様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ５ 月 １５ 日現在

機関番号：２４４０３
研究種目：基盤研究(C)
研究期間：2011～2014
課題番号：２３５０００２２

研究課題名（和文）ウェブリンク構造の時系列データからのマイニング―表現モデルとアルゴリズム―

研究課題名（英文）Web structure mining from time series web data---models and algorithms---

研究代表者
宇野　裕之（Uno, Yushi）

大阪府立大学・理学（系）研究科（研究院）・准教授

研究者番号：６０２４４６７０

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：ウェブのリンク構造は通常ウェブグラフとして表現される．本研究は，リンク構造が動的に変化し成長を続ける様子に着目し時系列データとしてとらえ，リンク構造の単にある瞬間のスナップショットを表すウェブグラフに代わり，ウェブ時系列データを表現するモデルを構築すること，およびそのモデルをもとにウェブに潜在する知識を時系列データから発見するためのデータ・マイニングアルゴリズムの設計や開発を行った．またこのようなモデルや手法の他分野の問題への適用可能性を検証した．

研究成果の概要（英文）：The link structure of the Web is generally viewed as a webgraph. In this research, by focusing on that the Web is changing and evolving dymanically over time, we successfully constructed models that represents a time series data of the link structure of the Web instead of webgraphs, and also developed algorithms, based on the constructed models, that can find potential structure representing hidden knowledge in the time series link data. We also verified the possibility of applying our methods to the problems in other fields.

研究分野：離散構造とアルゴリズム

キーワード：アルゴリズム　ウェブグラフ　時系列　データマイニング　情報基礎　離散最適化

１．研究開始当初の背景

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，日常生活や社会に多大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リンク解析と呼ばれる分野で，ウェブのリンク構造を表すウェブグラフは，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，さまざまな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指摘したのは Jon M. Kleinberg [Authoritative sources in a hyperlinked environment. J. ACM, 1997] であり，彼がこれらを含む情報科学の基礎理論における多大な功績により，2006 年の国際数学者会議(ICM) において，Nevanlinna 賞を授与されたことが，この分野の重要性や将来性を証明している．また，チャンピオン検索エンジンである Google が，理論計算機科学分野の成果物であることも，いまやよく知られた事実である．

このように，ウェブグラフはウェブアルゴリズム設計のためのモデルとして最大限に利用されているが，それはリンク構造の静的スナップショットにすぎず，動的に変化し成長するリンク構造を十分に表現できていない．そこで本研究の最大の特徴は，ウェブのリンク構造に時系列という視点を導入することである．その上で，従来は主として静的なリンク構造上で行われていた構造マイニングやさまざまなウェブアルゴリズムをウェブ時系列データ上で動作させ，既存の手法では得られなかった未知の知識の発見を目指す．

もちろん，ウェブを時系列データと見なす研究は過去にも存在する([Kraft et al. TimeLinks: Exploring the link structure of the evolving Web. *Proc. 2nd WAW*, 2003], [Toyoda and Kitsuregawa. What's really new on the Web? *Proc. 15th WWW Conf.*, 2006] など). しかしながらこれらの研究は散発的で，しかもウェブページの出現キーワードの意味解析や，単にリンクの変化の統計情報を与えるなど，時系列データにリンク解析の立場からの研究は確立していないのが実情である．したがって本研究は，Kleinberg がその重要性を指摘したウェブの基礎理論の主要トピックであるリンク解析の分野の延長線上に，時系列という新たな視点とともに新しい領域を開拓・展開しようとするものである．

２．研究の目的

われわれは過去の研究で，ウェブグラフがもつ新たなスケールフリー性(自己相似性)の観察やその性質をもつネットワーク生成モデルの提案と理論的な解析を行ってきた．またウェブ構造マイニングに関しては，ウェブグラフに表れる頻出構造を同定し縮約ウェブグラフを提案した上で，そのモデル上での構造マイニングで新たな知識を発見できることを確認するなど，多くの成果を得てきた．一方で，これらの研究の過程で重要かつ示唆に富む新たな研究課題の存在も認識させられた．その最大のものが，「ウェブのリンク構造を時系列データと見なす」というものである．

研究背景を踏まえ，本研究で明らかにしたいことは，以下の事項である．

（１）ウェブ時系列データの表現モデルの考案・構築

① 時系列データの表現モデルそのもの：ウェブ時系列データは，単純には 1 つ以上の静的ウェブグラフの族と考えることができる．しかしながら，1 つでも数百億以上のノードからなるグラフを複数保持する表現は，物理的なメモリの限界とともに，ウェブアルゴリズムの効率的な動作が期待できない．動的に変化する差分だけを保持する方法などが考えられるが，アルゴリズムを想定した最適な表現方法を考案し確立する．過去に関連する研究が少なく，新規開拓を目指すトピックである．

