information structure and Web users' behavior (e.g. Web view log data, micro-blog data), obtained user data through questionnaire, and executed integrated analysis on them.
Consequently, we obtained various findings through data such that there is a gap between information wanted to know and information wanted to inform, and that, through using Web portal sites, unexpected contact to various information occurs. Moreover, we proposed and studied various methods for advanced information access such as information recommendation and information retrieval based on the information obtained through the integrated analysis.

研究分野：　総合領域

科研費の分科・細目：　情報学・メディア情報学・データベース

キーワード：　Ｗｅｂ情報構造　データ統合　情報アクセス　Ｗｅｂ利用者行動　Ｗｅｂ閲覧履歴　マクロブログデータ　Twitterデータ　アンケート調査

１．研究開始当初の背景

(1) Web に関しては，クローリングデータを用いたリンク解析等の構造解析や，サーチエンジンの検索ログ，サーバログあるいはクライアントログなどの閲覧履歴を用いた利用者の情報検索・閲覧行動分析など，データに基づく様々な研究が行われてきた。一方で，利用者行動分析については被験者モニタリングやアンケート調査などに基づく研究も行われている。研究代表者及び連携研究者らも，これまでそれぞれに，Web 構造解析，検索・分類手法，情報抽出・活用，マイニング，利用者行動分析等の多様な研究を進めてきた。

(2) しかし，これらの従来研究の多くは，単一データまたは比較的親和性の高いデータの組合せに基づいており，Web の特性の限られた側面しか捕らえることができず，また関連分野への応用展開も困難である。

(3) そこで，Web の構造や利用者の情報検索・閲覧行動を総合的に理解し，その成果を高度な情報アクセス技術の実現やその他の応用に展開するため，従来別々に収集・解析されてきた Web 情報構造と Web 利用者行動に関連する様々なデータを統合する総合的なアプローチをとることが求められていた。

２．研究の目的

(1) 本研究では，まず Web 情報構造や Web 利用者行動に関連する様々なデータを収集し，それらを相互に照合・対応付けして統合した後，Web の利用者行動の側面を含め総合的観点から信頼性の高い Web 情報構造の分析を行い Web の実態を解明し，次に情報アクセスに関連する各種の具体的応用としての観点から統合データを詳細に分析してその結果を情報アクセス技術等に応用して高度化を図ることを目的とする。

(2) 本研究はデータに基づいた分析を基本とするので，クローリングデータを始めとする Web 構造を反映した複数のデータや，Web 閲覧履歴データのような Web 利用者行動を反映した複数のデータを収集し，URL や検索語の抽出・照合，アクセスや検索の頻度分析等を通じて，利用者の同定や検索・閲覧・発信の時系列化およびセッションの抽出等を含む統合データを構築するための手法を開発する。

(3) 統合したデータに対し，各種モデルに従って統計的分析や共通のデータ解析を行う。統合データを総合的に用い，検索・閲覧について総合的な利用者挙動の分析を行いて，既存の情報探索行動の研究結果との比較・検討を行う。

(4) 統合したデータから利用者モデルを構築し，個々の利用者の背景知識やコンテクストを推定する手法を開発する。また，これらに基づく情報推薦や知識検索等，情報アクセス関連技術の高度化に応用するための研究を行う。

(5) さらに，総合的 Web 情報構造分析と自然言語処理を組み合わせることで，Web 上の言説空間の分断化の研究などの情報発信者行動の実証的分析において，情報発信者だけでなく閲覧者の視点を加えることにより，社会科学研究における応用可能性を拡張する。

３．研究の方法

(1) 国立情報学研究所の研究者が中心となって，研究の基盤となる各種データの導入・収集を以下のように行った。

① 稼働中の Web クローラにより，継続してデータを収集し，各 Web ページの URL，参照元 URL，アンカーテキスト等の抽出，PageRank 等のリンク解析等を行った。

