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　元 年　 ６ 月　 ７ 日現在

機関番号：１７７０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００４２１

研究課題名（和文）オープンデータ活用のための名前空間の整合性の研究

研究課題名（英文）A study on consistency of name space for utilizing open data

研究代表者
渕田　孝康（Fuchida, Takayasu）

鹿児島大学・理工学域工学系・准教授

研究者番号：７０２５３９１１

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：国や地方自治体が保有するデータを公開し広く社会で活用しようというオープンデータ政策が推進されているが、現実には自治体ごとに公開しているオープンデータのフォーマットや内容が異なることから活用されていない状況が多くみられている。
本研究では国内のオープンデータの連携可能性を評価するため、RDF形式の述語に着目し、オープンデータとして公開されているCSVファイルのデータから述語ベクトルを生成し、オープンデータどうしの類似度を数値で評価する手法を提案した。
全国の約1万のオープンデータを対象に実験し、オープンデータ間の類似度をある程度数値によって表現できることを示した。

研究成果の学術的意義や社会的意義
国の方針としてオープンデータ推進政策がすすめられているが、現状では地方公共団体が公開しているオープンデータが活用されていないのが問題となっている。この問題はオープンデータの公開形式がバラバラで統一性がないことが原因の一つであると考えられる。
本研究ではこの問題を解決する方法として述語ベクトル法と呼ぶ手法を提案した。この手法はオープンデータのデータそのものに着目し、各列のデータが何を表しているかをベクトルの形で表現するものである。本手法を用いてオープンデータ間の類似度を算出し、数値でオープンデータの連携化の姓を示すことが可能となることを示した。

研究成果の概要（英文）：An open data policy is being promoted to disclose data held by national and local governments and make them widely available in society, but in reality it is used because the format and content of open data published to each locality is different There are a lot of situations that are not present.
In this research, in order to evaluate the linkability of open data in this country, focusing on predicates in RDF format, generate predicate vectors from the data of CSV files published as open data, and calculate the similarity between open data Proposed a method to evaluate in
We experimented with about 10,000 open data of the whole country and showed that the similarity between open data can be expressed numerically to some extent.

研究分野：情報工学

キーワード：オープンデータ　述語ベクトル　RDF　CSV　連携可能性　コサイン類似度

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

近年、ビッグデータやオープンデータに代表される様々な形態のデータの、新しい切り口からの活用が注目を集めている。世界的には、2013 年 6 月に開催された G8 サミットにおいて首脳宣言にオープンデータの推進が盛り込まれ「オープンデータ憲章」が採択・合意されている。これを受けて日本政府においても同年 6 月に「世界最先端 IT 国家創造宣言」が閣議決定され、その中でオープンデータ・ビッグデータの活用の推進が盛り込まれている。

日本政府は、2014 年 10 月に行政のオープンデータカタログサイトである data.go.jp の正式運用を開始した。このサイトは日本の省庁をはじめとするさまざまな組織のデータ公開を目的としており、データ以外にも利用者のコミュニケーションや開発者向けの情報なども公開している。このサイトでは 14,000 件を超えるデータが掲載されている。また、同じくオープンデータの活用を促進することを目的として linkdata.org というサイトもオープンしており、データ公開を積極的に進めている地方自治体のデータが多く公開されており、3,500 件を超えるデータが公開中である(2015.10 現在)。

このように、オープンデータの公開と活用は、日本政府が推進する大規模プロジェクトという側面を持つが、その反面、実際の公開と活用は進んでいないのが現状である。
例えば、政府が公開している data.go.jp には 14,000 件を超えるデータが公開されているが、そのフォーマットは、PDF が約 7,300 件、HTML が約 5,600 件、XLS が約 3,000 件、CSV が約 650 件と、データフォーマットの上位はすべて第 1～3 段階のものである。この中で RDF 形式のデータはわずかに 2 件、XML 形式にいたっては 0 件であり、「オープンデータとして公開されている」とは言えない状況である。

またもうひとつのデータ公開サイトである linkdata.org においては、データ自体は RDF 形式で公開されているので第 4 段階ではあるが、名前空間が外部の他のデータとの連携が取れない形で公開されているものがほとんどであり、第 5 段階とは言えない。この点が現在のオープンデータの抱える最大の問題点であり、本研究で取り扱う課題もこの名前空間に関する事項である。また、このサイトでは公開データを用いたアプリの作成もサポートされているが、複数のデータを取り扱う際の名前空間の整合性の保持は人が行っているのが現状である。

