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　 ６ 月　 ２ 日現在

機関番号：２５４０３
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５００１７８

研究課題名（和文）省メモリＷｅｂマイニング手法の開発とクラウドコンピューティングへの応用

研究課題名（英文）Development of memory-saving and time-efficient Web mining strategies and its applications on cloud computing

研究代表者
内田　智之（Tomoyuki, Uchida）

広島市立大学・情報科学研究科・准教授

研究者番号：７０２６４９３４

交付決定額（研究期間全体）：（直接経費）　　4,000,000円

研究成果の概要（和文）：閲覧したWebコンテンツに共通する木構造的特徴を表す項木パターンとその簡潔データ表現を定式化し、画像等から抽出できるグラフ構造的特徴を順序グラフパターンとして定式化した。簡潔データ構造を用いて、閲覧したWebコンテンツに頻出する項木パターンを過不足なく枚挙する省メモリかつ高速なWebマイニングアルゴリズムを開発した。さらに、計算量的機械学習理論に基づき、順序グラフパターンに対する効率的なパターンマッチングアルゴリズムを与え、正データからの多項式時間帰納推論可能性を示した。これらの結果に基づき、クラウド・コンピューティングに基づくユーザ・オリエンティッドな情報検索・提示システムを設計した。

研究成果の概要（英文）：The purpose of this research is to present memory-saving and time efficient Web mining strategies for extracting graph structured features common to webpages and apply proposed algorithms to information retrieval systems based on collaborative cloud computing. Firstly, we defined a succinct data representation for a term tree pattern representing tree structured features common to webpages. Secondly, we defined ordered graph patterns expressing graph structured features extracted from images in Webpages. Thirdly, we proposed efficient pattern matching algorithms for term tree patterns and incremental polynomial time enumeration algorithms using succinct data structures. Fourthly, based on computational machine learning, we showed polynomial time inductive inferability of ordered graph patterns from positive data by giving polynomial time pattern matching algorithms. Finally, we applied our proposed algorithms for information retrieval systems based on collaborative cloud computing.

研究分野：計算機科学

キーワード：グラフアルゴリズム　データマイニング　知識発見　Webマイニング　クラウド・コンピューティング　機械学習　情報基礎　情報検索

１．研究開始当初の背景

高速インターネット網の整備、コンピュータの高速化や補助記憶装置の大容量化に伴い、画像を多用した Web ページがストレスなく情報携帯端末で表示できるようになってきていた。それに伴い、キーワード以外に画像や映像といったマルチメディアデータも検索対象としたいというニーズが高まっていた。情報爆発時代でかつユビキタスネットワーク社会でもある昨今、小型ノートパソコンや携帯情報端末を介して、マルチメディアデータを含む Web ページから欲する情報を迅速かつ正確に得る手法、およびクラウド・コンピューティング上での情報検索・提示する方法が求められていた。

データをできる限りメモリ上に置くことで高速処理を可能とする簡潔データ構造および超簡潔データ構造の理論研究が急速に進み、データ圧縮、文字列検索、ゲノム解析等への応用研究も行われ始めていた。また、機械学習による画像・映像解析手法の高度化が図られ、自動でメタデータ(グラフでモデル化できるものを含む) を抽出する手法が実用に向けて大きく前進していた。グラフ構造データから頻出する部分構造を高速に発見する、機械学習に基づいたグラフマイニング手法やデータストリームからのグラフマイニング手法の研究は応募者らを含め盛んに行われていた。Web ページのグラフ構造は順序木で、Web ページ内の画像やリンクなどのレイアウト情報は平面グラフや順序グラフでモデル化できる。そこで、Web ページのもつ構造や内包するテキスト・画像・映像を解析することにより得られる特徴を一体的に捉えることができる Web マイニング手法の開発が求められていた。本研究のような、使用できるメモリ容量や CPU パワーなどに制限のある携帯情報端末を念頭に置いて、マルチメディアデータを含む Web ページに対する省メモリ Web マイニング手法を開発し、さらにユーザ・オリエンティッドな情報検索・提示システムに活用したシステムはまだ少なかった。

２．研究の目的

多くの Web ページはコマーシャル情報を含んでおり、表示画面が小さい携帯情報端末などでは閲覧中のページ内で欲しい情報を見つけ出すのに時間がかかってしまうことが少なくない。ユーザの閲覧履歴等から欲しい情報やノイズとなる情報を切り分けることができれば、画面サイズが制限された携帯情報端末でも欲しい情報を得やすくできる。さらに、ユーザが操作している携帯情報端末上であたかも実行しているかのように、Web 上で公開されている各種サービスを利活用できるようにしたクラウド・コンピューティングが増えてきている。