② パネル約 20,000 名による Web 閲覧ログデータ（閲覧先及び参照元の URL，アクセス時刻，閲覧時間，閲覧者プロファイル等が含まれる）を導入し，Web 上の主要なサービスに対するアクセスログの抽出・分析等を行った。

③ 研究用に公開されている知識検索サイトの質問回答データおよび Web 検索サイトの検索語頻度データを導入し，上記 Web 閲覧ログデータとの突き合わせ等を含む統計分析を行った。

④ ブログサービスを運用している Web 調査会社に委託し，利用者が行った具体的な投稿に対応させて，記事 URL，意図，記事タイプ，内容種別等，及びトラックバックに関する意図や種別等のデータ，並びに利用者プロファイル等に関するデータをアンケート調査により収集した。

⑤ Twitter の API を利用して低サンプル率のサンプルデータを収集するとともに，サンプル中の任意のツイートに対してリプライ関係にあるツイート（ツイート連鎖）を収集する手法を開発した。また，事例として衆議院選挙等を対象として，関連のツイート連鎖を収集するためのフィルタ条件を構築し，Twitter データ提供会社に委託して大規模データの収集を行った。

(2) 国立情報学研究所の研究者が中心となり，適宜連携研究者からの協力を得ながら，上記で導入・収集したデータの統合を行った。

① 各データ中の URL，タイムスタンプ等を

正規化し，相互に照合可能とした。

② 前記情報に基づき，各データに付与された利用者 ID 間の対応付を行った。

③ データ中で参照されている URL をクロールし，本文テキストを抽出して文書ベクトルとして蓄積した。

④ 各データ中の検索語を正規化し，照合を行うとともに，検索語と閲覧ページを対応づけた。

⑤ 各データを統計的に分析して一致度や代表性を検証し，データの収集方法等に問題がないことを確認した。

(3) 統合データを用いて，本研究に参加した各研究者が，情報アクセス技術等に関する各種の応用を高度化するための研究を行った。また，全体研究会を開催し，情報と課題の共有を行って連携を深めた。

４．研究成果

(1) 各データの統合処理過程において様々な角度からデータ間の関連等を分析し，次のような知見を得た。

① Twitter データと Web 閲覧ログデータにおいて，Web ニュース閲覧中のマイクロブログ（Twitter）上での発信を分析し，例えば，記事に関するツイートは必ずしも記事内容に対する関心や利用者の日常の興味を反映しておらず，知りたい情報と知らせたい情報の間に乖離があることを明らかにした。

② Web 検索データと Web 閲覧ログデータにおいて，検索クエリには一定の言語的修飾構造があることを示し，Web 検索行動と検索式の関係を分析した。

(2) 統合データを利用した，情報アクセス技術等の高度化に関して，下記のような成果を得た。

① Twitter 上の言及に基づくニュース記事推薦手法について，ツイート発信者の背景知識に基づくプロファイルと，ツイート発信の前後時刻における Web 閲覧およびツイート発信履歴に基づく文脈とを用いて，閲覧者の興味により適したニュース記事を推薦する手法を提案した。

② マイクロブログと知識検索サイトとの効果的な連携の実現を目指して，Twitter からの質問応答型知識の獲得や潜在的情報要求の抽出の手法を提案した。

③ Web 閲覧行動を利用して情報アクセスを高度化するため，Web 閲覧行動の統計情報に基づく可視化および推薦に関する研究を進めた。

④ 統合データによる社会科学研究の応用として，ブログサイトでの情報交換における情報発信者の行動について，情報発信者アンケート調査等を行い，その結果を一般利用者の Web 閲覧ログデータ等と対応付けて分析することにより，より客観的な観点のデータを取り込んで，情報発信者行動の実証的分析をおこなうための手法を開発した。

(3) 本研究で採用したデータ統合手法や統合データの分析手法をより高度化するための基本技術の研究として，以下のような成果を得た。

① 利用者（著者）の同定において ID，あるいは URL やタイムスタンプ等の情報を補完するための手法として，テキストの特性に基づき著者推定を行うための手法を開発した。