② 自己相似性を持つネットワークモデルの考案：現実に自然発生する（複雑）ネットワークには，スケールフリー性など従来のネットワークには見られない著しい性質が観察されるが，ウェブグラフに見られる再帰的な階層構造（より一般的には自己相似性）の生成メカニズムを十分に説明するモデルが欠如している([Bharat et al. Who links to whom: mining linkage between web sites. *Proc. IEEE ICDM*, 2001]). この事実を説明可能なモデルを考案することが，時系列データの表現方法につながると考えている．

③ ウェブグラフの圧縮技法の開発と利用：ウェブデータは極めて巨大であり，時系列にしたがい保持するには相当の困難が予想される．そこでウェブグラフの可逆な状態での圧縮を考える．これには，テキストなど 1 次元データとは異なる圧縮技術が必要であるが，近年大きな進展は伺えない．ここでは，ウェブグラフの性質も利用した効率的な圧縮技法を考案し，時系列データの簡潔な表現方法の開発に役立てたい．

④ 単一ウェブグラフに対するウェブアルゴリズムの時系列データへの拡張：既存のウェブアルゴリズムを時系列データに対応させるために，動的なものへの拡張を目指す．この方向の研究としては，たとえば[Gorke et al. Dynamic graph clustering using

minimum-cut trees. *Proc. WADS*, 2009]がある．

（2）ウェブ時系列データからのデータマイニング—実用アルゴリズムの開発と実装

項目 1 で構築する表現モデルにもとづき，リンク構造の時系列データから，単なるスナップショットからだけでは発見できない未知の知識を得たい．すなわち，時系列データからの構造マイニングである．ここでは，構造マイニングの方法の一つとして，頻出構造を特定しそれらを列挙するというアプローチをとる．(1) 頻出構造の同定：連続するスナップショットにおける頻出構造とは何か，その根本的な再定義が必要となる，(2) 列挙アルゴリズムの構築：同定された構造を数学的に厳密に定式化した上で，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，ことなどが目的となる．

リンク解析やその関連分野は，誕生から 20 余年が経過し一定の成熟期を迎えている．しかしながら，検索エンジンやランキング，データマイニングなどの既存の多くのウェブアルゴリズムは，静的なリンク構造を対象にし，リンク構造の動的な変化を考慮したものはほとんどなかった．そこに時系列という視点を導入するのが本研究の最大の特色である．その結果，時系列データからは，静的リンク構造では発見が困難な，例えばウェブ上での流行の変化のような新しい知識の発見が期待でき，これはリンク予測などを通じて社会現象の解明などに役立つことが期待される．さらに重要な点は，単にウェブの分野だけにとどまらず，時系列データは生物学，化学，物理学などあらゆる分野に存在し，その表現方法や効率的な処理が求められている．本研究の成果は，そのような多分野の問題を解決する糸口を与える可能性があるという点において，極めて意義が深いと考える．

３．研究の方法

本課題の研究期間 4 年のうち，前半 2 年を当初計画遂行フェーズ，後半 2 年を必要があれば目標を再設定した上での計画完遂フェーズと位置づける．1 年目：時系列データのモデル化，表現方法を考案し構築する．一方で，計算機環境を整備し，予備実験を開始する．2 年目：提案モデル上で，実際のウェブ時系列データからの構造マイニングを行う．この際，過去にわれわれが開発したアルゴリズムを時系列データに対応させるとともに，専用の手法を新たに開発する．3 年目：前半の総括，目標の再設定を行い，研究成果は順次公表する．その上で，2 年目の実験を継続するとともに，前半で開発したマイニングアルゴリズムを高速化，効率化する．そのための，ウェブグラフデータ圧縮技術を考察する．4 年目：圧縮表現されたウェブ時系列データからの知識発見を主題とするとともに，過去 3 年の各トピックでの総仕上げを行う．

