様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　５月２２日現在

機関番号：１３９０３
研究種目：　基盤研究（Ｃ）
研究期間：2010～2012
課題番号：　２２５００１２８
研究課題名（和文）閲覧者の観点によるＷｅｂ情報構造化のためのＷｅｂページ分割アルゴリズムの実現
研究課題名（英文）Realizing a Web Page Segmentation Algorithm considering Users Viewpoints
研究代表者
新谷　虎松（SHINTANI　TORAMATSU）
名古屋工業大学・工学研究科・教授
研究者番号：　00252312

研究成果の概要（和文）：Web 情報は、意味的な構造を持たないテキストであり、計算機を用いて情報の統合や検索をするためには、多くの課題を解決する必要がある。本研究では、既存の Web 情報を閲覧者の観点で構造化し、効果的な情報閲覧を支援するための新たな技術として、HTML を意味的な構造へと変換するための新たな Web ページ分割アルゴリズムを開発した。本アルゴリズムの応用として、エージェント技術に基づく知的 Web ブロック管理機構を実装した。本技術により、Web ページから特定の Web コンテンツを高い精度で収集可能になり、また、既存の Web コンテンツの再利用性を向上させ、Web ページの閲覧性を効果的に改善できる。

研究成果の概要（英文）：We need to solve a lot of problems to realize more practical information search and integration because information on the Web is represented as texts without semantics. In this study, we developed a new algorithm to translate existing Web pages into structured and semantic Web contents considering users viewpoints in order to support effective Web browsing. The algorithm can be applied to an implementation of a Web block management system. Our method can extract meaningful information from Web pages precisely and can improve the reuseness and readability of existing Web pages.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,100,000	330,000	1,430,000
2011 年度	1,200,000	360,000	1,560,000
2012 年度	900,000	270,000	1,170,000
総　計	3,200,000	960,000	4,160,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：知的エージェント、Web ページ分割

１．研究開始当初の背景

Web ページで代表される Web 技術の利点は、情報発信を容易にできることであるが、機械的な情報閲覧支援の実現という視点からは多くの課題がある。例えば、Web 情報は、意味的な構造を持たないテキストであり、計算機を用いて情報の統合や検索をするためには、多くの課題を解決する必用がある。Web 情報の解析技術に関連して、セマンティック Web の研究[参考文献１]があるが、現実

的な問題に対処するためには、さらに多くの時間が必用である。現状、既存の一般的 Web 情報の効果的な閲覧支援のためには、自然言語処理や機械学習技術などを基盤として、既存の Web 情報を閲覧者の観点で構造化するための新たな技術の確立が必要である。例えば、既存 Web 情報は、Web ページ検索システム、コンテンツフィルタリングシステム、情報抽出システム等で全文検索を行った際に、メインコンテンツ以外の文字情報がノイズとなり、精度が低下する。

Web ページから特定のコンテンツを抽出する手法として、Web ラッパーがある。Web ラッパーの欠点は、 HTML 構造の変化に脆弱で、記述形式が多少でも変化するたびに再構築する必要がある点である。 Web ラッパーを自動で生成する研究も盛んに行われているが、適用できる Web ページに制限が多く，また精度も不十分である。近年、必要な部分を抽出するために Web ページをまとまったコンテンツ単位に分割する Web ページ分割アルゴリズムの研究が盛んである。一般に、Web ページ分割の解釈は，人間ですら一貫性がないため，機械学習やヒューリスティクスを用いた手法など、様々な試みがある。

本研究は、ユーザの観点から Web ページの分割に着目した点、および Web 情報再利用機構において Web ブロックを様々な情報端末に適した形態で再利用する点が独創的である。本研究により、既存の Web コンテンツの再利用性が向上し、また Web ページの閲覧性を効果的に改善できる。