② 統合データを表現するためには複数の異なるタイプのエンティティ間の関係を表現するグラフが有用であり，これに適用可能なグラフ分析の新たな手法の研究を進めた。

③ 利用者の背景知識をプロファイル化するために必要となる技術として，文書の高速クラスタリング手法の研究を進めた。

④ 統合データから得た情報に基づき複数情報源へのアクセスを効率的に実現するための手法として，Web サービスの部品化と再構成のための手法の研究を進めた。

(4) 以上のように多くの研究成果が得られており，多様なデータを統合して情報アクセスに活用するという本研究のアプローチの有効性が示された。本研究の終了時点ではこのような研究アプローチが一般化してきており，本研究が先導的役割を果たしたと評価できる。

(5) 今後は本アプローチをより発展させ，具体的な情報アクセスシステムの実現とともに，より大規模なデータへの対応や非構造データへの対応を進めていく必要がある。

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６４件）

① Tetsuro Kobayashi and Kazunori Inamasu: "The Knowledge Leveling Effect of Portal Sites", Communication Research, OnlineFirst Version, 25 May 2014, p.1-21 (2014). (DOI: 10.1177/0093650214534965)（査読あり）

② Sorn Jarukasemratana, Tsuyoshi Murata, Xin Liu: "Community Detection Algorithm based on Centrality and Node Closeness in Scale-Free Networks", 人工知能学会論文誌 Vol.29, No.2, pp.234-244 (2014). (DOI: 10.1527/tjsai.29.234)（査読あり）

③ 池松恭平, 村田剛志, "3 部モジュラリティの改善とその最適化手法", 人工知能学会論文誌 Vol.29, No.2, p.245-258 (2014). (DOI: 10.1527/tjsai.29.245)（査読あり）

④ Hao Han: "Extracting News from Server Side Databases by Query Interfaces", Journal of Computer Information Systems, Vol.54, No.2, p.57-65 (2014).（査読あり）

⑤ Hao Han, Yinxing Xue, Keizo Oyama: "Mashup technology: Beyond open programming interfaces", Computer, Vol.46, No.2, p.96-99 (2013). (DOI: 10.1109/MC.2013.429)（査読あり）

⑥ Hao Han, Hidekazu Nakawatase, and Keizo Oyama: "Context Oriented Analysis of Interest Reflection of Tweeted Webpages based on Browsing Behavior", Proceedings of 15th International Conference on Information Integration and Web-based Applications & Services (iiWAS2013), p.34-43 (2013). (DOI: 10.1145/2539150.2539181)（査読あり）

⑦ Kyohei Ikematsu and Tsuyoshi Murata: "A Fast Method for Detecting Communities from Tripartite Networks", Proceedings of the 5th International Conference on Social Informatics (SocInfo 2013), LNCS 8238, p.192-205. (2013) (DOI: 10.1007/978-3-319-03260-3_17)（査読有り）

⑧ Mitsumasa Kubo, Ryohei Sasano, Hiroya Takamura and Manabu Okumura: "Generating Live Sports Updates from Twitter by Finding Good Reporters", Proceedings of Web Intelligence 2013, p.527-534 (2013). (DOI: 10.1109/WI-IAT.2013.74)（査読あり）

⑨ 塚本悠馬, 笹野遼平, 高村大也, 奥村学："マイクロブログ上の告知投稿に対する非明示的な関連投稿の収集", 情報処理学会研究報告. 自然言語処理研究会報告, 2013-NL-214(14) (2013)（査読なし）

⑩ 韓浩, 中渡瀬秀一, 大山敬三: "ウェブページのツイート行動への関心反映度に関するブラウズ行動のコンテクストに注目した分析", 情報処理学会研究報告. 情報学基礎研究会報告, Vol.2013-IFAT-111, No.28, p.1-6 (2013).（査読なし）

