様式 C-19

# 科学研究費補助金研究成果報告書

平成２３年３月３１日現在

機関番号：25403
研究種目：基盤研究（C）
研究期間：2008 ～ 2010
課題番号：20500140
研究課題名（和文）　大規模マルチメディアコンテンツからのデータマイニングとその応用

研究課題名（英文）　Data mining from large multimedia contents and its applications

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

研究成果の概要（和文）：
マルチメディアを含む大規模なコンテンツにおける異種属性（たとえば、キーワード、画像内オブジェクト、グラフ構造など）の相関を特徴として抽出するグラフマイニング手法を開発した。特に、マルチメディアコンテンツ内で頻出するキーワード（単語や画像内オブジェクトなど）とそれらキーワードの出現位置との関係を表す木パターンに対して、機械学習理論に基づいた多項式時間パターン照合アルゴリズムを提案し、効率的なグラフマイニングアルゴリズムを提案した。また、ユーザ・オリエンティッドな情報検索システムの基礎を構築するために、抽出されたパターンの管理方法や高速・省メモリパターン照合アルゴリズムを提案した。

研究成果の概要（英文）：
We proposed graph mining techniques for extracting all correlations of different kinds of attributes (e.g., keywords, objects in images, and graph structures) from large amounts of contents including multimedia. We formally defined a tree pattern, which can represent characteristics such as the usage of keywords (e.g., words, objects in image, and the structural relations among nodes in which characteristic words appear) in multimedia contents. Based on computational machine learning, we proposed efficient pattern matching algorithms and efficient graph mining algorithms for tree patterns and graph patterns. Moreover, in order to construct a theoretical base of user-oriented information retrieval system for multimedia contents, we proposed efficient methods of managing all extracted tree patterns.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| ２００８年度 | 1,300,000 | 390,000 | 1,690,000 |
| ２００９年度 | 1,200,000 | 360,000 | 1,560,000 |
| ２０１０年度 | 1,000,000 | 300,000 | 1,300,000 |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 3,500,000 | 1,050,000 | 4,550,000 |

研究分野：計算機科学
科研費の分科・細目：情報学・知能情報学
キーワード：グラフアルゴリズム、機械学習、データマイニング、情報検索システム

## １．研究開始当初の背景

多くの家庭でインターネット環境が整ったことや、コンピュータの高速化、補助記憶装置の大容量化、携帯電話を含む通信技術の進歩により、テキストのみならず静止画や動画像などを有する Web ページ(マルチメディアコンテンツ)が多くなってきている。たとえば、NHK オンラインサイト

(http://www.nhk.or.jp/)では、ニュース記事をテキストと動画の両方で配信しつつ、フラッシュを使用した動画や静止画をふんだんに利用したサイトになっている。さらに、近年、インターネットを介して録画予約ができるハードディスクレコーダなど、家電製品とインターネットの融合が進みつつある。これらの現状を踏まえると、今後マルチメディアコンテンツは増加すると考えられる。

機械学習に基づいた、大規模な構造データから頻出する部分構造を高速に発見する手法やデータストリームからのマイニング手法の研究は申請者らを含め盛んに行われている。しかし、本研究のテーマである、グラフ構造と複数の異種属性を融合した特徴的なパターンを効率よく発見するマイニング手法の研究はまだ少なく、さらにコンテンツの構造を含むことでユーザ・オリエンティッドな検索システムの基盤を構築しようとする試みは少ない状況であった。

２．研究の目的

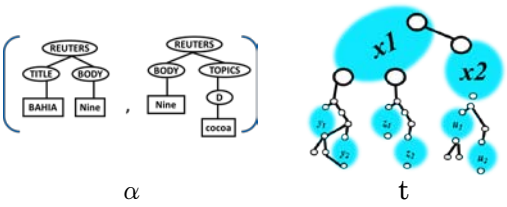
ユーザが最近閲覧した静止画や動画を含むマルチメディアコンテンツにおいて共起する異種属性（たとえば、キーワード、動画像内オブジェクト、Web ページ内の位置、グラフ構造など）の相関を特徴として抽出するデータマイニング手法を開発し、ユーザ・オリエンティッドな高精度情報検索システムの基礎を構築することを目的とする。

３．研究の方法

(1) HTML/XML などで記述されている Web ページのグラフ構造は順序木で表現することができる。また、静止画や動画などの時系列データでは、P ツリーと呼ばれる木構造を用いることで、それらの有している特徴（オブジェクトの位置情報やオブジェクトの変化量など）を表現することができる。このことより、マルチメディアコンテンツそのもののデータモデルとして順序木を用いる。Web ページ内のレイアウト情報や動画・静止画から抽出できるグラフ構造のデータモデルとして外平面的グラフや TTSP グラフも扱う。

