様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月 １１ 日現在

機関番号：１２６１２
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ１１５４７

研究課題名（和文）セマンティックWebデータの多様性に対する推論と学習の基盤技術

研究課題名（英文）A Method of Reasoning and Learning for Various Data on the Semantic Web

研究代表者
兼岩　憲（Kaneiwa, Ken）

電気通信大学・大学院情報理工学研究科・教授

研究者番号：００３４２６２６

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では、セマンティックWebにおけるオントロジーやグラフデータに対する機械学習の２つの方法を新たに提案した。１つ目の方法は、オントロジーの公理から概念包含やエンティティ間関係を推論することが可能である。２つ目の方法は、グラフデータの構造的特徴を訓練してクラス分類タスクを実行できる。これらの２つの方法は従来手法の精度よりも高いパフォーマンスをもたらすことを実験で評価している。また、機械学習の前処理として、オントロジーやグラフデータから特徴を抽出できるように高速なRDFデータストアを改良している。

研究成果の学術的意義や社会的意義
ここ数年で画像認識を中心に機械学習の成果が大きく社会で応用され、続いて人が解釈して入力したWeb規模の事実データがもたらす機械学習の成果に期待が膨らむ。そのためセマンティックWebの膨大なデータを検索するだけでなく、その知識から自動的に推論する機械学習メカニズムを提案した本研究の成果がその技術確立に寄与する。

研究成果の概要（英文）：We have proposed two methods of machine learning for ontology and graph data on the Semantic Web. The first method enables us to infer concept subsumption and entity relations from ontology axioms. The second method can train the structural features of graph data for node classification tasks. We have evaluated that our methods outperform the accuracies of conventional methods. For the preprocessing of machine learning, we have improved a fast RDF store system that extracts features from ontology and graph data.

研究分野：知識表現と推論

キーワード：セマンティックウェブ　RDF

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景
(1) 近年、スマートフォンなどを用いてWebから必要な情報を入手することが容易になり、知識を見つけることよりも今後は膨大な知識をどのように活かすかへ関心が移っている。即ち、機械学習技術の飛躍的な発展により、膨大なWebデータを利用して多くの知識処理を人間の代わりに自動化できないかと考えられている。

(2) しかしその実現には、次の３つの大きな現実的・技術的問題が存在する。まず、Web上の情報は未だに多くがドキュメント中心であり、計算機が自動処理できる構造化データとは言いがたい。次に機械学習によって画像や音声に対する分類や認識の能力が飛躍的に上がったが、（開世界の）セマンティックWebデータに対するマイニングや推論を可能にする機械学習には技術的な難しさが残る。さらに、Webデータから自動的に推論して現実的な応用問題を解決する際に、対象分野に固有のドメイン知識を用意する必要がある。

２．研究の目的
(1) 従来ドキュメント中心のWebでは人が読むためのWebページが公開されていたが、セマンティックWebではインターネット上にとてつもなく大きなデータのWeb空間を構築しようとしている。本研究では、そのようなデータのWebを構成するリンクトデータ（セマンティックWebデータ）を自動的に活用するために検索・推論・機械学習を備えた処理エンジンを開発する。リンクトデータはグラフ構造によって異なるデータセットがグローバルに連結するが、このグラフ構造を推論や機械学習に適用するのは容易ではない。

(2) 本研究では、理論上のグラフ構造でなく、データベース上に実在する大規模かつ冗長なグラフ構造から価値ある意味データを抽出し、単なる検索を超えて質問の答えを推論・学習するメカニズムを提案する。そのために、従来のデータベースとは違ったセマンティックWeb専用のデータストア、推論エンジン、および機械学習メカニズムを新たに独自開発する。

３．研究の方法
(1) 本研究では、Web上で無限に広がっていくセマンティックWebのデータ規模やリンク構造から機械学習や推論を可能にするメカニズムを提案する。その実現のために、以下の研究方法をとる。

(2) １つは、RDFデータをなす（主語、述語、目的語からなる）三つ組（トリプルと呼ぶ）から学習や推論に不要な部分をスキップする特徴抽出手法の研究である。その手法によりRDFの意味データから適切に訓練して、問題に適用した高い精度をもつ学習器を生成する。例えば、人物に関する事実がRDFで記述されているとき、それらの優劣などを人物の特徴データから訓練する。そのとき、トリプルによって各人物の情報が不統一なリンクでつながって書かれており、グラフ構造内を探索して優劣を学習する際に（不要な情報を除いた）本質的な特徴パターンを抽出できるグラフカーネルを設計する。本アプローチでは、ある程度自由度をもったグラフカーネルを用意し、データの特徴抽出に適すように事前学習を行う。即ち、汎用グラフカーネルを設計してそれを様々な対象問題に適用できるようにパラメータを事前学習して、そのパラメータを入力したグラフカーネルを使って対象問題を解く学習器を生成する。これは一種の多層学習であり、ディープラーニングが成功した考えにも通じる。