そこで、ユーザのネットサーフィンにおける閲覧履歴等から欲しい情報やノイズとなる情報を切り分けるために、Web ページやマルチメディアデータの持つグラフ構造をはじめとする特徴を一体的に抽出する省メモリ Web マイニング手法の開発を行い、クラウド・コンピューティングに基づいたユーザ・オリエンティッドな情報検索・提示システムへ応用することが本研究課題の目的である。

３．研究の方法

研究目的を達成するために、以下の方法で研究を遂行した。

1) ユーザが閲覧した Web コンテンツの木構造的特徴を表現する木パターンとその簡潔データ表現について定式化する。
2) ユーザが閲覧したマルチメディア Web コンテンツの木構造的特徴を表すのに適した項木パターンや、Web コンテンツ内の画像等のマルチメディアの構造的特徴を表すのに適した順序グラフパターンに関する計算量的機械学習理論の構築を行う。
3) 2 で得られる計算量理論的機械学習理論に基づき、閲覧した Web ページに共通する木構造的特徴を表す極大項木パターンを枚挙する方法と画像等がもつグラフ構造的特徴を表す極大順序グラフパターンの発見手法の開発を行う。
4) 特徴的なグラフ構造パターンを管理する方法の開発を行う。さらにその省メモリ化を図る。
5) ４までに得られた研究成果を、クラウド・コンピューティングに基づくユーザ・オリエンティッドな情報検索・提示システムへ応用する。

４．研究成果

以下に本研究で得られた成果を述べる。

1) ユーザが閲覧した Web コンテンツの特徴を表現する木パターンとして、過去に提案している（順序）項木パターンと VLDC 木パターンについて検討し、それらの簡潔データ表現について定式化を行った。項木パターンとは、2 つのノードからなる変数のみを有する木構造パターンである。項木パターンの変数を多ノードからなる変数に拡張したものを多ポート項木パターンという。本研究では省メモリでかつ高速な手法の開発を目的としているため、まず項木パターンに対する簡潔データ表現の定式化を行った。その定式化に基づいて、

多ポート項木パターンの簡潔データ表現の定式化を行った（雑誌論文 3、学会発表 5 など）。図 1 に多ポート項木パターンの例を、図 2 にその簡潔データ表現を与える。なお、図 1 では、変数を□で、各ノードは○で表している。また、○の中の記号はノード ID を、□の横にある記号は変数 ID を表している。□の中にある記号は変数ラベルを、辺のそばにある記号は辺ラベルを表す。図 2 において、簡潔データ表現における各ノードと変数に対する位置を下線およびノード ID と変数 ID で示している。変数を構成するノード v_1, v_4 および v_5 は簡潔データ表現上では特殊記号 $で表されている。さらに、VLDC 木パターンとは、Variable Length Don't Care と呼ばれる構造変数を有する木パターンであり、その簡潔データ表現は順序木に対する簡潔データ表現をそのまま適用することができるため、VLDC 木パターンについても検討した（雑誌論文 5 など）。

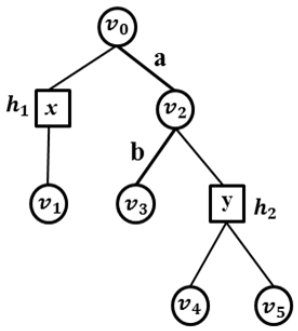

図 1　多ポート項木パターン t

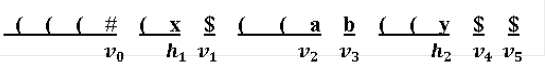

図 2　t の簡潔データ表現

2) 計算量的機械学習理論の観点から、Web コンテンツのもつ木構造的特徴を表す項木パターン・多ポート項木パターン（雑誌論文 8 など）や画像のグラフ構造的特徴を表す順序グラフパターン（雑誌論文 1 など）に対する多項式時間パターンマッチングアルゴリズムを提案し、正データからの多項式時間帰納推論可能性についての成果を得た（学会発表 1 など）。さらには、これらの研究成果をもとに、擬似的に木構造を有する Tree Contraction パターン（雑誌論文 2,4 など）や木幅が制限されたグラフに対する Graph Contraction パターン（雑誌論文 6 など）に応用した（研究発表 4）。
3) 与えられた順序木の集合を S とする。項木パターン t が次の条件を満たすとき、t は極大であるといい、t により生成される言語 $L(t)$は極小であるという。条件:S を包含し、言語 $L(t)$に真に含まれる言語を生成する項木パターンが存在しない。2 で得られた研究成果に基づき、極大項木パターンを過不足なく枚挙する逐次多項式時間アルゴリズムを提案し、実装を行い、その効率性を示した（研究発表 3 など）。さらに、この結果を多ポート項木パターンの枚挙アルゴリズムに拡張した（研究発表 5 など）。多ポート項木パターンの枚挙過程を示した例を図 3 に与える。

