公開

研 究 終 了 報 告 書

「オープンデータ利活用のためのデータ検索エンジンの構築」

研究期間: 2018 年 10 月～2022 年 3 月
研 究 者: 加藤 誠

1. 研究のねらい

世界規模の社会問題を解決するために，また，未解決問題を各国の研究者が協調して取り組めるようにするために，オープンサイエンスが世界規模で推進されてきている．オープンサイエンスとは，オープンデータ（調査・研究データの公開）を含む概念であり，データの公開によって研究成果の幅広い活用が図られ，市民の参画（シチズンサイエンス）が促進されることが期待されている．

しかしながら，オープンデータの活用については多くの課題が残されている．国際的動向を踏まえたオープンサイエンスの推進に関する検討会では，日本で運用されている機関リポジトリ数は世界最多であるが，オープンデータの活用が進んでいないと報告されている．多くの機関がデータを公開している点は必ずしも良いことではなく，オープンデータが散逸しており，異分野の研究者や一般市民などの利用者が容易にアクセスできる基盤が整っていない．

そこで、本研究ではデータ検索エンジンを構築し「世界中のオープンデータを整理し，世界中の人々がアクセスできて使えるようにする」ことを目的とする．各国・各機関が公開することで散在するオープンデータに対してインデックスを作成し，オープンデータへのアクセスの一元化を図る．また，研究者や一般市民などの対象利用者に生じる様々な情報要求に対応できるよう，一般的な Web 検索エンジンで用いられるキーワードクエリを指向したインデックスおよび検索モデルを構築する．更に，公開されているデータのメタデータを解析するだけでなく，データ自体を分析することで，より高度な事実検証型クエリ（例：音楽業界が低迷しているか）やデータ質問型クエリ（例：最も犯罪率の多い都市はどこか）なども処理できるようにする．また，各データから得られる可能性のある知見やデータの傾向をデータの要約として検索結果中に示すことにより，利用者のデータ探索およびデータ理解を支援する．これらの実現によって，オープンデータへの容易なアクセスを実現し，より多くの人々が公開されるデータを活用し，研究の推進，新たな価値の創造，データに基づく意思決定ができるようになることが最終的な目標である．

2. 研究成果

(1)概要

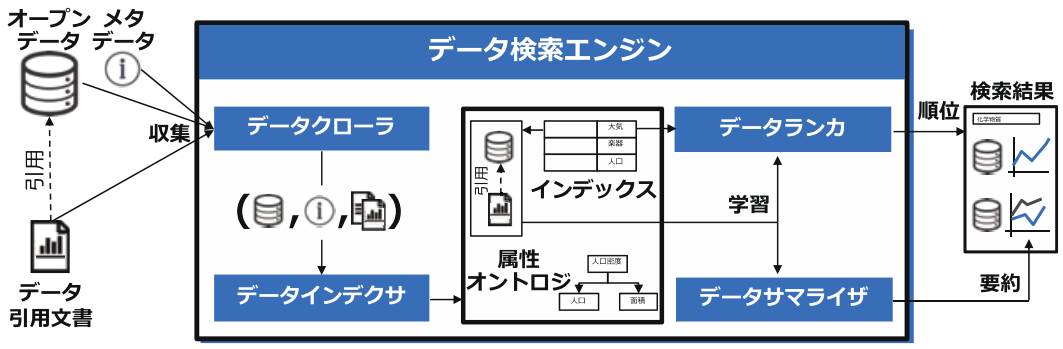

図 1. データ検索エンジンのアーキテクチャ

データ検索エンジンを実現するために，図 1 に示す一般的な検索エンジンを構成する 4 つのコンポーネント，データクローラ，データインデクサ， データランカ， データサマライザ，ごとに研究を行った．

A. データクローラ

A.1. オープンデータ Web サイトの発見： フォーカストクローラを構築しデータ Web 上に点在するオープンデータの効率的な収集方法について提案を行なった．最終的に Web 上にて公開されている 18,903 ドメイン，1,979,149 データを収集し索引付けを行なった．