また、このような名前空間の統一が図れないもうひとつの原因として、オープンデータを作成して公開する人が、名前空間に関する知識を持っていないことが挙げられる。行政組織等でオープンデータを作成し公開するのは組織の一般職員であり、データ内で使用される言葉に関する専門的な知識を持ち合わせていないケースがほとんどである。このため、組織が持つデータがどの名前空間に属するのかを判断することができず、結果として独自の名前空間を作成してしまうことで、名前空間の統一が図れない状態を作りだすことになる。

このような問題を解決するため、情報処理推進機構(IPA)と経済産業省では共通語彙基盤と呼ぶ、日本語の個々の単語の表記・意味・データ構造を統一し、互いに意味が通じるようにするための仕組みの導入を進めている。2013 年 10 月時点で、コア語彙 2.2 と情報交換パッケージ(IEP)が公開されている。

２．研究の目的

本研究は、オープンデータの公開と活用の推進を目的とし、以下の 3 点を明らかにする。

(1) 現在公開されているオープンデータで使用されている名前空間を調査し、IPA が進める共通語彙基盤に照らして統一化可能な名前空間をリストアップする。それに基づき、すでに公開されたオープンデータの LOD 化が可能となることを明らかにする。
(2) これから公開されるオープンデータを、標準化された名前空間を活用した形で公開するために、専門の知識がない人でも名前空間を適切に選択可能なツールを作成する。実際にこのツールを鹿児島市役所等において使用してもらい、共通化された名前空間によるオープンデータの公開が可能となることを明らかにする。
(3) 名前空間を共通化したオープンデータを機械的に連携することで、これまで知られていなかった新しいデータの創出を行う。このような作業は従来人が発想と閃きを用いて発見的手法で行っていたが、それが機械的に行えることを明らかにする。

３．研究の方法

国及び地方自治体において多くのオープンデータが公開されている。これらのデータは自由にダウンロードして利活用することが認められているものであるため、本研究においても研究リソースとして積極的に利用する。また、オープンデータの中には日本語の文字で表現されているデータも多く含まれている。これらのデータを機械的に処理するために、Word2Vec という技術を利用する。Word2Vec は大量の日本語文章を入力として学習を行うニューラルネットワークの一種であり、単語の前後関係から単語間の類似性を抽出しベクトル化する技術である。この技術を使用し、日本語 Wikipedia から抽出した約 2 億 6000 万単語の文章を学習し、32 万単語を 200 次元ベクトルに変換した。また、述語ベクトル法については、約 300 の軸を使用した。これらの軸は、Judges1（単一の単語を含むかどうか）、Judges2（複数の単語のどれかを含むかどうか）、Judges3（正規表現にマッチするかどうか）の 3 つの条件によって指定し、それらの条件は人間の判断によって決定した。

４．研究成果

本研究においては、主に以下の 2 つについて研究成果が得られた。

(1) 述語提案のための述語ベクトル生成

自治体が持つデータを公開しオープンデータとして活用可能にするためには、その連携について考える必要がある。連携できないデータを公開していても、単独データ活用以上の活用方法につながらず、せっかくのオープンデータを生かすことが出来ないからである。この問題に対応するため、オープンデータのデータそのものに着目し、CSV データの列データをベクトル化する手法である述語ベクトル法を提案した。述語ベクトルとは、図 1 に示すようにオープンデータの CSV ファイルの 1 つのデータに対して「軸」と呼ばれる判定関数を適用して 0,1 ベクトルを生成し、それを平均して得られるベクトルである。軸をうまく選択すれば、各列が表しているデータが何であるかを表現するベクトルが得られる。今回は軸として人為的に選択した約 300 の条件を使用した。また対象オープンデータとしては日本政府が公開している DATA.GO.JP というカタログサイトから全国の 626 個のオープンデータの CSV ファイルを収集して利用した。

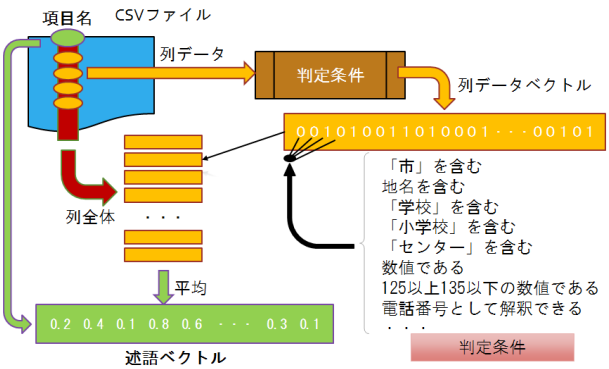

図 1　述語ベクトル法

これらのオープンデータを述語ベクトル法でベクトル化し、ヴォード法を用いた階層的クラスタリングを行った。クラスタ数はエルボー曲線を用いて決定し、視覚化には MDS 法を用いた。得られたカテゴリの例を図 2 に示す。この図より、述語ベクトルを用いることで自動的に意味が近いベクトルを抽出可能であることが分かる。ただし、このデータは人為的に選択された 300 次元の軸を用いてクラスタリングされた結果であり、かなり広範囲なカテゴリが抽出されている。より限定的なカテゴリを抽出するためには、軸の精査が必要になると考えている。