２．研究の目的

本研究の目的は、閲覧者が一般的に意味のあるかたまりとして認識する Web コンテンツ（例えば、ニュース記事、検索バー等）を自動抽出するために、HTML を意味的な構造へと変換する Web ページ分割アルゴリズムを実現することである。ここでの意味的な構造を Web ブロックと呼ぶ。Web 情報を、Web ブロックとそれらの関係を意味的にモデル化することで、より知的な Web 情報閲覧支援機構の実現が期待できる。また、類似する Web ページから類似コンテンツを自動抽出する技術が実現可能になる。具体的に、本研究では、提案アルゴリズムの有用性を実証するために、エージェント技術に基づく知的 Web ブロック管理機構の実現を目指す。エージェント技術を Web ブロック管理に応用することで、Web ページを効果的に Web ブロックへ変換し、再利用が可能になる。また、近年、携帯電話からの Web 閲覧が急増しており、既存の携帯電話も考慮した新たな Web コンテンツ作成技術や効果的な Web 情報閲覧支援技術に関連して知的な Web アプリケーションの開発技術の確立を目指す。

具体的には、本研究において、以下の 3 つの項目を明らかにする。研究項目１は、本研究のコアの部分であり、研究項目１の成果を用いて研究項目２を推進する。逆に研究項目２から研究項目１に関連して新たな手法を得る可能性もある。研究項目３は、本研究成果の有用性の実証実験として意味がある。

【研究項目１】Web ページ分割アルゴリズムの設計

Web ページ上の情報を閲覧者の観点から Web ブロックに分割する Web ページ分割アルゴリズムを設計する。Web ページ上の意味ある情報単位として Web ブロックを定義する。Web ブロックには静的なコンテンツだけでなく、Web サービスのような動的なコンテンツも考慮する。Web ブロックはユーザの何らかの好みや意見を反映していると考えられるため、ユーザプロファイルの構築や推薦アルゴリズムなどについて検討することで、その性質を明らかにする。人間のように Web ページを断片化するために必要な知見をヒューリスティクスとして蓄積する。また、Web ブロックから Web ブロック間の関係を推測するためのアルゴリズムを設計する。これらのアルゴリズムを用いて、Web ページから Web ブロックを知的に抽出するアルゴリズムを実現し、その有効性を明らかにする。

【研究項目２】Web エージェントに基づく Web ブロック管理機構の設計

Web 情報を閲覧者の観点から構造化し、それを適切な時間で処理・再利用するために、Web エージェントを利用した Web ブロック管理機構を実現するための基礎技術を開発する。従来のマルチエージェントシステムの実装のように粗結合なシステムが適切なのか、運用の容易さを考慮してクラウドコンピューティング上に、マルチエージェントシステムを構築すべきかを明らかにする。

【研究項目３】モバイルエージェント技術に基づく Web 情報再利用機構の試作と評価実験

生成した Web ブロックを携帯電話やスマートフォンなどの多様な情報端末に適したレイアウトやデータフォーマットに変換し、Web ブロックの再利用性を高めるための実装技術を確立し、Web 情報再利用機構を試作・評価する。ここでは、モバイルエージェント技術を応用して、Web ブロックを携帯電話用データとそのデータの操作に適した携帯電話用のプログラムを含むモバイルコンテンツに変換する。これにより受信側の携帯電話に合わせて Web ブロックを最適化し、元の Web ページにあった属性や機能（スタイル、リンクやフォームなど）を再現することが可能になる

３．研究の方法

Web 情報を閲覧者の観点で構造化し、情報閲覧を支援するための新たな技術として、HTML を意味的な構造へと変換するための新たな Web ページ分割アルゴリズムの実現に向けて、次の３つの研究項目を推進した。研究項目１は、本研究のコアの部分であり、研究項目１の成果を用いて研究項目２を推進した。研究項目３は、本研究成果の実証実験である。

【研究項目１】Web ページ分割アルゴリズムの設計

【研究項目２】Web エージェントに基づく Web ブロック管理機構の設計

【研究項目３】モバイルエージェント技術基づく Web 情報再利用機構の試作と評価実験

以下、各項目について説明する。

（１）Web ページ分割アルゴリズムの設計

Web ページ分割アルゴリズムを設計するために、まずは、人間による Web ページ分割に関する調査を実施した。人間がどのように Web ページを意味あるブロック(Web ブロック)として分割しているのかを調査した。ここでは､多数の Web ページを人手によって Web ブロックへ分割させ、具体的なデータ（Web ページ分割事例）を収集した。Web ページの収集を目的として、独自の Web クローラーを開発した。本クローラーは、元の Web ページの収集に加え、効果的な表示処理のために予め Web ページをレンダリングし画像としてデータベースに保存するようにした。

