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　5 日現在

研究種目：基盤研究（C）
研究期間：2007～2009
課題番号：19500016
研究課題名（和文）　ウェブ構造マイニングのアルゴリズムとその効率化に関する研究

研究課題名（英文）　Studies on Algorithms for Web Structure Mining and their Efficiency

研究代表者
宇野　裕之（UNO YUSHI）
大阪府立大学・理学系研究科・准教授
研究者番号：60244670

研究成果の概要（和文）：ウェブ構造マイニングは，ウェブのリンク構造を表現するウェブグラフを対象に，コミュニティの発見などを目指す．本研究では，コミュニティを表す可能性がある頻出構造を特定し，隠れた情報の効率的な発見を試みた．その結果，そのような構造を一まとめにする縮約ウェブグラフを提案し，縮約を反復して行ったグラフから有用な情報を繰り返し発見することに成功した．また，縮約ウェブグラフにおけるさまざまなスケールフリー性を観察し，その現象を説明するネットワークモデルを提案した．

研究成果の概要（英文）：One of the objectives of web structure mining is to find hidden communities from the Web based on the webgraph which is a model for representing the link structure of the Web. In this research, we tried to identify frequent substructures on the Web that represents communities. We propose “contracted webgraphs”, where such substructures are shrunk into single nodes in the original webgraph. As a result, we verified that we can find new information in such contracted webgraphs by extracting frequent substructures repeatedly from them. We also observed several new scale-freeness in those webgraphs, and then proposed a new network model that can explain such phenomena.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	1,500,000	450,000	1,950,000
２００８年度	1,100,000	330,000	1,430,000
２００９年度	700,000	210,000	910,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：総合領域
科研費の分科・細目：情報学・情報学基礎
キーワード：離散構造，アルゴリズム，ウェブ（WWW），リンク解析，データマイニング

１．研究開始当初の背景

ウェブのリンク構造とウェブアルゴリズム

ウェブは 1990 年代初頭の誕生以来，人々の予想を超える急速な発展を遂げ，情報発信や検索の手段として不可欠なものとなり，われわれの日常生活や社会にさまざまな恩恵と影響をもたらすと同時に，数多くの新しい研究分野を創出しつづけている．ウェブに関する研究分野も拡大を続けて多岐に渡り，その中心は応用的な分野であるが，その一方で，ウェブの応用的な利用を支える基礎的，理論的な研究が存在する．その中心となるのが，ウェブのリンク構造を有向グラフとしてとらえるウェブグラフに関連する分野である．ウェブグラフは，ウェブ上で動作する検索エンジン，クローラやページランキングなどのウェブアルゴリズム設計のための，最も基本的なモデルとなっている．この分野の重要性を最初に示したのは Jon M. Kleinberg であり，彼がこのことをはじめとする情報科学の基礎理論における功績により，2006 年の国際数学者会議において，ネヴァリンナ賞を授与されたことが，この分野の重要性を示している．

このような分野の中で本研究では，ウェブ上で動作するウェブアルゴリズムや，その中でもとくに，次に説明するウェブ構造マイニングとそのアルゴリズム，ならびにそれらの効率化など，ウェブグラフの応用的な利用の側面に焦点をあてる．ここでとくに留意すべき点は，アルゴリズムのスケーラビリティである．ウェブアルゴリズムの多くは，基本的にはグラフ・ネットワークのアルゴリズムであるが，それらがウェブ規模のグラフで動作する実装が伴わなければ，実用上は無意味であるという点で，既存のアルゴリズムとは本質的に異なる．

２．研究の目的

ウェブ構造マイニング

ウェブを巨大なデータベースとみなして，主として単純な照合で情報を得るのが検索であるのに対して，それだけでは発見できない埋もれた二次的な情報を発見することをウェブマイニングとよぶ．その中でも，ウェブのリンク構造に注目して(すなわちウェブグラフ上で)行うマイニングを，とくにウェブ構造マイニングとよぶ．構造マイニングは，ウェブ上で特定の話題に興味を持つ隠れたコミュニティの発見などを目指し，その成果は社会現象の解明や，構造にもとづく検索エンジンの効率化などへの利用が期待できる．

一般に，コミュニティを構成するページは互いに緊密に，あるいは規則的にリンクを張ると想像され，ウェブグラフの中で密な部分グラフや，固有の特徴的な部分グラフを構成していると考えられる．従って，これらの仮定のもとでウェブ構造マイニングは，ウェブグラフからそのような部分グラフを抽出することで達成されることになる．

本研究課題は，Kleinberg がその重要性を指摘し，いまやリンク解析などの用語とともに，重要な位置を築いている研究分野のまさに延長線上にあり，そのような新たな課題に取り組み，解決することを目的とするものである．具体的には，本研究の目的は以下のとおりである．