(2) 動画・静止画のもつ特徴とそれらが出現するマルチメディアコンテンツのグラフ構造の特徴との相関を統一的に表現するために、単語と木構造の相関関係を表現できる単語間連結経路パターン(Consecutive Path Pattern, CPP)とその拡張である単語間木構造パターン(Tree Association Pattern, TAP)を、さらには構造的変数を有する項木(Term Tree)をそれぞれ知識表現として用いる。図 1 に CPP と項木の例を与える。

(3)過去に構造的変数をもつ木やグラフをそれぞれ項木と項グラフという木パターンとグラフパターンとして定式化した。さらに、Formal Graph System(FGS)という項グラフを項としてもつ一階述語論理を提案した。これら項木や項グラフをパターンとみなすことでグラフ言語（項木言語や項グラフ言語）を定義することができる。これらグラフ言語を対象にした機械学習可能性について計算量理論的な検討を加えることにより、項木や項グラフに関する高速・省メモリパターン照合アルゴリズムの設計を行う。これにより、大規模マルチメディアコンテンツを扱えるようにする。




α　　t

図 1：CPP α と項木 t

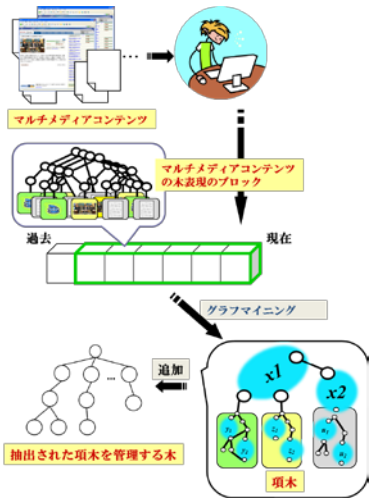



図 2：開発するシステムのイメージ

(4) 上述の 3 で開発するパターン照合アルゴリズムを用いて、知識表現として木パターン(CPP、TAP および項木)をもつマルチメディアコンテンツからのオンライングラフマイニング手法の設計および実装をそれぞれ行う。併せて動画・静止画および Web 内のレイアウト情報等の特徴を項木および項グラフとして抽出するグラフマイニング手法の設計および実装を行う。

(5) マイニング結果としての木パターン(CPP,TAP,項木)を効率よく管理するデータ構造を設計・実装する。さらにパターン照合アルゴリズムやグラフマイニングアルゴリズムの高速・省メモリ化を図る。これにより、本研究成果の応用として、ユーザ・オリエンティッドな検索システムの基盤を構築する。

４．研究成果

開発したグラフマイニングシステム(マルチメディアコンテンツのデータモデルや知識表現としての項木およびユーザ・オリエンティッドな高精度情報検索に必要なパターン管理木の流れ)のイメージを図２に示す。

以下に本システムを開発する際に得られた研究成果を述べる。

(1) マルチメディアコンテンツから順序木を構築するために XML パーザー Xerces(http://xerces.apache.org/xerces-e/)を、マルチメディアコンテンツにある画像やリンクされている動画から特徴を抽出するために画像処理・画像認識用ライブラリ OpenCV(http://sourceforge.net/projects/opencvlibrary/)を用いることにより、インターネット上の実データを実用的な時間で順序木やグラフ構造へ変換できることが確認できた。

(2) マルチメディアコンテンツから特徴的な単語とそれら単語の出現位置関係を特徴として統一的に抽出するために、2004 年に研究代表者らが既に提案している単語間連結経路パターン(CPP)を、CPP より厳密な表現が可能な単語間木構造パターン(TAP)へと拡張した。さらに、マルチメディアコンテンツをブラウザ上で表示する際のレイアウト情報の多くは外平面的グラフや TTSP グラフでモデル化することができるため、それぞれブロック保存型外平面的グラフパターン(Block-Preserving Outerplanar Graph Pattern, BPO グラフパターン)や TTSP 項グラフパターンを提案した。

(3) 機械学習に基づいたグラフマイニング手法を開発するために、まず項木に対する質問学習可能性および多項式時間機械学習可能性を明らかにした。木構造パターン以外にも、TTSP 項グラフパターンを構文木に一意的に変換する手法と、TTSP 項グラフパターンに対するパターン照合アルゴリズムを与えることで、TTSP 項グラフパターンに対する正データからの多項式時間帰納推論可能性を明らかにした。さらに、BPO グラフパターンの多項式時間照合アルゴリズムを与えることにより、BPO グラフパターンに対する正データからの多項式時間帰納推論可能性も明らかにした。

(4) マルチメディアコンテンツから、構造的特徴を表す極大頻出 TAP を枚挙する効率的なグラフマイニング手法を提案した。さらに、3 で得られた項木、TTSP 項グラフパターン、BPO グラフパターンに対する多項式時間パターン照合アルゴリズムを用いて、それぞれ極大な頻出パターンを発見するグラフマイニングアルゴリズムを提案し、計算機上に実装した上で評価実験を行い、それぞれのグラフマイニングアルゴリズムの有効性を示した。

