科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１１　日現在

機関番号：　１２１０１
研究種目：　若手研究(B)
研究期間：　2010～2013
課題番号：　２２７００００２

研究課題名（和文）グラフ上に現れる記号列に対する文法処理手法の開発とその応用

研究課題名（英文）Development of a grammatical processing method for symbols on graph structures and its applications

研究代表者
藤芳　明生（Fujiyoshi, Akio）

茨城大学・工学部・准教授

研究者番号：００３２３２１２

交付決定額（研究期間全体）：（直接経費）　　2,700,000円、（間接経費）　　810,000円

研究成果の概要（和文）：グラフ上に現れる記号列に対する文法処理手法の開発を行った。文字列を対象とする文法処理手法は確立されているが、グラフ上の記号列を対象とする研究はほとんど行われてこなかった。本研究課題では、ラベル付き多重有向グラフに対し木オートマトンによって受理される全域木を発見する線形時間アルゴリズムの開発に成功した。この結果は、応用研究として開発している数式OCR及び化学構造式OCRの精度及び速度を向上させために、大変有効である。

研究成果の概要（英文）：This study develops a grammatical processing method for symbols on graph structures. Though grammatical processing methods of string have already been established, a method for symbols on graph structures has not been studied yet. This study successfully develops an algorithm solving the membership problem of labeled multidigraphs of bounded tree-width for a spanning tree automaton. This result can be applied to the developments of robust and efficient mathematical OCR and chemical structure formulae OCR. The developments of these OCR are also the purpose of this study.

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：　情報学基礎　形式言語理論　グラフ理論　アルゴリズム　数式OCR

１．研究開始当初の背景

本研究課題を開始する以前から、文字列を対象とする文法処理の研究は、活発に行われている。例えば、文脈自由言語には、CYK 法やアーリー法等の古典的アルゴリズムが存在し、最近では、一般化 LR 法等のより高速なアルゴリズムの提案も行われている。また、大量の文字列データに文法処理を施すことによるデータ検索を効率的に行おうとする研究も進んでいる。

一方、グラフ上に現れる記号列に対する文法処理の研究は、ほとんど進んでいなかった。これは、文字列上では効率良く解ける問題の多くが、グラフ上では NP 困難となり、現在のコンピュータの能力を持ってしても計算に天文学的時間が必要だとされていたためである。しかしながら、近年、グラフ・アルゴリズムの分野において、木幅（Tree-Width）を制限したり、特定の特徴を持つグラフに限定したりすると、これまで NP 困難とされていた問題も線形時間や多項式時間で解けることが分かってきている。そこで、グラフ上に現れる記号列に対する文法処理にも同様の手法が適応できるのではないかと予想される。

研究代表者は、本研究課題を開始する以前から、直並列グラフに限定すれば、グラフ上に現れる記号列に対する文法処理が線形時間や多項式時間で行えることを既に発見していた。最新のグラフ・アルゴリズムの手法を取り入れることで、より広いグラフ・クラスについても、グラフ上に現れる記号列に対する効率の良い文法処理手法が開発できると考えていた。

グラフ上に現れる記号列に対する文法処理手法は、様々な応用分野に適応できる。その中でも、本研究課題では、数式 OCR 及び化学構造式 OCR に注目した。論文や特許などの大量の科学技術情報がインターネット上に公開され、必要な情報を簡単に検索できるようになった。しかし、それらの多くは PDF ファイル形式等であるため、数式や化学構造式が図形として記録されている。そのため、数式や化学構造式を利用した検索はほとんど不可能な状態である。数式や化学構造式を電子的に取り扱うためには、それらを数式 OCR や化学構造式 OCR を用いて検索可能な形式で電子化することが必要である。

数式 OCR では、スキャンされた画像中の記号の接続関係を解析し、接続関係の候補を表現するグラフを構成し、グラフ上に現れる記号同士の接続関係の中から最適な組み合わせを求める。文法的に正しい記号の接続関係を求めるためには、効率の良い文法処理手法が求められていた。同様に、このような技術は化学構造式 OCR の開発にも役立つものと考えられる。

２．研究の目的

本研究課題の目的は、第一に、研究代表者が積み重ねてきた形式言語理論の研究成果を活用し、グラフ上に現れる記号列に対する文法処理手法の開発を行うことである。文字列に対する文法処理の研究は既に多く行われているが、本研究課題ではグラフを文法処理する方法を考える。特に、以下に上げる課題について重点的に取り組むこととした。