A.2. データ引用の収集： Web にて公開されているオープンデータを引用する文書を収集し，主に数値がどのように引用されているかを分析，また，引用元が明示されていない場合に自動的に引用元を発見する技術を開発した．

B. データインデクサ

B.1. 属性オントロジの構築： 検索キーワードと属性名のミスマッチを減らすために，オープンデータに含まれる属性間の関係性を見いだし，同一の属性の発見や他の複数の属性の値から計算することのできる属性を特定し属性オントロジを構築した．

B.2. データ構造の理解： データに付与されるメタデータだけでは検索の索引語として不十分であるため，スキーマが明示されていない表形式データから，グラフニューラルネットワークを利用して精度良く見出しを抽出する方法を開発した．

C. データランカ

C.1. キーワードクエリと属性の対応付け： データ検索において入力されるキーワードクエリと収集されたデータの索引語の言語表現が異なるという問題が存在する．この課題に対して，(1) ゼロショット学習，および，(2) 文脈誘導学習，という 2 つのアプローチを提案し評価を行なった．

C.2. データ検索のためのテストコレクション構築： 情報アクセスのための評価キャンペーンである NTCIR において，Data Search タスクを運営し，データ検索のためのデータセットの構築お

よび検索手法の分析を行なった.

D. データサマライザ

D.1. 適切なデータ視覚化方法の推定: 本課題では, データに対して適切な視覚化方法を推定するというタスクに取り組んだ.

D.2. 言語によるデータの説明: 本課題では, 文書中に含まれる表形式データが与えられた時に, 文中からそのデータを説明するのに必要な箇所を特定し, 表形式データの説明文を生成するモデルの提案を行なった.

（2）詳細

A) データクローラ

A.1) オープンデータ Web サイトの発見

オープンデータは異なる組織によって公開されているため, 今後新たに公開されるデータを含め網羅的にオープンデータを収集するためには, データが公開されている Web サイトを効率的に発見できる必要がある. 一般的にこれを実現する方法として, リンクを選択的に辿るフォーカストクローリングが提案されているが, Web の規模に対してオープンデータが公開されている Web サイト数は少なく, ある話題を網羅的に収集する目的で用いられるフォーカストクローリングでは効率的な収集は見込めない.

そこで, 数ステップ先にデータが存在する可能性があるかを推定し, Web グラフ上に点在するデータを効率的に収集するためのクローラの開発に取り組んだ. 具体的には, Deep Q-Network を利用することによって, 長期的な報酬を予測しながら Web グラフを遷移するデータクローラを開発した. 200 万ページから構成されるデータクローリング用のデータセットを構築し, 一般的なフォーカストクローリングタスクで用いられるクローラと比較した結果, Deep Q-Network を利用したクローラは従来のクローラよりも 10 倍以上効率的にデータの収集が可能であることがわかった. 実際に Web 上でクローリングを行ったところ, 本研究期間中に, 18,903 ドメイン, 1,979,149 データを収集することができた.

A.2) データ引用の収集

オープンデータの一部を引用することを, データ引用と呼び, 本研究課題では一般の Web ページでのデータ引用を収集することで, 後述するテストコレクション構築に利用し, また, Wikipedia 中でデータ引用を行なっている箇所を抽出しデータ引用元特定タスクのためのデータセットを構築した.

一般の Web ページでのデータ引用を収集するために, Amazon が提供する Common Crawl データセットのうち, 2019 年に収集された約 100 億 Web ページ（非圧縮で約 800TB）のデータから, e-Stat にて公開される日本語データの引用 137,388 件, および, 国外のデータポータルサイト中のデータに対する引用 3,604,487 件を収集した.

また, Wikipedia 記事の各セクションから オープンデータへのリンクを自動的に抽出し, データ引用を人手によって特定した. このデータセットを用いて, 引用元が与えられたときに, 自動的に引用先を特定する方法について研究を行なった（中野 優, 加藤 誠. クエリと文書のフィー

ルドを考慮した被引用統計データの検索. 情報処理学会論文誌 データベース 14 (4), pp. 49-60, 2021).