(3) もう１つは、大規模かつ複雑なRDFデータから計算コストの高いグラフカーネルやそれを用いた機械学習を現実時間で解くために、高速なRDFストアを実装する。研究代表者が開発したFROSTは高速検索と高いデータ圧縮（省メモリ化）を実現している。圧縮後のRDFデータは濃縮された本質的な特徴を表しており、高速検索と合わせてグラフカーネルの計算にどう貢献するか明らかにしていきたい。この点は、特にRDF専用のデータベースを自前で開発している研究室ならではの成果である。

４．研究成果
(1) 初年度は、本研究の目的である膨大なセマンティックWebデータから機械学習や推論を自動的に行うための問題点とその問題点を解決する方法論の基礎を分析した。セマンティックWebのRDFデータは様々な分野や作成者によるデータセットから作られており、機械学習や推論への適用にはその多様性と冗長性が問題である。そうしたデータから推論・学習を実現するために、異なるデータセットを融合させる方法を提案した。特に、対象（エンティティ）の同一性を示すプロパティを用いて、DBpediaやWikidataなどの異なるデータをつなげて個々のデータセットでは推論できない検索を実現している。その際、複数のキーワードから出発してリンクトデータ（RDFのリンク構造）を辿って意味構造による検索結果を出力する。この意味構造は次年度以降

のリンクトデータからの特徴抽出に寄与する技術である。
　また、公開されているリンクトデータは膨大だが内容に偏りがあり、機械学習や推論に不足するデータが考えられる。その解決のために、自然言語テキスト文から自動的に RDF データやオントロジーを作成する技術を進展させた。従来のように自然言語文から述語項構造を抽出する研究の発展として、文意を表した知識構造を名前空間を付与した知識データとして構築している。この研究成果により，生成された RDF データと公開済みのリンクトデータを合わせれば機械学習と推論の精度を向上させると考えられる。

(2) ２０１９年度は、リンクトデータの大規模性や多様性から学習・推論するためのプロトタイプの実装とその性能の分析を行った。RDF データセットから２値分類を行うための訓練データを用いて、RDF データを構成する主語、述語、目的語からなる三つ組（トリプルと呼ぶ）から学習や推論に不要な部分をスキップして特徴ベクトル化する手法を開発した。この特徴ベクトルは、トリプルから様々な組み合わせで部分構造を取り出し対象データ（主語リソース）がそれぞれの部分構造をもつかどうかをベクトルの要素で表現する。その結果、性質の異なる多様な RDF データに対して、機械学習（ニューラルネットワークなどの適用）に必要な特徴ベクトルを抽出できる。この抽出手法は、RDF データ特有のデータ設計に基づいた特徴表現と情報利得率による特徴選別を用いている点が新しい。
　本手法のプロトタイプを実装して、実際に正例と負例に分けた RDF の訓練データ（１０種類のデータ）を用いて性能実験を行った。実験では、特徴ベクトルの次元、情報利得率、割引率などのパラメータを変えて深層ニューラルネットワークを用いて学習し、テストデータを用いて学習結果の正解率を分析した。その実験結果の分析により、実際のデータセットによって高い正解率をもたらす対象データを特徴付けるデータが、属性の種類（述語）、属性の値（目的語）や属性の種類と値（述語と目的語の組）などと異なることが明らかになっている。

(3)２０２０年度は、計算コストの高い機械学習の特徴抽出において DBpedia などの現実のリンクトデータを利用するために高速な RDF ストアを実装した。この RDF ストアは、研究代表者がこれまで開発してきた FROST の改良版である。その RDF ストアは既に高いデータ圧縮と効率的な検索メカニズムを備えていたが、機械学習を想定したとき訓練データの生成に特化した機能が必要である。RDF データは関係データベースなどとは異なりスキーマレスとグラフ構造により非常に柔軟なデータ構造を備えており Web データに有効であるが、その反面、直接データを機械学習へ適用することを難しくしている、従って、機械学習に必須となる訓練データを構築する際に、RDF データから特徴ベクトルを抽出する検索処理を効率化した。その手法は、RDF ストアにトリプル（RDF データを構成する主語、述語、目的語からなる三つ組）を格納する際に、主語、述語、目的語という順にインデックスを作成していたのを、主語、目的語のみの順序にしている。これにより、RDF データのグラフ構造を抽出するアルゴリズムの計算量が大きく減少した。前年度に開発した RDF データからの特徴ベクトルの抽出をこの RDF ストア上で再実装して、より高速に特徴ベクトルの構築ができることを実験して確認できている。
　さらに RDF ストアの開発に加えて、セマンティック Web データを応用するために、ドメインオントロジーの構築と利用を調査している。そのオントロジーを利活用するために、記述論理による推論エンジンのプロトタイプを独自に実装して OWL オントロジーに対する概念の包摂関係を判定する実験を行った。