- ウェブグラフの構造のさらなる解明と理論モデルの構築
- ウェブ構造マイニングのための実用アルゴリズムの開発と実装
- 開発アルゴリズムを利用したウェブ上での実際に有用な情報の発見

３．研究の方法

ウェブグラフは，ウェブページのリンク構造をモデル化したものである．2006 年現在でのウェブページの総数は，少なくとも 200 億ページはあると推定されており，その数は現在もなお爆発的に増加しているだけでなく，そのリンク自身も生成死滅を繰り返し，そのトポロジーは常に変化しつづけている．またその位相構造は，古典的なランダムグラフとは異なり，スケールフリー性やスモールワールド性などの，さまざまな特徴的な性質を持つことがこれまでの研究で明らかになってきている．

ウェブに対するアルゴリズムを研究・開発することは，基本的にはグラフ・アルゴリズムを設計することに違いないが，上記の (i) 規模が極めて大きいこと，および (ii) 成長性 を考慮するとき，小規模なネットワークに対しては効率的に動作する既存のアルゴリズムであっても，ウェブ規模のグラ

フではそれが望めないという状況がしばしば起きる．すなわち，ウェブアルゴリズムに対しては，常にそのスケーラビリティが求められている点において，従来アルゴリズムとは本質的に異なる問題を抱えている．このとき，この問題を克服する手がかりとして，(iii) 特徴的な位相構造 の性質を利用することが重要かつ不可欠であると考える．

これらを踏まえ，研究目的の１項目目に関して予備知識を得る．すなわち，社会学，生物学や交通ネットワークなど，現実社会で自然発生的に生成されるネットワークは一般に複雑ネットワークと総称され，ウェブグラフもその代表的なものの一つである．複雑ネットワークは，スケールフリー性やスモールワールド性など，従来の古典的なネットワークには見られない著しく異なった性質が観察され，その生成メカニズムに関しては，これまで主として物理学や数学の分野で活発な研究があり，多くのグラフモデルが提案されている．

ところがそれらには，ウェブグラフに見られる階層構造(より一般的には自己相似性)を十分に説明するモデルが欠けていることが判明しつつある．このことを厳密に確認・検証し，過去の研究で欠如が明らかになった「ウェブグラフの階層構造(ある種の自己相似性)を組み込むことが可能なグラフモデル」について考察・考案を試みる．これは，開発するウェブアルゴリズムの性能評価や動作検証の目的にも必要であり，そのようなモデルを考案し提案する．

二つ目の研究目的であるウェブ構造マイニングのための実用アルゴリズムの開発・実装，およびそれらのスケーラビリティを含めた性能向上，効率化に関しては，以下のような方法をとる．

本研究における構造マイニングの方針としては，実際にウェブグラフに頻出する特徴的な構造を発見，同定し，主としてそれらを列挙することによりマイニングを達成するというアプローチをとりたい．
－頻出構造の同定：はじめに，実際のウェブデータにもとづくウェブグラフを観察することで，頻出構造を同定することを目指す．
－列挙アルゴリズムの構築：つづいて同定された構造を数学的に厳密に定式化し，それらを列挙するアルゴリズムを設計，提案し実用化する．
－アルゴリズムの効率化，スケーラビリティの向上：一般に，グラフ中の頻出構造はそれ自身が指数個存在する可能性があり，アルゴリズムの実行時間が入力多項式ではなく，出力多項式時間にならざるを得ないケースが多い．このようなアルゴリズムは，ウェブ規模のデータに対してはほとんど無力である．そこで，スケーラビリティの向上を図る必要が生じるが，本研究では，対象となる頻出構造を列挙が容易な(出現個数がある程度制限されるような)構造として定式化することを有望視する．また，少なくとも適切なデータ構造を用いた高度に洗練された実装は常に必要である．

研究計画全体をとおして，
・大規模データを効率的に処理できるための計算機環境の導入，
・外部専門家知識の要請，
・大学院学生の助力，
・研究成果の社会への還元
をあわせて予定する．

４．研究成果

ウェブのリンク構造を有向グラフとしてモデル化したウェブグラフについて，その研究の創始より 10 余年あまりが経過したいま，われわれはそのモデルを再評価するとともに，ウェブグラフで動作するさまざまなアルゴリズムの設計・開発やそれらの効率化を目指す．なかでも，ウェブにアップロードされている情報を巨大なデータベースと見なし，そこから有益な情報を発見するデータマイニング技術は重要かつ不可欠である．今年度は，そのモデル上で行われるウェブ構造マイニングに関する研究を重点的に実施し，以下に示すような結果を得た．

（１）実際のウェブデータから構築されたウェブグラフを観察することで，孤立クリークや孤立スターと呼ばれる頻出部分構造が存在することを発見・同定した．その上で，これらの大きさの分布は，よく知られているスケールフリー性を満たすことが確認され，ウェブグラフの新たな理論モデルの手がかりとなった．