B) データインデクサ

B.1) 属性オントロジの構築

検索キーワードと属性名のミスマッチを減らすために, オープンデータに含まれる属性間の関係性を見いだし, 同一属性と上位下位属性を大量に取得する方法を提案しその性能評価を行った. 両関係性の判定には, 属性の取る値間の関係性に基づいた手法を考案した. 同一属性抽出においては, まず既存の表解釈手法を利用して, 表中の各タプルが表すエンティティを特定する. 異なる表中の 2 つの属性において, 同一のエンティティを表すタプルの多くが同じ数値を含むのであれば, それらの属性は同一であると判定する. 上位下位関係の抽出では, まず 1 つの表中の属性の集合から, 上位下位関係が成立している可能性が高い属性の対を抽出する. タプルごとに, すべての下位属性候補の数値の和と上位属性候補の数値が一致すれば, それらに上位下位関係が成立していると判断する.

実験では, Common Crawl にて公開されている 2018 年 10 月から 2019 年 4 月の間に収集された約 200 億件の Web ページを利用し, 数値を含み十分な大きさを持った 5,783,365 個の表を関係抽出に用いた. 結果として, 19,893 個の数値属性の同一性と 8,118 個の数値属性の上位下位関係を発見することができた. 同一性の正解率は 71.4%, 上位下位関係の正解率は 59.2%であった(藤岡 周平, 加藤 誠, 吉川 正俊. 表からの量的データ属性間の関係抽出. 情報処理学会論文誌データベース(TOD) 13(3), pp. 10-21, 2020).

B.2) データ構造の理解

メタデータに記載されていないデータのスキーマを抽出するために, 表形式データの属性を特定しそれらの階層関係を発見する方法について提案および評価を行なった. 我々は, 表形式データを表現するスプレッドシートをグラフデータに変換し, グラフニューラルネットワークを適用することで, スプレッドシートの大域的な特徴を考慮して見出し等を認識する方法を提案した. また, 見出し認識の教師データを用意するためには大きな労力が必要となるため, 教師なし表現学習を行うことによって有効なセル表現を学習し, 少数の教師データであっても効果的な学習が可能となる方法を提案した. 実験では, 見出し階層の認識と見出しの識別を行い, GNN による提案手法の評価を行った. 実験の結果, GNN の教師あり学習手法が, 従来の機械学習による手法よりも見出し階層の認識で高い精度を達成できることを示した(笹治 拓矢, 加藤 誠. グラフニューラルネットワークを用いた表形式データの見出し認識. 電子情報通信学会論文誌 D, Vol.J105-D, No.5, pp.360-371, 2021).

C) データランカ

C.1) キーワードクエリと属性の対応付け

キーワードクエリと属性の対応付けにおいて, 属性名とその属性値を表す言語表現が異なるという問題が存在する. そこでまず, 我々はテキストによって記述される属性を, 関係分類タスクとして捉え, ゼロショット学習によって属性を同定する問題に取り組んだ. ゼロショット学習

を想定したのは，データ検索エンジンが保有するデータに含まれるすべての属性に対して訓練データを用意することは現実的ではないためである．我々は，属性を知識ベースのグラフ構造から得られる埋め込みによって表現し，ディープニューラルネットワークモデルと組み合わせることによって，訓練データのない属性の同定を可能とした．実験では，独自データセット（WEB19）および公開データセット（NYT10）にて，訓練データがある場合とない場合を比較し，訓練データがない属性であっても訓練データのある属性に匹敵する精度で同定できることを明らかにした（代表的な論文（原著論文）発表 (1)）．