(4) ２０２１年度は、RDF データやオントロジーによる２つの異なるレベルのセマンティック Web データを学習して、リンク関係、論理的関係、およびクラス属性を推定する学習型推論システムを開発した。

①リンクトデータを表現する RDF データ（主語、述語、目的語からなるトリプル）に対する埋め込み学習とともに、オントロジーを表現する記述論理の知識ベースに対する埋め込み学習を実装した。これにより、セマンティック Web データから記述論理の推論アルゴリズムを使って論理的に推論するだけでなく、訓練済みの学習器から実データに不足している知識を新たに推定できるようになっている。即ち、従来の推論システムでは論理的な帰結しか導けないのに対して、学習型推論システムでは知識ベースに含まれていない概念のインスタンスデータ、個体間の二項関係、概念間の包含関係などを導くことができる。今回、RDF データとオントロジーの両タイプに機械学習を適用することで、ドメインオントロジーをリンクトデータに融合させた知識ベースでの推論が期待できる。

②さらに、前年度までに開発した特徴抽出の手法により、リンクトデータと類似性の高いグラフデータベースに対して、不要な知識をスキップしてデータを特徴ベクトル化する手法を適用した。これは人間関係や科学論文関係などのようなネットワーク構造で表されるグラフデータにおいて、各対象物（人間や科学論文）がどのようなクラスに属すか学習する。本分野で用いられているグラフニューラルネットワーク（GNN）と本研究が提案している特徴抽出の手法を組み合わせて高い推定能力（正解率）を実現した。

(5)最終年度は、これまで開発した学習型推論システムを用いて推論や学習の検証実験を行った。この検証実験は、①リンクトデータによる RDF データに対するオントロジーの推論タスク、および②グラフ構造をもつデータセットに対するクラス分類タスク、といった機械学習について２つのタスクの評価を行っている。

①１つ目の検証実験では、RDF データ形式上の OWL オントロジーとして記述されている GO(遺伝オントロジー)、FoodOn（食品オントロジー)、HeLis（健康オントロジー）を用いて実用的なオントロジーデータで学習型推論システムの有効性を評価した。本研究の成果として、オントロジーから記述論理による推論アルゴリズムによって推論できない、機械学習による包含関係の推定を実現している。特に、オントロジーに含まれる注釈文（自然言語文）から単語の共起や文脈を抽出する埋め込み手法や、オントロジーの論理的な公理を使って機械学習する精度の向上を提案している。その結果、従来手法よりも高い正解率でオントロジー上の推論タスクを実現させている。

②２つ目の検証実験では、科学論文の引用関係、Web 上のリンク関係、Wikipedia 内の人物などの関係を表すグラフデータセット（Pubmed，Texas，Actor など）を用いて各ノードに対するクラス属性を推定する能力を評価した。本研究の学習型推論システムは、グラフデータからグラフ構造を直接的に抽出し有益な情報だけを特徴選択する手法を提案している。この手法をグラフニューラルネットワーク（GNN）へ適用したとき、クラス属性の分類タスクで高い正解率を実現することを確認した。この成果の大きな意義は、グラフデータ上で隣接するノードが異なるクラス属性をもつ異種性グラフと呼ばれるデータセットでも機械学習の精度を低下させない点にある。