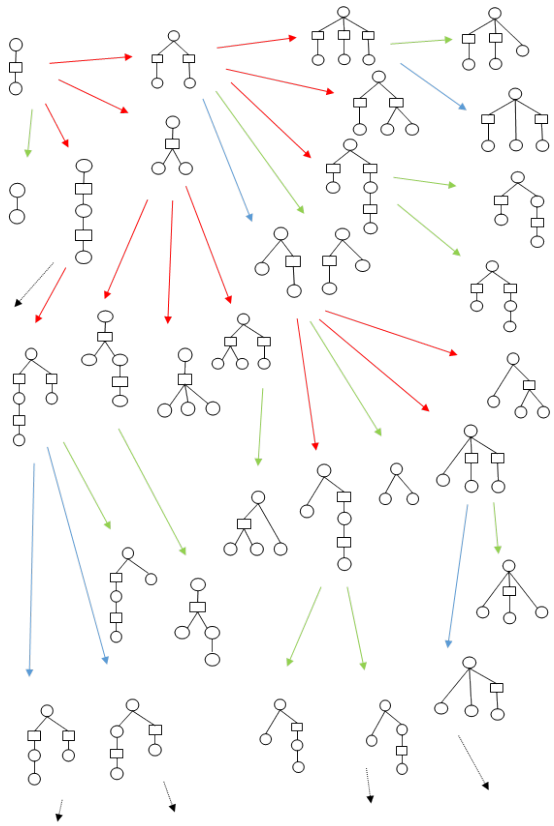

図 3　多ポート項木パターンの枚挙過程

4) マイニング結果として発見された特徴的な項木パターンは図 3 で示された枚挙過程を示した木で管理される。この木の各ノードには親の簡潔データ表現の位置 i にある最右パス上のノードあるいは変数に対して操作 A を行うことで得られるということを示すペア(i,A)だけが保持されることで省メモリ化を図っている。なお、操作 A には、(1) 変数に辺を代入する操作と (2) 変数を追加する操作の 2 種類がある。
5) これまでの研究成果に基づき、ユーザのネットサーフィンの閲覧情報からユーザの嗜好パターンを学習するクラウド・コンピューティングに基づくユーザ・オリエンティッドな情報検索・提

示システムの開発を行った。過去に研究実績がある単語間木構造パターンを木パターンとした情報検索・提示システムをまず開発した。その全体概念図を図４に与える。単語間木構造パターンとは、葉には単語あるいは単語のリストが、内部ノードには点ラベルがラベルづけられている順序木のことである。このシステムは、ラウド・コンピューティングに基づき、携帯型情報端末iPad等を念頭にiOS上で稼働するアプリとして実装した。サーバ側では、クライアントからのユーザの閲覧情報等やユーザの嗜好パターンを表す単語間木構造パターンを収集・管理するために MySQL を用い、クライアントとサーバ間のキーワードや閲覧情報の通信処理には PHP を用いた。

単語間木構造パターンの出現位置を獲得するアルゴリズムはすでに開発済みであった。しかし、項木パターンが順序木にマッチするか否かの判定には出現位置の特定は不要であるため、項木パターンの出現位置獲得アルゴリズムを開発する必要がある。そのため、与えられた項木パターン t と t にマッチする順序木 T が与えられたとき、t の各葉が対応する T の葉を確定していくことで、すべてのマッチング関数を枚挙するアルゴリズムを開発し実装を行った。しかし、研究機関内では、上記の情報検索・提示システムのサーバ側にそのプログラムを移植し、4で開発した管理方法を MySQL で実現するまでには至らなかった。今後そのシステムのプロトタイプの実装を速やかに行う予定である。また、多ポート項木パターンと VLDC 木パターンの出現位置獲得手法についても検討する予定である。

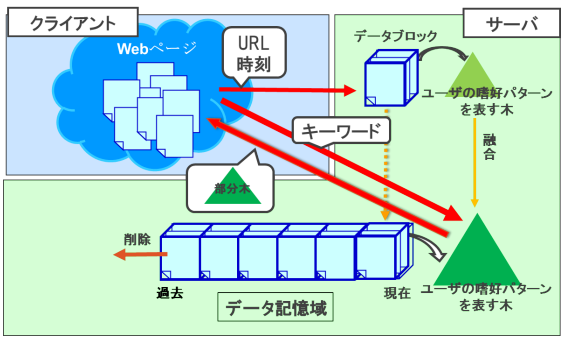

図４　システムの全体概念図

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８件）

① T. Hino, Y. Suzuki, T. Uchida and Y. Itokawa, Polynomial Time Pattern Matching Algorithm for Ordered Graph Patterns, Proc. ILP 2012, LNAI 7842, Springer Berlin Heidelberg, 査読有, 2013, pp.86-101.DOI: 10.1007/978-3-642-38812-5_7