⑪ 内藤慎也, 江口浩二："閲覧履歴グラフに基づく正則化リンク解析を用いたロバスト推薦", 日本データベース学会論文誌, Vol.12, No.1, p.7-12 (2013),（査読あり）

⑫ Xin Liu, Tsuyoshi Murata, Ken Wakita: "Extracting the multilevel communities based on network structural and nonstructural information", Proceedings of the 22nd international conference on World Wide Web (WWW 2013 (poster)), pp.191-192 (2013).（査読あり）

⑬ Hao Han, Hidekazu Nakawatase, and Keizo Oyama: "An Exploratory Analysis of Browsing Behavior of Web News on Twitter", Proceedings of 2012 ASE/IEEE International Conference on Social Informatics (ICSI 2012), p.86-95 (2012). (DOI: 10.1109/SocialInformatics.2012.41)（査読あり）

⑭ 中渡瀬秀一, 大山敬三: "短文投稿型 Web サービスからの Q&A 型知識抽出に向けて", 電子情報通信学会技術研究報告（思考と言語研究会）, Vol.112, No.339 (TL2012-36), p.13-16 (2012).（査読なし）

⑮ Sorn Jarukasemratana and Tsuyoshi Murata: "Visualizing Web Structure based on Browsing Sessions", Proceedings of the 10th Asia Pacific Conference on Computer Human Interaction (APCHI2012), 2p. (2012)（査読あり）

⑯ Hao Han, Yinxing Xue, and Keizo Oyama: "Client-Side Rendering Mechanism: A Double-Edged Sword for Browser-Based Web Applications", Proceedings of 24th International Conference on Software Engineering and Knowledge Engineering (SEKE 2012), p.124-130 (2012).（査読あり）

⑰ 中渡瀬秀一, 大山敬三: "Web 視聴記録等を用いた利用者行動因子の抽出", 電子情報通信学会技術研究報告（人工知能と知識処理研究会）, Vol.112, No.94 (AI2012-06), p.31-32 (2012).（査読なし）

⑱ Shinya Naito and Koji Eguchi, "Robust Recommendations using Regularized Link Analysis of Browsing Behavior Graphs", Social Computing, Behavioral-Cultural Modeling and Prediction, Lecture Notes in Computer Science, Vol.7227, p.339-347 (2012). (10.1007/978-3-642-29047-3_41)（査読有り）

⑲ 井上雅翔, 山名早人："品詞 n-gram を用いた著者推定手法－話題に対する頑健性

の評価", 日本データベース学会論文誌, Vol.10, No.3, pp.7-12 (2012). (査読あり)

⑳ Manabu Okumura, Tetsuya Motegi, Tetsuro Kobayashi, Keizo Oyama, Takahisa Suzuki: "Can We Predict Political Poll Results by Using Blog Entries?", Proceedings of 2012 45th Hawaii International Conference on System Sciences (HICSS-45), p.1785-1789 (2012). (DOI: 10.1109/HICSS.2012.145) (査読あり)

㉑ Xin Liu and Tsuyoshi Murata, "Detecting Communities in K-Partite K-Uniform (Hyper)Networks", Journal of Computer Science and Technology, Vol.26, No.5, pp.778-791 (2011). (DOI:10.1007/s11390-011-0177-0) (査読あり)

㉒ 中渡瀬秀一, 大山敬三: "検索クエリを用いた情報の下位範疇化", 電子情報通信学会技術研究報告, Vol.111, No.70 (AI2011-3), p.13-14 (2011). (査読なし)

㉓ Tsuyoshi MURATA: "A New Tripartite Modularity for Detecting Communities", Computer Software, Vol. 28, No. 1, pp. 154-161 (2011). (DOI: 10.11185/imt.6.572) (査読あり)