（１）既存の文字列文法、文字列オートマトン、木文法、木オートマトンを用い、グラフ上に現れる記号列の中から文法的に正しい文字列分割または全域木を抽出する方法の開発。

（２）効率の良い認識アルゴリズムを伴うグラフ文法、グラフ・オートマトンの開発。

（３）大量のグラフデータに文法的な特徴量を与えることによるデータ検索手法の開発。

様々なクラスのグラフに対し、効率の良いアルゴリズムの開発に努め、計算量の解析も同時に行うこととした。

第二は、本研究の理論的成果を応用し、数式 OCR 及び化学構造式 OCR の精度及び速度を向上させることである。これらの OCR は、紙に印刷された膨大な量の学術情報を電子化するために必要なソフトウェアである。

３．研究の方法

本研究課題は、文法処理アルゴリズムの開発に関する理論研究と、それを数式 OCR 及び化学構造式 OCR の開発に役立てるという応用研究である。それぞれ分けて記述する。

（１）理論研究として行う文法処理アルゴリズムの開発は、解決が必要とされている問題を分析し、抽象化することから始める。問題をうまくグラフ上の問題に抽象化できれば、そのグラフ上の問題を解く手法が既知でないかを調べる。これには、まずは文献を検索するが、それでも見つからない場合には専門家に尋ねるなどする。既知の問題でないならば、新しいアルゴリズムの開発を行う。アルゴリズムが実装に向いているか、計算量が膨大にならないか等にも注意する。

（２）応用研究として行う数式 OCR 及び化学構造式 OCR の開発は、現在、国内で最も先進的な OCR システムの開発を行っている九州大学大学院・数理学研究院・鈴木研究室と協力し、研究を進めて行く。まず、数式や化学構造式に現れる記号間の正しい接続関係を定義する文法を設計する。これには、数式コーパス及び化学構造式コーパスを利用し、機械学習と手作業を組み合わせて、最適な文法を求める。次に、スキャンされた画像から、それぞれの記号の認識候補及び記号間の接続関係候補をグラフで表現する方法を考え、文法的に正しい数式や化学構造式を表現する部分グラフを抽出するアルゴリズムを設計し、実装する。スキャンする画像がきれいであるとは限らず、大量のノイズが乗っていたり、文字が滲んでいたりすることもある。そのような画像にも対応できるように、様々な記号の認識候補の中から最適な物を選択で

きるようにアルゴリズムを設計する。

４．研究成果

本研究課題は、理論研究として、文法処理アルゴリズムの開発と、その応用研究として、それを数式 OCR 及び化学構造式 OCR の開発を行った。各年度のそれぞれの研究成果は以下の通りである。

（１）平成 22 年度

理論研究として、木幅（Tree-Width）の制限された一般のグラフに対し木オートマトンによって受理される全域木を発見する線形時間のアルゴリズムを開発するという成果を出した。直並列な無閉路有向グラフに限定すれば同様のアルゴリズムは既に開発していたが、その結果を一般のグラフに拡張した意味は大きい。この結果は数式 OCR だけでなく、化学構造式 OCR、化学式検索などに応用が可能である。

応用研究として、数式 OCR の文法処理アルゴリズムのプロトタイプの作成を行った。文法モデルとして、研究代表者が得意とする単項・線形・文脈自由木文法を導入した。この文法モデルには入力サイズの三乗時間で構文解析するアルゴリズムが得られているため、それの実装を行った。実装に用いた文法には、数式 OCR の認識精度を改善させるために様々なチューニングを行った。

また、数式 OCR の文法処理アルゴリズムに応用するために、拡張した最小全域木問題を考え、それを解くアルゴリズムの提案を行った。最小全域木問題は頂点ラベルなしグラフ上で考えるものであるが、この拡張した最小全域木問題では頂点ラベル付グラフ上で議論を行う。辺を結ぶ頂点のラベルの選択次第で、辺の重みが変わるというように拡張されているのである。数式 OCR では、頂点のラベルとは、各文字の認識候補を表し、辺の重みは、位置関係やバイグラムに基づいた認識結果の尤もらしさを表している。この拡張した最小全域木問題を解くための線形時間のアルゴリズムを開発し、提案を行った。

