研究論文

述語ベクトルを用いたオープンデータ間の連携度の 算出手法の提案

A proposal of calculation method for degree of cooperation between open data using predicate vector

久永忠範^{1*}, 湊田孝康²

Tadanori HISANAGA^{1*}, Takayasu FUCHIDA²

1 鹿児島大学大学院理工学研究科

Graduate School of Science and Engineering, Kagoshima University

〒890-0065 鹿児島市郡元1-21-40

E-mail: forever@forever.co.jp

2 鹿児島大学大学院理工学研究科

Graduate School of Science and Engineering, Kagoshima University

〒890-0065 鹿児島市郡元1-21-40

E-mail: fuchida@ibe.kagoshima-u.ac.jp

*連絡先著者 久永忠範

近年, 多くの行政・団体がオープンデータの公開, 活用に取り組んでいるが政府の推進する「オープンデータは, 機械判読可能で人手を多くかけずにデータの2次利用が可能である」というデータ活用までには至っていない. 本研究では, 地方公共団体のオープンデータの可能性を探るために, 開示されているCSVデータの項目名, 列データをベクトル化することによりデータの連携度を測る述語ベクトル法を提案する. またその述語ベクトル法を用いて, 各データ間の連携可能性について実験を行い, オープンデータ間における項目名のみの連携度と列データのみの連携度を比較した. 結果として列データのみから算出した連携度が高いオープンデータは特定の列の類似度が高くなり, データ連携を取りやすいことが示された.

In recent years, many administrative organizations and local governments are engaged in opening and utilizing open data but until the data utilization has not reached where the government promotes "Open data can be secondarily used without much manpower being machine-readable". In this research, we propose a predicate vector method that measures the similarity of data by vectorizing item names and item values of CSV data disclosed to explore the possibility of open data of a local government. In addition, using the predicate vector method, we experimented on the possibility of cooperation between

each data, and compared the degree of cooperation of each item name and the degree of cooperation of each column data among open data. As a result, it was shown that open data with a high degree of cooperation calculated from column data has a high degree of similarity of a specific column, making it easy to establish data cooperation.

キーワード: オープンデータ, CSV, ベクトル化, 類似度, 述語ベクトル法, 連携度

Open Data, CSV, Vectorization, Degree of similarity, Predicate vector method, Degree of cooperation

1 はじめに

近年、ビッグデータやオープンデータの活用が推進され、国や地方公共団体をはじめ多くの団体がオープンデータの公開、活用に取り組んでいる。

これらの開示されたデータ形式は、特定のアプリケーションに依存した形式やCSV形式のファイルが未だに多く、2012年7月に策定された「電子行政オープンデータ戦略」[1]において情報の連携・共有の必要性が明示されている。この戦略の意義・目的として

- ・行政の透明性、信頼性の向上
- ・公共データ活用に対しての国民の参加、官民協働の推進
- ・国全体の経済の活性化、国や地方公共団体における行政の効率化

の3つが挙げられている。

またこの戦略においては「オープンデータは機械判読に適したデータ形式で二次利用が可能な利用ルールで公開されたデータである必要があり、それによって人手を多くかけずにデータの二次利用を可能にする」とされている。

また政府は、2014年10月にデータカタログサイト「DATA.GO.JP」[2]の本格運用を開始した。このサイトは、国・地方公共団体・独立行政法人等のもつオープンデータの

データカタログサイトとなっており、各データサイトの名称・組織名・ライセンス・APIの有無・主な分類・概要・更新日等が揭示され、オープンデータの活用を推進している。