初年度(23 年度) は，時系列データをいかに表現するかというモデル化そのものに取り組む．時系列データは，単純にはスナップショットであるウェブグラフ 1 つ以上からなるグラフ族と考えられるが，1 つでも数百億以上のノードからなるグラフを複数保持する表現は，物理的なメモリの限界とともに，さまざまなウェブアルゴリズムの効率的な動作が期待できない．ノードやリンクの動的に変化する差分だけを保持する方法などが考えられるが，個別のアルゴリズムを想定した最適な表現方法を複数考案しなければならない．この際，過去に欠如が指摘される「自己相似性(ある種の階層構造) の組込みが可能なウェブグラフの生成モデル」(たとえば [Bharat et al. Who links to whom: mining linkage between web sites. *IEEE ICDM*, 2001]) についても，過去の研究を継続し取り組む．これは，時系列データの新たな表現方法となる可能性があるだけでなく，今後検討を予定する各種ウェブアルゴリズムの性能評価や動作検証の目的にも役立つと考えるからである．

また 2 年目以降に中心的に実施予定の時系列データからのマイニング実験のための計算機を予め導入し予備実験を開始する．

計画 2 年目の 24 年度は，研究目的の中でウェブ時系列データからの知識発見を目的としたデータマイニングの実用アルゴリズムの開発と，実際のウェブデータからのマイニングに取り組む．本研究での構造マイニングは，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，それらの列挙により達成するアプローチをとりたい．その際，(1) 頻出構造の同定については，連続するスナップショットでの頻出構造をモデル化し，その厳密な定義を数学的に与える必要がある．また(2) 列挙アルゴリズムについては，時系列における頻出構造がグラフの部分構造という既存の概念とは異なる可能性があり，それに対応する列挙の新たな枠組みを構築した上で，個別のアルゴリズムを設計する必要がある．離散構造の列挙理論に数多くある研究成果 ([Avis et al. Reverse search for enumeration. *Discr. Appl. Math.* 65, 1996] ほか) は，設計に際し可能な範囲で用いるが，時系列データ上の頻出構造の新たな定義に応じてその可能性や限界を検証した上で，可能性を欠くものには新たなアイデアによる再設計が必要となる．このため関連する項目 1.4 を同時に目標に設定する．すなわち，ウェブのスナップショットを想定した既存アルゴリズムを，時系列データにも適用可能となる改良や再設計することである．これらは動的アルゴリズムの分野と関連があり，今後の重要な潮流を形成すると考える．

計画 3 年目の 25 年度は，ウェブグラフの圧縮技法の開発と利用にも着手する．具体的には，単一のウェブグラフを圧縮表現する方法を開発することで，複数のウェブグラフ情

報を持つ時系列データのコンパクトな表現が可能になると考え，過去 2 年に開発したマイニングを初めとする各種ウェブアルゴリズムの高速化，効率化を目指すものである．テキスト情報など 1 次元データの圧縮技法に関する研究は，歴史も古く数も多い．これに対して(ウェブ) グラフの圧縮は，たとえば [Suel et al. Compressing the graph structure of the Web. *Proc. IEEE DCC*, 2001] や [Brisaboa et al. *k*2-trees for compact Web graph representation. *Proc. SPIRE*, 2009] などがあるが，実際のウェブデータに対する圧縮性能の評価は定まらず，ウェブアルゴリズムに採用され効果をあげている例も少ない．また，グラフは通常の 1 次元データとは異なり，それを表現するデータ構造にも極めて高い自由度があり，このため圧縮にもさまざまなアプローチが考えられる．本研究課題では，構造マイニングの成果を持ち込むことでウェブデータ圧縮に新しいアイデアを導入したいと考えている．

４．研究成果

ウェブのリンク構造は通常ウェブグラフとして表現される．ウェブ構造マイニングの目的の一つは，ウェブに潜在するコミュニティなどの知識をその構造にもとづき発見することである．われわれは 23 年度までの研究で，それらに固有の構造を見つけ出し，そのような構造を縮約する縮約ウェブグラフの概念や，縮約ウェブグラフから構造マイニングを行う手法などを提案した．また，縮約ウェブグラフが極めて特徴的な自己相似性を持つことも明らかにした．

平成 23 年度は，時系列データをいかに表現するかというモデル化そのものに取り組んだ．時系列データは，単純にはスナップショットであるウェブグラフ 1 つ以上からなるグラフ族と考えられるが，1 つでも数百億以上のノードからなるグラフを複数保持する表現は，物理的なメモリの限界とともに，さまざまなウェブアルゴリズムの効率的な動作が期待できない．ノードやリンクの動的に変化する差分だけを保持する方法などが考えられるが，個別のアルゴリズムを想定した最適な表現方法を考案しなければならない．

研究ではウェブグラフの時系列データを，グラフとグラフの各枝が存在する時区間のペアとして定義することに成功した．さらに，ウェブグラフで重要な意味を持つとされるクリークのうち，長い時間存在するクリークを効率的に列挙する方法も考察した．これらの結果は，単にウェブのスナップショットだけでは得られない新たな知識を得ることにつながると考える．