また，キーワードクエリと複合的な属性が対応する場合についても研究を行なった．例えば，例えば，「幸福度」というクエリに対しては，「GDP」や「自殺者数」などといった統計データの属性が幸福度を推定，算出する上で有用となりうる．しかしながら，この問題を既存のランキング学習の問題として扱うためには，やはり，訓練データが不足してしまうという問題が起こりうる．この問題に対処するために，我々は文脈誘導型学習と呼ばれる学習方法を提案した．文脈誘導型学習は，パラメータを決定する際に，訓練事例だけでなく属性の文脈を利用する．文脈は追加の情報をモデルに与え，過学習を防止して学習を正しい方向に誘導する．実験では，国，都道府県，カメラの 3 つのクラスを対象とし 158 種類の対応関係を学習した．実験の結果，文脈誘導型学習は既存のランキング学習手法よりも統計的有意に高い精度を示した（Makoto P. Kato, Wiradee Imrattanatrai, Takehiro Yamamoto, Hiroaki Ohshima, Katsumi Tanaka. Context-guided Learning to Rank Entities. Proceedings of the 42nd European Conference on IR Research (ECIR 2020), pp. 83-96, 2020）．

C.2） データ検索のためのテストコレクション構築

情報アクセスのための評価キャンペーンである NTCIR において，Data Search タスクを運営し，データセットの構築および検索手法の分析を行なった．日英の 2 つの言語を対象としたデータセットを構築し，日本語のデータコレクションとしては e-Stat，英語のデータコレクションとしては data.gov をクロールして用いた．テストコレクションには，主に事実検証型クエリ（ある属性値が特定の傾向を示すかどうかを検証するクエリ）を対象とした IR サブタスク，データ質問型クエリ（関係モデルで表現できるデータに対する SQL クエリへ変換可能なクエリ）を対象とした QA サブタスクを評価するためのデータが含まれている（代表的な論文（原著論文）発表 (2)）．

D） データサマライザ

D.1） 適切なデータ視覚化方法の推定

データに対して適切な視覚化方法を推定するというタスクに取り組んだ．このタスクでは，クエリと表形式データが与えられており，両者の関係性から互いのどの部分に着目して要約方法を推定すべきかを決定するような，双方向アテンションを用いたモデルを提案した．提案モデルでは，まずクエリ中の各単語と表データ中の各列の見出し，および，列の統計的な特徴（平均値や分散等）を求め，互いの類似性に基づいて，クエリからどの列にアテンションをかけるべきか，また，列からクエリ中のどの単語にアテンションをかけるべきかを決定する．その後，クエリ中の語の埋め込み，および，列の埋め込みの重み付き平均を取り，その結果を用いて要約方法を推定するモデルとなっている．データセットとしては，Tableau Public にて公開されている表

形式データを利用し，要約表現の予測結果は双方向アテンションを用いたモデルの結果がベースラインを大きく上回り，提案モデルの中でも双方向アテンションを用いたモデルが最も高い性能を示した（代表的な論文（原著論文）発表 (3)）.

D.2) 言語によるデータの説明

文書内の表形式データから，事前学習済み言語モデルを用いて説明を生成する際に，文書中のどの文を追加して説明文生成を行うべきかを検証した．実験では，表をクエリとした文の検索問題を考え，得られた文から論文中に含まれる表から表のキャプションを復元できるかを評価した（Junjie H. Xu, Kohei Shinden, Makoto P. Kato. Table Caption Generation in Scholarly Documents Leveraging Pre-trained Language Models. Proceedings of the 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE 2021), pp. 963-966, 2021）.

本さきがけ研究を通じ，特にデータ検索のテストコレクション構築において，国際的かつ学際的な共同研究が実現した．オーストラリアのオープンデータ検索について研究をしている Ying-Hsang Liu 氏（Oslo Metropolitan University）と，アメリカの大学図書館におけるオープンデータ活用に携わっている Hsin-Liang Chen 氏（Philadelphia College of Osteopathic Medicine）と共同研究を行い，データ検索のテストコレクション構築に関する論文の執筆（代表的な論文（原著論文）発表 (2)），および，Association for Information Science and Technology (ASIS&T)においては 2020 年と 2021 年にデータ検索に関するパネル討論を行なった.

3. 今後の展開

本研究期間中で達成できなかった，実サービスの運用が今後の目標であると考えている．既にデータ自体の収集はできているため，研究トピックとしては扱えなかった，データの分類やタグづけ，フォーマットの異なるメタデータの認識などの実装が行えれば，検索サービスとして運用できると考えている.