このようにオープンデータの施策は、政府や地方公共団体等が推進しているが、より具体的な社会課題の解決や利用促進を行うために2016年5月に「オープンデータ2.0」[3]を情報通信ネットワーク社会推進戦略本部（IT総合戦略本部）が決定した。この「オープンデータ2.0」は、官民一体となったデータ流通を促進し、課題解決のためのオープンデータの「実現」を目指すものである。また2020年までにオープンデータの更なる進化を図る施策として、①政策課題を踏まえた強化分野の設定、②国及び地方公共団体におけるオープンデータの取り組みの推進、③地方公共団体における防災等の地域を跨いだ共通的な分野における取組の推進が掲げられた。特に③においては、政府CIOによる地方公共団体への訪問や、オープンデータ伝道師の制度を活用した地方公共団体への人材派遣、オープンデータパッケージの提供及びオープンデータ100の活用事例の提示のみならず、地方の特性に応じた課題解決に向けた取り組みを支援することで、地方公共団体への普及啓発や利活用に向けた取り組みを推

進することが謳われている。

しかし国の機関、全国の自治体のオープンデータの開示状況、データ形式の調査より、データ連携、機械判読が容易な例は未だに少なく、既存のデータをそのまま掲示してあったり、画像やPDFを張り付けてあるオープンデータも見受けられる状況である。また機械判読可能なCSV形式であっても、項目名としてデータを作成する担当者の主観的表現や、当該行の固有の表現が用いられており、語彙の統一性に欠けている状況も多くみられる。

オープンデータは、機械判読の容易性、著作権等の扱いにより、その開放性の程度が異なっている。これらのオープンデータの形式をTim Berners-Lee[4]が図 1に示す5つの段階とそれに値するデータ形式を提言した。

第1段階から第3段階は、オープンライセンスで提供されていたり、特定のアプリケーション、構造化されたデータとして公開されているデータ形式であり、第4段階から第5段階は、物事の識別にURIを利用したり、他のデータとリンクしているデータ形式である。この段階のRDF形式やLOD形式のデータ作成には、ある程度のITやプログラムの知識が必要であり一般の地方公共団体職員がこの形式のデータを作成することは少々難しいと思われる。

図 1 オープンデータの5つの段階

2018年に総務省が、オープンデータを推進する地方公共団体職員の人材育成のためにオープンデータ研修ポータル[5]を開設した。このサイトはオープンデータの推進に必要な知識・技術等を習得するための各種研修教材等を利用できるポータルサイトである。データ作成の基本からe-learningによる研修も盛り込まれている。このような支援研修により、今後はより多くの地方公共団体職員が機械判読可能なデータ作成を行うことが可能になる。

本研究においては、一般的なデータ形式として開示され易い第3段階におけるCSVデータの連携に着目した。人による視覚的なデータ連携ではなく、機械判読による連携を行うために、本研究では統計処理を用いた述語ベクトル法により、オープンデータ間の連携度の算出方法を提案する。

以降、第2章では、当研究における関連研究の紹介、第3章では提案手法である述語ベクトル法の説明とそれを用いた連携度の計算式を提案し、第4章では提案手法による地方公共団体のオープンデータの連携度の実験と考察、そして第5章では実験結果のまとめと今後の課題について述べる。

2 関連研究

2013年6月にG8サミットで合意された「オープンデータ憲章」[6]により、世界中で政府の持つデータの積極的なオープン化が進展してきている。日本政府もこれに伴い「世界最先端IT国家創造宣言」[7]を閣議決定し、官民にわたる多くの組織がオープンデータを活用し、分野を超えた情報交換を行うために、独立行政法人情報処理推進機構を中心に共通の単語の表記・意味・

データ構造を統一するために「共通語彙基盤」の構築を進めている。ここでは、情報連携に不可欠な基本情報やツールを提供する共通語彙基盤サイトIMI (Infrastructure for Multi-layer Interoperability) [8]を開設している。データ連携に欠かせない共通語彙やIMIを基にしたデータモデル記述(DMD)も提供している。またデータ提供者が利用する機能で、表計算ソフトやCSVなどのデータをDMDに定義された構造化データ(JSON-LD, Turtle, RDF/XML, XML)へ変換するツールも用意されているが、一般の地方公共団体の職員が活用するにはある程度の知識が必要である。

また最近では、地方公共団体のオープンデータの防災におけるデータ活用の研究も進んでおり、特に2011年の東日本大震災や2016年の熊本地震などでは、SNSによる情報発信や情報共有による災害時の対応に多くのデータが活用された。浦田ら[9]は、防災情報のオープンデータ化を推進するために、自治体職員の負担軽減や費用対効果を考え、自治体におけるオープンデータ推進への受容性向上を目的とした研究を進めている。