平成 24～25 年度は，前年度までにひきつづき時系列データをいかに表現するかというモデル化そのものに取り組んだ．すなわち，ウェブグラフの時系列データを，グラフとグラフの各枝が存在する時区間のペアとして定義する方法をさらに拡張することを考え，長い時間存在するクリークをより効率的に列挙する方法を考察した．また，それ以外の重要な構造の時系列データ上での定義も新たに手がけた．

一方 25 年度は，ウェブリンク以外にも本質的に時系列と同様の性質を持つデータを取り上げ，それらのモデル化に着手した．その一例として，生物の遺伝子構造の類似性による繋がりが，その閾値により離散連続的に変化する様子に着目し，そこに潜む重要構造のモデル化を行うと同時に，実データによるマイニングの実験を開始した．

最終年度である 26 年度は，これまでに提案した時系列モデルを発展させ，より一般的なモデルへと拡張させることに成功した．具体的には，従来よく研究されてきたグラフ系列における頻出構造に代わり，われわれはグラフ系列中で一定期間以上の長期に渡ってある性質を保ち続ける部分構造（保存構造(preserving structure)と呼ぶ）に注目した．研究では，実用上も重要な連結成分とクリークに着目し，グラフ系列中に長期間存在するそのような構造を効率的に列挙するアルゴリズムを提案した．それらは多項式遅延時間アルゴリズムであり，グラフの各枝が高々1 つの時区間を持つという仮定のもとで計算時間は，連結成分の列挙が 1 つあたり O(|V||E|^3) 時間，クリークの列挙が 1 つあたり O(min{Δ^5, |E|^2 Δ}) というものである．ただし入力のグラフを G = (V,E)とし，G の最大次数をΔとしている．

また，ウェブグラフに関連して派生するさまざまなグラフ問題に着目し，それらを効率的に解くアルゴリズム群を開発した．たとえばその一つはベクトル支配点集合問題と呼ばれるものであり，それは入力として頂点数 n のグラフ G = (V, E) と需要ベクトルと呼ばれる n-次元非負ベクトル d = (d(1), d(2), . . . , d(n))が与えられたとき，S の中に少なくとも d(v)個の隣接頂点を持つような最小サイズの頂点部分集合 S を見つけるという問題である．この問題に対して，入力グラフが限定枝幅（したがって限定木幅）を持つときには多項式時間アルゴリズムが存在することや，そのことを利用してグラフが平面的であるときにはこの問題が解の大きさ k に関して劣指数時間固定パラメータ容易な問題であるという結論を得た．

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 14 件）

① Takeaki Uno and Yushi Uno. Mining preserving structures in a graph sequence.

Lecture Notes in Computer Science［査読あり］, Vol. 未定 (D. Du, D. Du and X. Du (eds.)), 2015.
DOI: 未定

② Toshimasa Ishii, Hirotaka Ono and Yushi Uno. (Total) Vector domination for graphs with bounded branchwidth. Lecture Notes in Computer Science［査読あり］, Vol. 8392 (A. Pardo and A. Viola (eds.)), pp. 238--249, 2014.
DOI: 10.1007/978-3-642-54423-1_21

③ Yushi Uno and Fumiya Oguri. Contracted webgraphs---scale-freeness and structure mining---. IEICE Transacions on Information and Network Science［査読あり］, Vol. E96-B, 2766--2773, 2013.
DOI: なし

④ Takeya Shigezumi, Yushi Uno and Osamu Watanabe. A new model for a scale-free hierarchical structure of isolated cliques. Journal of Graph Algorithms and Applications［査読あり］, Vol. 15, 661--682, 2011.
DOI: 10.7155/jgaa.00243

〔学会発表〕（計 7 件）

① Takeaki Uno and Yushi Uno. Mining graph structures preserved long period. The 17th International Conference on Discovery Science, 2014年10月08日, リュブリアナ（スロベニア）

② Toshimasa Ishii, Hirotaka Ono and Yushi Uno. (Total) Vector domination for graphs with bounded branchwidth. The 10th Latin American Theoretical Information Symposium, 2014年4月3日, モンテビデオ（ウルグアイ）

③ Yushi Uno and Fumiya Oguri. Contracted webgraphs: structure mining and scale-freeness. 5th International Frontiers of Algorithmics Workshop + 7th International Conference on Algorithmic Aspects of Information and Management, 2011年5月28日, Jinhua（中国）

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

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