（２）平成 23 年度

理論研究として、ラベル選択付最小全域木問題を一般化したラベル選択付最小全域部分グラフ問題に取り組んだ。最小連結全域部分グラフ問題とは、各辺に「繋ぐ場合の重み」と「繋がない場合の重み」の 2 種類を与え、選ばれた辺の「繋ぐ場合の重み」の合計と選ばれなかった辺の「繋がない場合の重み」の合計の和が最小となるような連結全域部分グラフを求める問題である。ラベル選択付最小全域木問題が NP 困難であるため、この問題を拡張したラベル選択付最小連結全域部分グラフ問題も同様に NP 困難である。しかし、入力のグラフの木幅を 2 以下に制限した場合であれば、この問題が線形時間で解けることを示した。更に、線形時間で動作する非常にシンプルなアルゴリズムの提案を行った。この結果は数式 OCR だけでなく、化学構造式 OCR、化学構造式検索などに応用が可能である。

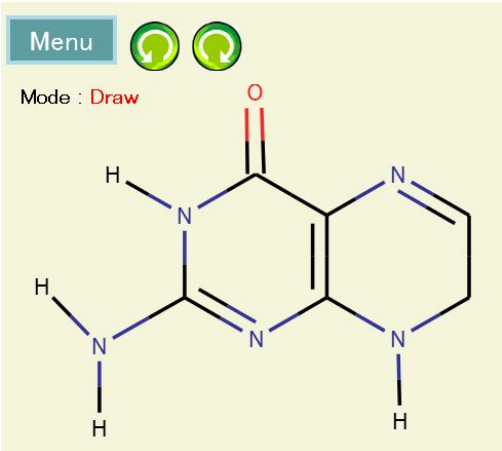
文法処理を行う上で、多くの認識候補が存在する場合、文法によってそれらの順序付けを行うことが必要になる。つまり、最適な認識候補を出力するための順位が必要となるのである。本研究では、ラベル選択付最小全域部分グラフ問題を解くことで、その順位付けを行うことを検討している。

応用研究として、昨年度までに開発された認識アルゴリズムと、数式 OCR 開発の経験と技術を応用し、化学構造式 OCR の開発を行った。日本の公開特許公報中には多数の化学構造式が存在するが、現在、手作業での電子化作業が行われており、時間と費用が膨大なものとなっている。社会貢献の度合いは、数式 OCR より化学構造式 OCR の方が大きいのではないかと考え、開発の重点を化学構造式 OCR におくこととした。

（３）平成 24 年度

理論研究として、一般のグラフに対し木オートマトンによって受理される全域木を発見する問題が線形時間 FPT（Fixed Parameter Tractable）であることを証明した。線形時間 FPT であるとは、すなわち、線形時間でこの問題を解決するアルゴリズムが存在することを意味する。木幅が制限されたグラフに限定したり、直並列な無閉路有効グラフに限定したりすれば線形時間のアルゴリズムはすでに発見していたが、一般的のグラフに対してこの問題が線形時間で解決可能であることの証明を行った。国際会議で論文発表を行った。この結果は数式 OCR だけでなく、化学構造式 OCR、化学式検索などに応用が可能である。

応用研究として、昨年度までに開発された認識アルゴリズムと、数式 OCR 開発の経験と技術を応用し、化学構造式 OCR の開発を引き続き行っている。化学構造式 OCR の認識結果をスマートフォンやタブレット端末で確認するためのツールとして、Web アプリケーションの化学構造式エディタ MolTouch を開発した。MolTouch は茨城大学の Web サーバ上で公開を開始している。



Moltouch

また、応用研究の一つとして、PDF ファイルのレイアウト解析プログラムの開発を行っている。これは、OCR の技術を応用したものである。

（４）平成 25 年度

理論研究として、前年度までの結果を更に一般化し、ラベル付き多重有向グラフに対し木オートマトンによって受理される全域木を発見する問題が線形時間 FPT（Fixed Parameter Tractable）であることを証明したことである。ラベル付き多重有向グラフは、ラベル無しグラフ、無向グラフ、単純グラフの全てを一般化したグラフであり、これに対する結果は大変強力であるといってよい。つまり、本研究課題がこの数年で出した結果の多くを含むような一般化された結果となっている。この結果は、国際会議で発表予定であり、学術誌に論文を投稿準備中である。