次に、実験環境として、専用の Web アプリケーションを開発した。本アプリケーションは、Web ページのレンダリング結果の画像を被験者（研究室に所属する大学院生１０名程度）に提示し、簡単な操作で Web ページの分割方法や閲覧者の観点などを記録できるようにするシステムである。被験者には、Web ページの分割方法だけでなく、Web ページのそれぞれの断片の役割（メニュー、広告、タイトル、本文など)、および、断片間の関係（依存関係、無関係など）を簡便に記述させた。収集したデータに基づき、人間の Web ページ分割ための認識モデル、Web ページの断片の役割を推定するために必要な知見、および、Web ページの断片間の関係を推定するための知見を整理し､HTML を意味的な構造へと変換するための汎用的なヒューリスティクスとして蓄積した。

その後、アルゴリズムの設計を行った。最初に、Web ページ上の意味ある情報単位として Web ブロックを定義する。Web ブロックには静的なコンテンツだけでなく、Web サービスのような動的なコンテンツも考慮した。次に、Web ブロックの定義に基づく、Web ページ分割アルゴリズムの基盤を明らかにするために、予備的研究において、HTML におけるタグの構造だけでなく、レイアウトを考慮した Web ページ解析アルゴリズムに関する成果で得られた知見に基づき Web ページ上の情報を Web ブロックに分割する Web ページ分割アルゴリズムのコアを設計した。

本研究では汎用性を高めるための工夫として、上述のヒューリスティクスを考慮した新たな階層型の解析手法に基づくアルゴリズムを設計した。これにより、予備的研究での成果で課題であった汎用性の不足を解決することができた。

（２）Web エージェントに基づく Web ブロック管理機構の設計

Web 情報から Web ブロックを適切な時間で抽出・再利用するための Web ブロック管理機構を開発し､Web ブロックの組合せに基づく、知的な Web サービスの実現のための基盤を実装した。具体的には、Web ブラウザ上で、必要な情報をクリッピングする際に、Web ブロック抽出技術を利用することで、ユーザにとって容易な Web クリップ抽出、および、管理システムを実装した。また、Web 上から収集したニュース記事から Web ブロックを抽出し、ソーシャルメディア上の情報との関連づけを行うシステムを実装した。これらのシステムの実装において、Web ブロックを抽出するのみならず、Web ブロック間や他の情報源とのリンク付けを支援するための Web エージェントシステムを実装した。また、Web ブロックに対するアノテーションの付与を可能にすることで、Web ブロックの機械可読性を高め、必要に応じて Web ブロックを組み合わせて利用するための基盤を実装した。

（３）モバイルエージェント技術に基づく Web 情報再利用機構の試作と評価実験

Web ブロックをユーザの観点から構造化するためのテストベッドとして、携帯電話向けに Web 上の情報を構造化して配信するシステムを試作した。生成した Web ブロックを携帯電話、スマートフォン、タブレット端末などの多様な情報端末に適したレイアウトやファイルフォーマットに変換し、Web 情報の再利用性を高めるための実装技術を確立する。具体的には、モバイルエージェント技術に基づく Web ブロックのレイアウト最適化アルゴリズムとその操作プログラム生成アルゴリズムを設計し､実装･評価する。具体的には、モバイルエージェント技術を応用して、Web ブロックを携帯電話用データとそのコンテンツの操作に適した携帯電話用のプログラムを含むモバイルエージェントに変換する。これにより受信側の携帯電話に合わせて Web ブロックを最適化することが可能になる。