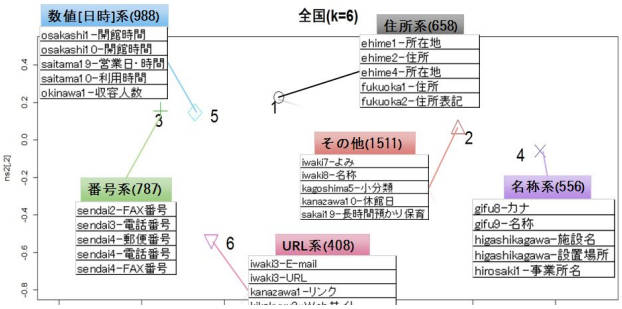

図 2　述語ベクトルのクラスタリング

(2) オープンデータの類似度算出

オープンデータを活用するには、それらのデータがどのくらい連携可能かを示す指標があると便利である。本研究では、全国で公開されているオープンデータについて、述語ベクトル法を用いて類似度を計算した。オープンデータ間の類似度の計算は以下の手順によって行った。まずオープンデータの各列の述語ベクトルを算出する。その際、項目名に当たる単語については今回の述語ベクトルでは使用していない。その後、2 つのオープンデータの列間の類似度を述語ベクトルのコサイン類似度によって計算する。最後に、以下の 3 つの手法でこれらの類似度の重み付き和を計算する。

A　単純平均。2 つのオープンデータのすべての列の組み合わせについて、列間類似度を平均した値

B　ガウス重み平均。2 つのオープンデータのすべての列の組み合わせについて、列間類似度が大きい順にガウス分布にしたがって重みづけて平均した値。ガウス分布の標準偏差は総列数の 1/10 に設定した。

C　しきい値平均。2 つのオープンデータのすべての列の組み合わせについて、しきい値を上回ったものについてだけ平均した値。しきい値は 0.5 とした。

全国の自治体が公開している 633 のオープンデータの CSV に対して、上記の 3 通り計算方法で類似度を計算した。得られた結果で類似度が最も大きかったものを図 3,4,5 に示す。

類似度 A については、新潟市の公園情報と高松

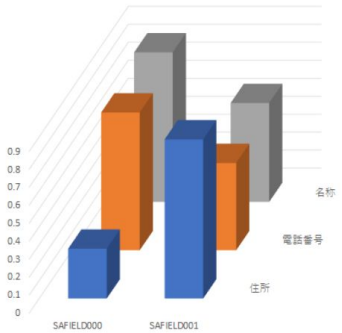

図 3　類似度 A の 1 位

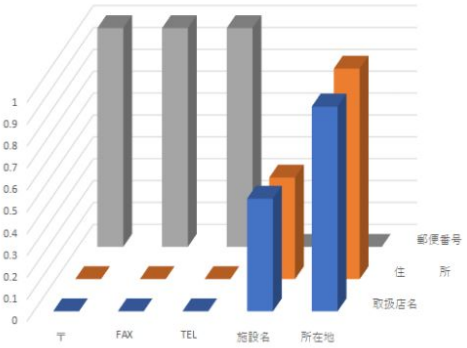

図 4　類似度 B の 1 位

市のコミュニティセンター等の施設情報であった。これら 2 つのオープンデータには施設名、住所、緯度・経度などの情報が含まれており、人間の感覚から見ても良く似たデータである。

類似度 B は、福岡市の小中学校に関するデータと札幌市の商業施設に関するデータであった。これらは郵便番号や住所が含まれるという意味で似ているが、同じ施設情報でもかなり異なるものである。

類似度 C はどちらも千葉市の避難場所に関するデータであり、フォーマットも同じであることから高い値となっている。

これらの結果より、述語ベクトルを用いて列間の類似度を算出し、それを用いてオープンデータ間の類似度を計算することは出来ているが、その類似度にはまだ改善すべき点が多くあると考えられる。以下にいくつかの改善点を挙げる。