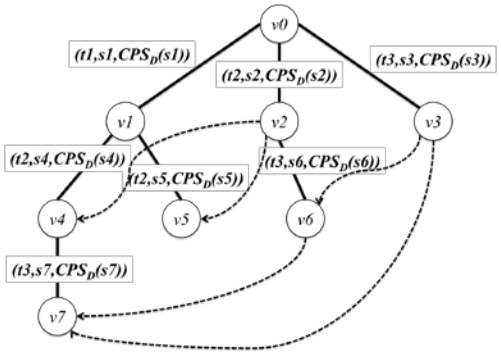



図 3：TAP を管理する TAP 木

(5) マルチメディアコンテンツから枚挙された極大頻出 TAP を管理する TAP 木（図 3 参照）を提案した。その TAP 木には、根から各ノードへのパス上にそのノードで管理する TAP 内に出現する CPP を配置することで効率的な管理を実現した。さらに、TAP を含むノードラベル付き順序木を、完全復元を保証して高圧縮する方法を提案し、計算機上に実装し評価実験を行ってその有効性を示した。また、順序木に対する簡潔データ構造を項木に対応させることで、簡潔データ構造を用いた項木に対する高速・省メモリパターン照合アルゴリズムを与えた。過去に提案されている項木に対するパターン照合アルゴリズムはボトムアップ方式であったが、本研究で開発したものはトップダウン方式である。両者の最悪時間計算量は変わらないが、実機での評価実験では、今回提案したアルゴリズムの方が定数倍速いことが確認できた。さらに、ユーザ・オリエンティッドな検索システムの基盤を構築するために、XSLT 変換言語(XSLT transformation language)を用いて、抽出された特徴的な単語間のグラフ構造パターンを表す CPP から構造変換をもたらすラッパーを半自動変換する手法を提案し、実装を行ってその有効性を確認した。

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 8 件）

① Y. Itokawa, M. Wada, T. Ishii and T. Uchida, Tree Pattern Matching Algorithm Using a Succinct Data Structure, Proc. International MultiConference of Engineers and Computer Scientists (IMECS2011), 査

読有, 2011, pp.206-211.

② Y. Itokawa, K. Katoh, T. Uchida, and T. Shoudai. Algorithm Using Expanded LZ Compression Scheme for Compressing Tree Structured Data, Lecture Notes in Electrical Engineering, 査読有, 52, 2010, pp.333-346.

③ Y. Itokawa, J. Miyosi, M. Wada and T. Uchida, Succinct Representation of TTSP Graphs and Its Application to the Path Search Problem, Proc. the Sixth IASTED International Conference on Advances in Computer Science and Engineering (ACSE 2010), ACTA Press, 査読有, 2010, pp.33-40 .

④ T. Uchida and K. Kawamoto, Algorithm for Enumerating All Maximal Frequent Tree Patterns among Words in Tree-Structured Documents and Its Application, International Journal of Database Theory and Application, 査読有, 2(4), 2009, pp.59-73.

⑤ S. Matsumoto, Y. Suzuki, T. Shoudai, T. Miyahara, and T. Uchida, Learning of Finite Unions of Tree Patterns with Repeated Internal Structured Variables from Queries, IPSJ Transactions on Mathematical Modeling and its Applications, 査読有, 2, 2009, pp.127-137.

⑥ H. Yamasaki, Y. Sasaki, T. Shoudai, T. Uchida, and Y. Suzuki, Learning block-preserving graph patterns and its application to data mining, Machine Learning, 査読有, 76, 2009, pp.137-173.

⑦ R. Takami, Y. Suzuki, T. Uchida, and T. Shoudai, Polynomial Time Inductive Inference of TTSP Graph Languages from Positive Data, IEICE TRANSACTIONS on Information and Systems, 査読有, E92-D, 2009, pp.181-190.

⑧ M. Nagamine, T. Miyahara, T. Kuboyama, H. Ueda and K. Takahashi, Evolution of Multiple Tree Structured Patterns from Tree-Structured Data using clustering, Proc. 21st Australian Joint Conference on Artificial Intelligence (AI-2008), 査読有, Lecture Notes in Artificial Intelligence Vol.5360, 2008, pp.500-511.

〔学会発表〕(計3件)

① 鈴木 祐介, Enumerating Maximally Frequent TTSP graph patterns, Seventh Workshop on Learning with Logics and Logics for Learning (LLLL 2011), 平成23年3月30日, 大阪大学中之島センター.

② 河野 達哉, 単語間木構造パターンを用いたTTSPグラフの頻出部分グラフ枚挙手法, 平成21年度 電気・情報関連学会中国支部第60回連合大会, 平成21年10月17日, 広島市立大学.

③ 和田 将信, TTSPグラフに対する簡潔表現とパス探索問題への応用, 平成20年度 電気・情報関連学会中国支部第59回連合大会, 平成20年10月25日, 鳥取大学

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)
九州大学・システム情報科学研究院・准教授
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