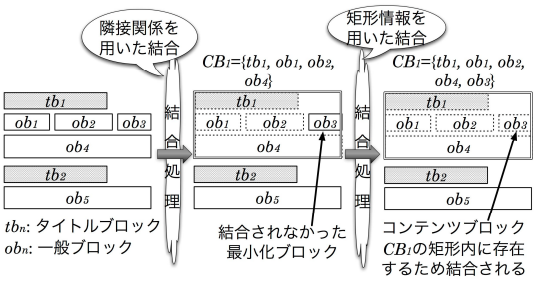

図 1　Web ブロック抽出のための結合ステップ

4．研究成果

Web ページを意味的なまとまりのあるコンテンツであるブロックに分割するためのアルゴリズムを実現した。本研究で開発した Web ページ分割手法の特徴は、図 1 のようにウェブページを細分化ブロックという単位まで分割した後に、Web コンテンツの見出しとなるようなブロック(タイトルブロック)に着目して細分化ブロックの結合を行うことにより、Web ページを意味的にまとまりのある単位へと分割する点である。

既存の Web ページ分割手法の多くが、面積や子ノード数など、コンテンツ量に依存する情報を用いて結合を行っていた。その結果、同一 Web サイト内の同じレイアウトの Web ページから異なる分割結果が得られるという問題が存在したが、本手法ではコンテンツ量に非依存な結合を行うために、この問題を効率的に解決している。まずは、分割処理に関して、細分化ブロックへの分割には、W3C が定義するブロックレベル要素を用いている。ブロックレベル要素とは、Web ページ上の矩形領域であり、子供の要素をその領域内に描画する要素である。次に、得られたブロックの役割を判定する。その後、タイトルブロックとそれに続くタイトルブロック以外のブロック（一般ブロック）を結合していく処理を行う。図 2 は、提案手法による Web ページ分割結果である。図 2 により、多様な Web ページに対して本 Web ページ分割手法が有効に機能していることがわかる。ここでは、ニュース記事、ブログ、e コマースサイト等の例を示している。

ここで重要な点としては、計算機によるタイトルブロックの抽出方法である。本研究では、計算機によるタイトルブロックの自動抽出を行うために，機械学習によって分類器を生成した。具体的胃は、J4.8 アルゴリズムによる決定木学習によって生成した分類器により、表 1 に示されるように、タイトルブロックおよび一般ブロックの抽出性能が、F 値で 77.8%および 89.3%となり、十分な性能を達成できた。

得られたタイトルブロックを用いて細分化ブロックの結合を行った結果、ニュースサイトのニュース記事部分に着目した場合、96.1%の精度でコンテンツ量に依存しない同一の分割結果が得られることを確認した。

図 2　提案手法による Web ページ分割結果

表 1　タイトルブロックの判定性能

	J4.8	RT	SVM
a:タイトルブロックを正しく判定した数	588	593	424
b:一般ブロックを正しく判定した数	1401	1338	1395
c:タイトルブロックを誤判定した数	194	189	358
d:一般ブロックを誤判定した数	141	204	147
P_{tb}: タイトルブロックの判定精度	0.807	0.744	0.743
R_{tb}: タイトルブロックの再現率	0.752	0.758	0.542
F_{tb}: タイトルブロックの F 尺度	0.778	0.751	0.627
P_{ob}: 一般ブロックの判定精度	0.878	0.876	0.796
R_{ob}: 一般ブロックの再現率	0.909	0.868	0.905
F_{ob}: 一般ブロックの F 尺度	0.893	0.872	0.847

上記手法を用いて、Web 情報から Web ブロックを適切な時間で抽出・再利用するための Web ブロック管理機構を開発し、Web ブロックの組合せに基づく、知的な Web サービスを試作した。具体的には、Web ニュースから重要な Web ブロックを抽出し、得られた Web ブロックと関連する情報を Linked Open Data として保存し、エンティティ同士をリンクして表示するシステムをタブレット端末上で実装し評価した。また、Web ブロック管理機構で得られた知見を活用した、Web サービスとして既存 Web ページ同期編集機構を試作し実際に運用している。

本研究の成果を、査読付き論文誌 4 編、および、国際会議 16 編、図書 1 冊として、発表し、国内外で高い評価を得ている。また、本研究により、Web ページ分割手法およびその応用に関する知見、ノウハウ、および、ソフトウェア資産を蓄積することができた。