4. 自己評価

当初の研究計画において提案した 11 個の研究トピックのうち，8 つについては成果をあげることができたため，研究面においては良好であったと考えている．データクローラの課題「異なるデータ公開方法への対応」については，データ自体を発見することに時間がかかったため取り組むことができず，データランカの課題「事実検証型クエリ」，および，「データ質問型クエリ」については，評価自体に大きな労力を割いたため，技術自体について十分な成果が挙げられなかった.

一方で，社会実装という点については不十分であったと考えている．データ検索エンジンの要素技術については研究が行えていたものの，それらを統合したシステムの構築・運用までには至っていない.

研究実施体制および研究費執行状況については適切であったと考えている．研究期間の後半には，研究補助員を追加して予算計画を一部変更したものの，実装やアノテーションなどの作業が多く必要になる本研究課題においては必要不可欠であったと思われる.

本研究課題の 1 つの成果物であるデータ検索のテストコレクションは，オープンデータの

利用者分析を研究対象とする海外の研究者との国際共同研究の結果として構築できたものである．世界に先駆けてデータ検索の評価用データを構築できた点は評価に値すると思われる．

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：10件

1. Wiradee Imrattanatrai, Makoto P. Kato, Masatoshi Yoshikawa. Identifying Entity Properties from Text with Zero-shot Learning. Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019). 2019. pp. 195-204.
キーワードクエリと属性の対応付けにおいて，属性名とその属性値を表す言語表現が異なるという問題が存在する．そこで，我々はテキストによって記述される属性を，関係分類タスクとして捉え，ゼロショット学習によって属性を同定する問題に取り組んだ．属性を知識ベースのグラフ構造から得られる埋め込みによって表現し，ディープニューラルネットワークモデルと組み合わせることによって，訓練データのない属性の同定を可能とした．
2. Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, Hsin-Liang Chen. A Test Collection for Ad-hoc Dataset Retrieval. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021). 2021. pp. 2450-2456.
日英の 2 つの言語を対象としたデータ検索のためのデータセットを構築し，74 検索システムの性能評価を行なった．このデータセットは，検索対象データとして e-Stat，および，data.gov にて提供されるオープンデータを用い，検索クエリとしては，コミュニティ質問応答サイトからデータ検索に関連した質問を抽出し人手で選定および整形したものを用いている．
3. Atsuki Maruta, Makoto P. Kato. Intent-aware Visualization Recommendation for Tabular Data. Proceedings of the 22nd International Conference on Web Information Systems Engineering (WISE 2021). 2021. pp. 252-266.
本論文では，与えられた検索クエリと表形式データから，クエリの意図に合致した視覚化方法を推定する問題について取り組んだ．特に，視覚化に用いられるグラフの種類と視覚化対象となる列を，クエリ中の各単語と表データ中の各列の見出しの類似性，および，列の統計的な特徴（平均値や分散等）に基づいて推定するモデルを提案した．

（2）特許出願

研究期間全出願件数：0 件（特許公開前のものも含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

1. Makoto P. Kato, Wiradee Imrattanatrai, Takehiro Yamamoto, Hiroaki Ohshima, Katsumi Tanaka. Context-guided Learning to Rank Entities. Proceedings of the 42nd European Conference on IR Research (ECIR 2020), pp. 83-96, 2020.
2. Wiradee Imrattanatrai, Makoto P. Kato, Masatoshi Yoshikawa. Identifying Entity Properties

from Text with Zero-shot Learning. Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019). 2019. pp. 195-204.

3. Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, Hsin-Liang Chen. A Test Collection for Ad-hoc Dataset Retrieval. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021). 2021. pp. 2450-2456.
4. Atsuki Maruta, Makoto P. Kato. Intent-aware Visualization Recommendation for Tabular Data. Proceedings of the 22nd International Conference on Web Information Systems Engineering (WISE 2021). 2021. pp. 252-266.
5. Junjie H. Xu, Kohei Shinden, Makoto P. Kato. Table Caption Generation in Scholarly Documents Leveraging Pre-trained Language Models. Proceedings of the 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE 2021), pp. 963-966, 2021.