応用研究として、これまでに開発された認識アルゴリズムと、数式 OCR 開発の経験と技術を応用し、化学構造式 OCR の開発を引き続き行っている。　日本の公開特許公報中には多数の化学構造式が存在するが、現在、手作業での電子化作業が行われており、時間と費用が膨大なものとなっている。社会貢献の度合いは、数式 OCR より化学構造式 OCR の方が大きいのではないかと考え、開発の重点を化学構造式 OCR におくこととした。

また、新しい応用研究として、マルチモーダル書籍の開発に着手した。マルチモーダル書籍とは、「見えない２次元コードが重ねて印刷された紙の書籍を読む」と「２次元コード読取装置で対応する音声を聞く」という動作を同時に行い、視覚と聴覚を複合的に用いることで、効率的な読書を可能とする紙の書籍である。高齢者や読字障害者だけでなく、一般の人々にも新しいスタイルの読書を提案するものである。 OCR の技術を応用し、マルチモーダル書籍の半自動作成システムを完成させた。これは、PDF ファイル等の文章に対し、レイアウト解析、文章自動抽出、読みの付与、音声合成等を半自動的に行うものであり、効率的なマルチモーダル書籍の作成を可能にする。マルチモーダル教科書の試作を行い、読字障害の小学生に提供し、実証実験中である。

マルチモーダル書籍の半自動作成システム

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藤芳衛，大澤彰子，小山田寛史，薬師寺駿介，青松利明，澤崎陽彦，藤芳明生，「文字認知障害者のための２次元コード方式のリスニングテストシステムの開発」，大学入試研究ジャーナル，no.23，113-120（2013）査読有

② Akio Fujiyoshi and Masakazu Suzuki, “Minimum Spanning Tree Problem with Label Selection”, IEICE Transactions on Information and Systems, vol.E94-D, no.2, 233-239 (2011) 査読有

③藤芳衛，南谷和範，藤芳明生，青松利明，澤崎陽彦，「読字障害者および重度の弱視者のための文字と音声のマルチモーダル問題の開発」，大学入試研究ジャーナル，no.21，181-190（2011）査読有

④ Akio Fujiyoshi, “Recognition of Directed Acyclic Graphs by Spanning Tree Automata”, Theoretical Computer Science, vol.411, no.38-39, 3493-3506 (2010) 査読有

⑤ Akio Fujiyoshi, Masakazu Suzuki and Seiichi Uchida, “Grammatical Verification for Mathematical Formula Recognition Based on Context-Free Tree Grammar”, Mathematics in Computer Science, vol.3, no.3, 279-298 (2010) 査読有

⑥Akio Fujiyoshi and Masakazu Suzuki, “A Variation of the Minimum spanning Tree Problem for the Application to Mathematical OCR”, Journal of Math for Industry, vol.2, 183-197 (2010) 査読有

〔学会発表〕（計　１１件）

①藤芳明生，藤芳衛，大澤彰子，「文字認知障害児の能動的な学習を可能にする文字と音声のマルチモーダル教科書の試作」，日本教育工学会 第 28 回全国大会 講演論文集，547-548，2012.9.15，長崎大学

②藤芳明生，鈴木昌和，「ラベル選択付最小連結全域部分グラフ問題と化学構造式ＯＣＲへの応用」，冬のＬＡシンポジウム，2012.2.1，京都大学数理解析研究所

③藤芳明生，「木オートマトンを用いた化学グラフのスクリーニング手法」，冬のＬＡシンポジウム，2011.2.3，京都大学数理解析研究所

〔産業財産権〕

○出願状況（計　１件）

名称：２次元コード読取装置
発明者：藤芳　明生
権利者：国立大学法人茨城大学
種類：特許

番号：特許出願２０１２－１１４１７８
出願年月日：２０１２年5月１８日
国内外の別： 国内

〔その他〕
ホームページ等
http://apricot.cis.ibaraki.ac.jp/moltou
ch/

６．研究組織
(1)研究代表者
藤芳　明生（FUJIYOSHI　AKIO）
茨城大学・工学部・准教授
研究者番号：００３２３２１２

(2)研究分担者
無し

(3)連携研究者
無し