５．主な発表論文等（全て査読あり）

〔雑誌論文〕（計４件）

1. 鈴木亮詞，村瀬隆拓，白松俊，大囿忠親，新谷虎松:タブレット端末のためのスマートサイネージシステムの実装について，コンピュータソフトウェア，Vol.30, No.2, pp.176-190, 2013/05.
2. 佐野博之，白松俊，大囿忠親，新谷虎松: Web ページ分割のための決定木学習を用いたタイトルブロック抽出.電子情報通信学会情報・システムソサイエティ和文論文誌，Vol. J95-D, No.4, pp. 909-918, 2012.
3. 平田紀史，白松俊，大囿忠親，新谷虎松，〝ユーザの観点に基づくイベント系列化を用いた Web ニュース記事閲覧支援システムの実装 ", 人工知能学会論文誌, Vol. 26, No. 1, pp.228--236, 2011.
4. Robin Swezey, Shun Shiramatsu, Tadachika Ozono and Toramatsu Shintani, "Intelligent Page Recommender Agents: Real-Time Content Delivery for Articles and Pages Related to Similar Topic", Lecture Notes in Computer Science, Vol. 6704, 173-182, 2011.

〔学会発表〕（計１６件）

1. Shin-ya Katayama, Takushi Goda, Shun Shiramatsu, Tadachika Ozono and Toramatsu Shintani, Fast Synchronization Mechanism for Collaborative Web Applications based on HTML5, In Proc of the 14th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2013), 2013（採録済発表予定）.
2. Ryota Inoue, Yudai Kato, Takushi Goda, Shun Shiramatsu, Tadachika Ozono, Toramatsu Shintani, "A Real-Time Collaborative Mechanism for Editing a Web Page and its Applications ", In Proceedings of The 2012 IEEE International Symposium on Parallel Architectures, Algorithms and Programming(PAAP '12), pp.186-193, Taipei, Taiwan, 2012/12/18.
3. Shun Shiramatsu, Norifumi Hirata, Robin M. E. Swezey, Hiroyuki Sano, Tadachika Ozono, and Toramatsu Shintani: Gathering Public Concerns from Web towards Building Corpus of Japanese Regional Concerns. In Proceedings of the 2012 IIAI International Conference on Advanced Applied Informatics (IIAI-AAI 2012), pp. 248-253, Fukuoka, 2012/9/21.
4. Norifumi Hirata, Hiroyuki Sano, Rrobin M. E. Swezey, Shun Shiramatsu, Tadachika Ozono, and Toramatsu Shintani: A Web Agent Based on Exploratory Event Mining in Social Media. In Proceedings of the 2012 IIAI International Conference on Advanced Applied Informatics (IIAI-AAI 2012), pp. 236-241, Fukuoka, 2012/9/21.
5. Tadachika Ozono, Robin M. E. Swezey, Shun Shiramatsu, Toramatsu Shintani, Ryota Inoue, Yudai Kato, Takushi Goda: A Real-Time Collaborative Web Page Editing System WFE-S based on Cloud Computing Environment. In Proceedings of the 2012 IIAI International Conference on Advanced Applied Informatics (IIAI-AAI 2012), pp. 224-229, Fukuoka, 2012/9/21.
6. Tadachika Ozono, Robin M. E. Swezey, Shun Shiramatsu, Toramatsu Shintani, Ryota Inoue, Yudai Kato, Takushi Goda: Differential Synchronization Mechanism for a Real-Time Collaborative Web Page Editing System WFE-S. In Proceedings of the 2012 IIAI International Conference on Advanced Applied Informatics (IIAI-AAI 2012), pp.242-247, Fukuoka, 2012/9/21.
7. Shun Shiramatsu, Robin M. E. Swezey, Hiroyuki Sano, Norifumi Hirata, Tadachika Ozono and Toramatsu Shintani: Structuring Japanese Regional Information Gathered from the Web as Linked Open Data for Use in Concern Assessment. In Electronic Participation. Proceedings of the 4th IFIP WG 8.5 International Conference, ePart 2012, Lecture Note in Computer Science, Vol. 7444, Springer, pp. 73-84, Kristiansand, Norway, 2012/9. 2012/9/3-5
8. Robin M. E. Swezey, Hiroyuki Sano, Norifumi Hirata, Shun Shiramatsu, Tadachika Ozono, Toramatsu Shintani: An e-Participation Support System for Regional Communities Based on Linked Open Data, Classification, and Clustering. In Proceedings of the 11th IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI-CC 2012), pp. 211-218, Kyoto, 2012. 2012/8/22-24
9. Robin M. E. Swezey, Shun Shiramatsu, Tadachika Ozono, and Toramatsu Shintani: An Improvement for Naive