５．主な発表論文等

〔雑誌論文〕　計8件（うち査読付論文　8件／うち国際共著　0件／うちオープンアクセス　6件）

１．著者名 Yuki Odaka, Ken Kaneiwa	４．巻 Vol.142
２．論文標題 Block-Segmentation Vectors for Arousal Prediction using Semi-supervised Learning	５．発行年 2023年
３．雑誌名 Applied Soft Computing	６．最初と最後の頁 Article 110327
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.asoc.2023.110327	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Yuga Oishi, Ken Kaneiwa	４．巻 Vol.11
２．論文標題 Multi-duplicated Characterization of Graph Structures using Information Gain Ratio for Graph Neural Networks	５．発行年 2023年
３．雑誌名 IEEE Access	６．最初と最後の頁 34421-34430
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2023.3264596	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yuga Oishi, Ken Kaneiwa	４．巻 Vol.11
２．論文標題 Hierarchical Model Selection for Graph Neural Netoworks	５．発行年 2023年
３．雑誌名 IEEE Access	６．最初と最後の頁 16974-16983
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2023.3246128	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 兼岩 憲, 山中 佑紀	４．巻 Vol.38 No.2
２．論文標題 複数の大規模RDFデータセットからの同値関係による集結パス検索	５．発行年 2023年
３．雑誌名 人工知能学会論文誌	６．最初と最後の頁 D-M53_1-9
掲載論文のDOI（デジタルオブジェクト識別子） 10.1527/tjsai.38-2_D-M53	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 兼岩 憲，平山 健太	4．巻 Vol.21-J
2．論文標題 重複排除を含むSPARQLクエリの等価変換と計算量	5．発行年 2023年
3．雑誌名 日本データベース学会和文論文誌	6．最初と最後の頁 Article No.1
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 兼岩 憲，長井拓馬	4．巻 35
2．論文標題 極小RDF推論に基づく記述論理SROIQの概念生成	5．発行年 2020年
3．雑誌名 人工知能学会論文誌	6．最初と最後の頁 B-J62_1-13
掲載論文のDOI（デジタルオブジェクト識別子） 10.1527/tjsai.B-J62	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 荒井 大地，兼岩 憲	4．巻 33
2．論文標題 RDFグラフの多様性に対する汎用カーネル関数	5．発行年 2018年
3．雑誌名 人工知能学会論文誌	6．最初と最後の頁 B-I12_1-14
掲載論文のDOI（デジタルオブジェクト識別子） 10.1527/tjsai.B-I12	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 廣橋美紀，兼岩 憲	4．巻 17-J
2．論文標題 RDF データに対するグラフパターンマイニング	5．発行年 2019年
3．雑誌名 日本データベース学会和文論文誌	6．最初と最後の頁 Article No.1
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計13件（うち招待講演　0件／うち国際学会　0件）

1．発表者名
山中 佑斗，白石 幸寛，兼岩 憲

2．発表標題
概念包含の論理的推論に対するロール埋め込み手法

3．学会等名
第127回 知識ベースシステム研究会（SIG-KBS）

4．発表年
2022年

1．発表者名
大石 悠河，兼岩 憲

2．発表標題
グラフ構造のIGR多重特徴化によるグラフニューラルネットワーク

3．学会等名
第127回 知識ベースシステム研究会（SIG-KBS）

4．発表年
2022年

1．発表者名
白石幸寛，兼岩 憲

2．発表標題
個体と包含関係の論理的推論に対する埋め込みモデル

3．学会等名
電子情報通信学会 人工知能と知識処理研究会（AI）

4．発表年
2021年

1．発表者名
大石 悠河，兼岩 憲

2．発表標題
グラフニューラルネットワークに対する階層的モデル選択

3．学会等名
第124回 知識ベースシステム研究会（SIG-KBS）

4．発表年
2021年

１．発表者名
南 陽太，兼岩 憲

２．発表標題
ノード特徴とグラフ構造を抽出したベクトルによるノード分類

３．学会等名
第124回 知識ベースシステム研究会（SIG-KBS）

４．発表年
2021年

１．発表者名
高橋 大樹，兼岩 憲

２．発表標題
モンテカルロ木探索による記述論理の充足可能性判定

３．学会等名
第51回セマンティックウェブとオントロジー研究会

４．発表年
2020年

１．発表者名
小高祐輝，兼岩 憲

２．発表標題
半教師あり学習による覚醒度推定と感情分析

３．学会等名
電子情報通信学会 人工知能と知識処理研究会

４．発表年
2020年

１．発表者名
平山健太，兼岩 憲

２．発表標題
SPARQLにおける集約の形式化とクエリ書き換え

３．学会等名
第48回セマンティックウェブとオントロジー研究会

４．発表年
2019年

１．発表者名
南 陽太，兼岩 憲

２．発表標題
RDFグラフから抽出した多様な特徴ベクトルによる教師あり学習

３．学会等名
第49回セマンティックウェブとオントロジー研究会

４．発表年
2019年

１．発表者名
末木顕人，兼岩 憲

２．発表標題
Wikipedia記事からの中間RDFグラフとDBpediaトリプルの抽出

３．学会等名
第45回セマンティックウェブとオントロジー研究会

４．発表年
2018年

１．発表者名
山中佑紀，兼岩 憲

２．発表標題
複数の大規模RDFデータセットを統合したキーワード検索

３．学会等名
第45回セマンティックウェブとオントロジー研究会

４．発表年
2018年

１．発表者名
田邉憲太朗，兼岩 憲

２．発表標題
日本語辞書の語義文からのオントロジー自動構築

３．学会等名
電子情報通信学会 人工知能と知識処理研究会（AI）

４．発表年
2019年

１．発表者名 鈴木　諒，兼岩 憲
２．発表標題 日本語文の係り受け木からの意味構造パターンマイニング
３．学会等名 電子情報通信学会 人工知能と知識処理研究会（AI）
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関