② Y. Yoshimura and T. Shoudai, Learning Unordered Tree Contraction Patterns in Polynomial Time, Proc. ILP 2012, LNAI 7842, Springer Berlin Heidelberg, 査読有, 2013, pp.257-272. DOI: 10.1007/978-3-642-38812-5_18

③ Y. Itokawa, M. Wada, T. Ishii, T. Uchida, Pattern Matching Algorithm Using a Succinct Data Structure for Tree-Structured Patterns, Intelligent Control and Innovative Computing, LNEE110, Springer US, 査読有, 2012, pp.349-361, DOI: 10.1007/978-1-4614-1695-1_27.

④ Y. Okamoto and T. Shoudai, Hard Optimization Problems in Learning Tree Contraction Patterns, Applied Computing and Information Technology,553, Springer, Studies in Computational Intelligence, 査読有, 2014, pp.77-90, DOI: 10.1007/978-3-319-05717-0_6

⑤ S. Nakai, T. Miyahara, T. Kuboyama, T. Uchida and Y. Suzuki, Acquisition of Characteristic Tree Patterns with VLDC's by Genetic Programming and Edit Distance, 2013 IIAI International Conference on Advanced Applied Informatics (IIAI AAI 2013), Conference Publishing Services (CPS), 査読有, 2013, pp.147-151, DOI: DOI: 10.1109/IIAI-AAI.2013.79.

⑥ T. Yamada and T. Shoudai, Graph Contraction Pattern Matching for Graphs of Bounded Treewidth, Latest advances in inductive logic programming, London : Imperial College Press, 査読有, 2014, pp.173-180, DOI: 10.1142/9781783265091_0018.

⑦ S. Nakai, T. Miyahara, Y. Suzuki, T. Kuboyama and T. Uchida, Acquisition of Characteristic Sets of Tree Patterns with VLDC's Using Genetic Programming and Edit Distance, Prof. 7th Inter. Work. Computational

Intelligence and Applications (IWCIA 2014), IEEE, 査読有, 2014, pp.113-118, DOI: 10.1109/IWCIA.2014.6988088.

⑧ Y. Suzuki, T. Shoudai, T. Uchida and T. Miyahara, An Efficient Pattern Matching Algorithm for Ordered Term Tree Patterns, IEICE Trans. Fundamentals, E98-A, No.6, IEICE, 査読有, 2015 (to appear).

〔学会発表〕(計5件)

① T. Hino, Y. Suzuki, T. Uchida, T. Miyahara, Ordered Graph Patterns Which Are Polynomial Time Inductively Inferable from Positive Data, Proc. 7th IADIS Information Systems Conference (IS 2014), IADIS, 査読有, 2014, pp.263-270.28 Feb.-02 March, Madrid, Spain.

② Y. Okamoto, K. Koyanagi, T. Shoudai and O. Maruyama, Discovery of Tree Structured Patterns Using Markov Chain Monte Carlo Method, Proc. 7th IADIS Information Systems Conference (IS 2014), IADIS, 査読有, 2014, pp.263-270.28 Feb.-02 March, Madrid, Spain.

③ Y. Itokawa, T. Uchida and M. Sano, An Algorithm for Enumerating All Maximal Tree Patterns without Duplication Using Succinct Data Structure, Proc. International MultiConference of Engineers and Computer Scientists 2014 Vol I (IMECS 2014), International Association of Engineers(IAENG), 査読有, 2014, pp.156-161, 18-20 March, Hong Kong.

④ 正代隆義, 内田智之, 多項式時間学習可能な木幅定数グラフ言語の形式体系について, 冬のLAシンポジウム, 2015年01月28日～30日, 京都大学 数理解決研究所.

⑤ 糸川裕子, 内田智之, 構造データからの頻出多ポート項木パターン枚挙アルゴリズム, 2015年度人工知能学会全国大会, 2015年05月30日～06月02日, 公立はこだて未来大学.

６．研究組織

(1)研究代表者
内田　智之（UCHIDA TOMOYUKI）
広島市立大学・情報科学研究科・准教授
研究者番号：70264934

(2)研究分担者
正代　隆義（SHOUDAI TAKAYOSHI）
九州国際大学・国際関係学部・教授
研究者番号：50226304

宮原　哲浩（MIYAHARA TETSUHIRO）
広島市立大学・情報科学研究科・准教授
研究者番号：90209932

(3)連携研究者
糸川　裕子（ITOKAWA YUKO）
広島国際大学・心理科学部・助教
研究者番号：40341234

鈴木　祐介（SUZUKI YUSUKE）
広島市立大学・情報科学研究科・助教
研究者番号：10398464