Bayes Text Classification Applied to Online Imbalanced Crowdsourced Corpora. The 25th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE 2012), In Modern Advances in Intelligent Systems and Tools, Studies in Computational Intelligence, Vol. 431, Springer, pp.147-152, Dalian, China, 2012. 2012/6/9-12
10. Tatiana Zidrasco, Victoria Bobicev, Shun Shiramatsu, Tadachika Ozono, Toramatsu Shintani: How to Reach it? Defining Language Features Leading to Agreement in Discourse. In Proceedings of the 8th International Conference Recent Advances in Natural Language, Processing (RANLP 2011), pp. 781-786, 2011. 12-14, September, Hissar, Bulgaria
11. Jun Takasaki, Tatiana Zidrasco, Shun Shiramatsu, Tadachika Ozono,Toramatsu Shintani: On Facilitating Argumentation by Designing Agreement-Oriented Rules. Proceedings of the 2010 International Congress on Computer Applications and Computational Science, pp.218-223, Singapore, 2010.
12. Shun Shiramatsu, Tadachika Ozono, Toramatsu Shintani, Hiroshi G. Okuno: A Corpus-based Analysis of Coreferential Recency Effect in Japanese Discourse for Tracking Dynamic Topic. In Proceedings of the 9th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2010), pp. 645-650, Yamagata, 2010.
13. Tadachika Ozono, Shun Shiramatsu, Toramatsu Shintani: Preventing Fake Web Pages Using Push Delivery - Defending against Theft Crawlers. In Proceedings of the 9th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2010), pp. 639-644, Yamagata, 2010.
14. Shun Shiramatsu, Jun Takasaki, Tatiana Zidrasco, Tadachika Ozono, Toramatsu Shintani, Hiroshi G. Okuno: System for Supporting Web-based Public Debate Using Transcripts of Face-to-Face Meeting. In Trends in Applied Intelligent Systems, Proceedings of the 23rd. International Conference on Industrial Engineering and Other Applications of Applied Intelligence Systems (IEA/AIE 2010), Part III, Lecture Notes in Computer Science , Vol. 6098, Springer, pp. 311-320, Cordoba, Spain, 2010.
15. Tatiana Zidrasco, Shun Shiramatsu, Jun Takasaki, Tadachika Ozono, Toramatsu Shintani: Building and Analyzing Corpus to Investigate Appropriateness of Argumentative Discourse Structure for Facilitating Consensus. In Trends in Applied Intelligent Systems, Proceedings of the 23rd. International Conference on Industrial Engineering and Other Applications of Applied Intelligence Systems (IEA/AIE 2010), Part II, Lecture Notes in Computer Science, Vol. 6097, Springer, pp. 575-584, Cordoba, Spain, 2010.
16. Norifumi Hirata, Shun Shiramatsu, Tadachika Ozono, Toramatsu Shintani: Generating an Event Arrangement for Understanding News Articles on the Web. In Trends in Applied Intelligent Systems, Proceedings of the 23rd. International Conference on Industrial Engineering and Other Applications of Applied Intelligence Systems (IEA/AIE 2010), Part II, Lecture Notes in Computer Science, Vol. 6097, Springer, pp. 525-534, Cordoba, Spain, 2010.

〔図書〕（計１件）

1. 新谷虎松，大囿忠親，白松俊：知識システムの実装基礎 - スライドで理解する人工知能技術．コロナ社，215p, 2012/10/22.

〔その他〕

ホームページ:

http://www-toralab.ics.nitech.ac.jp/index-j.html

６．研究組織

(1)研究代表者

新谷　虎松（Toramatsu SHINTANI）
名古屋工業大学・工学研究科・教授
研究者番号：00252312

(2)研究分担者

大囿　忠親（Tadachika OZONO）
名古屋工業大学・工学研究科・准教授
研究者番号：90324475

(3)連携研究者

なし