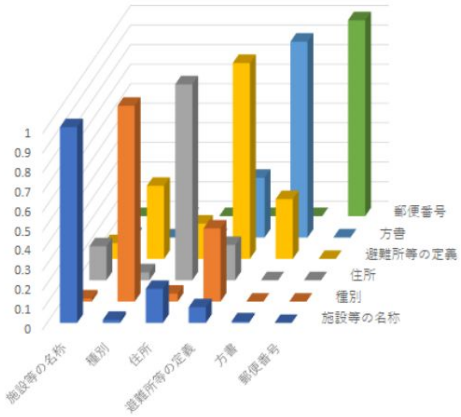

図 5　類似度 C の 1 位

① 数値だけからなるデータ列の処理
② Judges1,2,3 の中の軸の追加
③ 項目名の取り扱い

これらの課題は今後の研究の中で引き続き検証していく予定である。

５．主な発表論文等

〔雑誌論文〕（計 0 件）

〔学会発表〕（計 11 件）

① Bo Chen, Tadanori Hisanaga, Daisuke Noto, Daiki Tomari, Takayasu Fuchida, ”A comparison between Word2Vec and FastText regarding with suggestion of predicates of OpenData”, Proceedings of The Twenty-Fourth International Symposium on Artificial Life and Robotics, pp.227-230, Jan.23-25,2019 (2019.1)

② Tadanori Hisanaga, Bo Chen, Daisuke Noto, Daiki Tomari, Takayasu Fuchida, “Calculation of cooperation possibility of open data using predicate vector method”, Proceedings of The Twenty-Fourth International Symposium on Artificial Life and Robotics, pp.231-234, Jan.23-25,2019 (2019.1)

③ 陳　博・久永忠範・能登大輔・増田　翼・泊　大貴・満園大輔・渕田孝康,“Word2Vec を用いたオープンデータの述語提案手法の研究”, 第 71 回電気･情報関係学会九州支部連合大会論文集 03-1A-05 (2018.9)

④ 能登大輔・久永忠範・陳　博・増田　翼・泊　大貴・満園大輔・渕田孝康,“述語ベクトル法における軸の選定方法の研究”, 第 71 回電気・情報関係学会九州支部連合大会論文集 02-1P-01 (2018.9)

⑤ 久永忠範・陳　博・能登大輔・増田　翼・泊　大貴・満園大輔・渕田孝康,“述語ベクトル法を用いたオープンデータの連携”, 第 71 回電気・情報関係学会九州支部連合大会論文集 02-1P-02 (2018.9)

⑥ Tadanori Hisanaga, Takayasu Fuchida, Bo Chen, “A Proposal of Recommendation Method of Predicates for Open Data Using Statiscital Method”, Proceedings of Asia-Pacific Conference on Engineering and Natural Sciences, pp.107-108, Mar.13-16,2018 (2018.3)

⑦ Bo Chen, Takayasu Fuchida, Tadanori Hisanaga, Chong Guo, Daisuke Noto, “A study on learning method of Word2Vec for recommending predicates of Open Data”, Proceedings of The Twenty-Third International Symposium on Artificial Life and Robotics, pp.173-176, Jan.18-20,2018 (2018.1)

⑧ Tadanori Hisanaga, Takayasu Fuchida, Daisuke Noto, Bo Chen, Chong Guo, “A proposal of recommendation method of predicates for Open Data using statistical method”, Proceedings of The Twenty-Third International Symposium on Artificial Life and Robotics, pp.177-180, Jan.18-20,2018 (2018.1)

⑨ 久永忠範・渕田孝康・能登大輔・陳　博・郭　崇,“オープンデータにおける RDF 変換の研究”, 第 70 回電気・情報関係学会九州支部連合大会論文集 12-2A-07 (2017.9)

⑩ 陳　博・渕田孝康・久永忠範・能登大輔・郭　崇,“自治体オープンデータ向けの単語ベクトルの学習法の研究”, 第 70 回電気・情報関係学会九州支部連合大会論文集 12-2A-08 (2017.9)

⑪ Tadanori Hisanaga, Takayasu Fuchida, Chong Guo and Daisuke Noto, “A proposal of a method for converting into RDF in Open Data”, PROCEEDINGS OF THE TWENTY-SECOND

INTERNATIONAL SYMPOSIUM ON ARTIFICIAL LIFE AND ROBOTICS, pp.196-199, Jan.19-21,2017 (2017.1)

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
なし

６．研究組織

(1)研究分担者
なし

(2)研究協力者
なし