（２）大規模な実験を通して存在が確認された孤立クリークおよび孤立スターと呼ぶ頻出構造の中に，実際にコミュニティに対応する構造があること，しかしながらその大部分は単一ドメイン内部に存在するメニューやインデックス構造であることなどを確認した．

（３）特定の頻出部分構造のうち，有用な情報の発掘には直接的に無関係であると考えられるものを縮約することによって得られ

る「縮約ウェブグラフ」と呼ばれるモデルを提案した．さらにその操作を繰り返し反復することをあわせて提唱した．

（4）そのような縮約ウェブグラフにおいて，新たなさまざまなスケールフリー性を観察した．具体的には，縮約ウェブグラフにおける次数分布，孤立クリークおよび孤立スターの大きさの分布，およびそれらが縮約ウェブグラフにおいて反復して出現することなどである．

（5）4で観察された事実を説明可能な理論的なネットワークモデルを提案した．具体的には，clique replace model や hierarchical isolated clique model と呼ぶものである．その上で，これらのモデルが持つ性質に対する理論的な解析を行った．

（6）1億ノード 20 億リンク以上からなる実際のウェブデータに対してウェブグラフを構築し，そのウェブグラフに対して特定の頻出構造を用いて縮約ウェブグラフを繰り返し構築し，構造マイニングを反復する実験を行った．その結果，縮約前のグラフでは得られなかった新しい情報を発見することに成功し，本研究で提案する縮約ウェブグラフの有効性，妥当性を検証した．

今後の展望としては，ウェブデータ圧縮にウェブ構造マイニングの成果を持ち込むことで，きわめて効果的にウェブデータを圧縮することができる可能性があるのではないかと考える．具体的には，ウェブグラフ中の局所的に密な部分構造を独立して保持することで，データ量を大きく削減できることを予想する．また，グラフのスケールフリー性を利用することで，その圧縮効率を高くすることも期待できる．なお，これらの圧縮は当然可逆なものでなければならず，これらを採用することで，既存のウェブアルゴリズムの性能も向上させることが期待できる．

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　11 件）

① T. Shigezumi, Y. Uno and O. Watanabe. A new model for a scale-free hierarchical structure of isolated cliques. 査読有. Lecture Notes in Computer Science, Vol. 5942, pp. 216-227, Springer, 2010.

② Y. Uno, Y. Ota and A. Uemichi. Investigating the Web structure by isolates stars. 査読有. Transactions of the Japanese Society for Artificial Intelligence, Vol. 25, pp. 9-15, 2010.

③ T. Shigezumi, Y. Uno and O. Watanabe. A replacement model for a scale-free property of cliques. 査読有. Proc. 8th Cologne-Twente Workshop on Graphs and Combinatorial Optimization, pp. 285-289, 2009.

④ Y. Uno, Y. Ota and A. Uemichi. Web structure mining by isolates stars. 査読有. Lecture Notes in Computer Science, Vol. 4936, pp. 149-156, 2008.

⑤ Y. Uno, Y. Ota and A. Uemichi. Web structure mining by isolated cliques. 査読有. IEICE Transactions on Information and Systems, Vol. E90-D, pp. 911-930, 2007.

〔学会発表〕（計　14 件）

① T. Shigezumi, Y. Uno and O. Watanabe. A new model for a scale-free hierarchical structure of isolated cliques. The 4th Workshop on Algorithms and Computation, Bangladesh, 2010 年 2 月 11 日.

② 小栗 史弥, 清谷 竜也, 宇野 裕之. 孤立クリークおよび孤立スター縮約ウェブグラフにおけるウェブ構造マイニング. 電子情報通信学会 Technical Report, Vol. 109, No. 391, COMP2009-44, 福岡, 2010 年 1 月 25 日.

③ 宇野裕之. ウェブ・アルゴリズム―ウェブグラフの性質とその利用―. 日本オペレーションズ・リサーチ学会北海道支部平成 21 年度第 2 回講演会, 札幌, 2009 年 12 月 12 日.

④ T. Shigezumi, Y. Uno and O. Watanabe. A replacement model for a scale-free property of cliques. The 8th Cologne-Twente Workshop on Graphs and Combinatorial Optimization, Paris, 2009 年 6 月 3 日.

⑤ Y. Uno. Investigating web structure by cliques and stars. Kyoto RIMS Workshop on Acceleration and Visualization of Computation for Enumeration Problems, Kyoto, 2008 年 9 月 30 日.

６．研究組織

(1)研究代表者

宇野　裕之（UNO　YUSHI）
大阪府立大学・理学系研究科・准教授
研究者番号：60244670

(2)研究分担者

なし

(3)連携研究者

なし