㉔ Swit Phuvipadawat, Tsuyoshi MURATA: "Detecting a Multi-Level Content Similarity from Microblogs based on Community Structures and Named Entities", Journal of Emerging Technologies in Web Intelligence, Vol.3, No.1, p.11-19 (2011). (DOI:10.4304/jetwi.3.1.11-19) (査読あり)

㉕ 中渡瀬秀一, 大山敬三: "大規模閲覧記録に基づく利用者の Web 検索行動と検索式の調査", 人工知能学会「社会における AI 研究会」第 11 回研究会 (2011). (査読なし)

㉖ 中渡瀬秀一, 大山敬三: "検索クエリにおける修飾構造の調査", 電子情報通信学会技術研究報告, Vol.110, No.407(TL2010-54), p.49-52 (2011). (査読なし)

〔学会発表〕(計１９件)

① 中渡瀬秀一, 大山敬三: "マイクロブログメッセージからのグランブル抽出", 情報知識学会第 22 回 (2014 年度) 年次大会, 和歌山大学システム工学部 (2014.05.24).

② 奥谷貴志, 山名早人:"メンション情報を利用した Twitter ユーザプロフィール推定", DEIM2014, 淡路夢舞台＆ウェスティン淡路, 淡路島 (2014.3.3).

③ 中渡瀬秀一, 大山敬三: "ツィートテキストからの Q&A 型知識の抽出", 第 3 回知識共創フォーラム, 北陸先端科学技術大学院大学, 東京 (2013.03.02).

〔図書〕(計２件)

① Hao Han, Peng Gao, Yinxing Xue, Chuanqi Tao, and Keizo Oyama: "Analysis and Design: Towards Large-Scale Reuse and Integration of Web User Interface Components", p.133-162, in Ozyer, T.; Kianmehr, K.; Tan, M.; Zeng, J. (Eds.): "Information Reuse and Integration In Academia And Industry", Springer, 306p. (2013).

② Tsuyoshi MURATA: "Detecting Communities in Social Networks", p.269-280, in "Handbook of Social Network -- Technologies and Applications", Springer, 734p. (2010).

６．研究組織

(1)研究代表者

大山　敬三 (OYAMA, Keizo)
国立情報学研究所・コンテンツ科学研究系・教授
研究者番号：９０１７７０２２

(2)研究分担者

該当なし

(3)連携研究者

相澤　彰子 (AIZAWA, Akiko)
国立情報学研究所・コンテンツ科学研究系・教授
研究者番号：９０２２２４４７

宮尾　祐介 (MIYAO, Yusuke)
国立情報学研究所・コンテンツ科学研究系・准教授
研究者番号：００３４３０９６

孫　媛 (SUN, Yuan)
国立情報学研究所・情報社会相関研究系・准教授
研究者番号：００２４９９３９

小林　哲郎 (KOBAYASHI, Tetsuro)
国立情報学研究所・情報社会相関研究系・准教授
研究者番号：６０４５５１９４

韓　浩 (HAN, Hao)
神奈川大学・理学部・特別助手
研究者番号：２０６００８０２

岸田　和明 (KISHIDA, Kazuaki)
慶應義塾大学・文学部・教授

研究者番号：９０２３４２１０

山名　早人（YAMANA, Hayato）
早稲田大学・理工学術院・教授
研究者番号：４０２３０５０２

奥村　学（OKUMURA, Manabu）
東京工業大学・精密工学研究所・教授
研究者番号：６０２１４０７９

吉岡　真治（YOSHIOKA, Masaharu）
北海道大学・情報科学研究科・准教授
研究者番号：４０２９０８７９

石田　栄美（ISHITA, Emi）
九州大学・附属図書館研究開発室・准教授
研究者番号：５０３６４８１５

村田　剛志（MURATA, Tsuyoshi）
東京工業大学・情報理工学研究科・准教授
研究者番号：９０２４２２８９

江口　浩二（EGUCHI, Koji）
神戸大学・システム情報学研究科・准教授
研究者番号：５０３２１５７６