吉賀ら[10]は、水防情報を構成する河川の水位、雨量、潮位および風向風速情報をLinked Dataとして蓄積し、機械可読な標準的ウェブAPIにより配信提供するシステムについて検討している。これは、国土交通省や都道府県が提供する防災情報や水位情報が個別の表として提供されており、表同士を結び付けるリンクが付されていないものの、河川と水系は別サイトを参照しなければならないなど、閲覧者が直接見ることを想定しており、機械可読なデータとして公開されいないことが原因である。

菱田[11]は、地方公共団体のオープンデータには、同じ内容の項目であっても、いろいろな呼称の項目名が用いられて、データの内容から項目名の意味を推測する必要があるものもあると述べている。

これらの施策や関連研究においては、オープンデータの利活用の推進はなされているが、実際のオープンデータの連携や国の推進する機械判読可能なデータ活用には至っていないことがわかる。

3 提案手法

本章ではオープンデータ間の連携度を計算するために述語ベクトル法を提案し、それを用いて具体的な連携度の算出方法について説明する。述語ベクトルとは、オープンデータの1つの列の内容を数値化したベクトルのことである。述語ベクトルを計算するために、まず全国の地方公共団体が公開しているCSV形式のオープンデータを収集し、あらかじめ定義した約16,000の項目判定関数に基づき述語ベクトルを生成する。項目判定関数については後述する。得られた述語ベクトルからオープンデータの各列間類似度を計算し、オープンデータ間の連携度合いを測る。

3.1 オープンデータの収集

地方公共団体の保有するオープンデータをデータカタログサイト「DATA.GO.JP」に掲載されている359の地方公共団体のリンクサイトから約3万件のオープンデータを収集した。そのデータの内、CSV形式である24,913ファイルを抽出した。

3.2 項目名と列データの関連について

抽出したCSV形式のファイルを調べると、最初の行に項目名ではなく、データの内容の説明書きがあったり、ファイルの作成年

月日等が記載されているものもあった。また各地方公共団体の施設情報を比較してみると、1行目が項目名、2行目から項目値のファイルであっても、地方公共団体によって項目名の表記の仕方が違ったり、記号や英語表記のものも多くみられた。

別表1は、広島市オープンデータの区民文化センターの施設情報である。項目名は、名称、X座標、Y座標、所在地、TEL、FAX、URL、閉館時間、休館日の9つの要素から成り立っている。これに準ずる他の地方公共団体の施設情報の類似した項目名を調べた結果を表1に示す。

表1 広島市の施設情報項目名と他地方公共団体の類似した項目名

広島市施設情報の項目名	類似した項目名
名称	施設名, 建物名, 名前, 物名, 呼称
X座標	緯度, 北緯, 位置
Y座標	経度, 東経, 位置
所在地	住所, 場所名, 位置
TEL	電話, 電話番号, ナンバー
FAX	Fax番号, Faxナンバー
URL	ホームページ, ホームページアドレス, ウェブサイト
開館時間	営業時間, 運営時間, Open
休館日	休日, 休み, Closed

「名称」という項目名に対しても「施設名」「建物名」「名前」「物名」「呼称」等、数多くの表現が用いられていることがわかった。したがって、各オープンデータの列の項目名だけをみても、各列間の類似度を比較することは難しく、項目名よりは項目名の下に位置する列データを比較することが類似度を測る尺度になるのではな

いかと考えられる。

3.3 述語ベクトルの生成

述語ベクトルを生成するために、あるCSVファイルの項目名以下にある列データひとつひとつに判定条件を照らし合わせ、それと一致すれば1を、一致しなければ0を返し、その合計数を列データ数で割り平均を取る項目判定関数を準備する。表2にCSVファイルにおける項目名と列データを示す。

表2 CSVファイルにおける項目名と列データ

csvファイル					
項目名	施設区分	施設名	郵便番号	住所	電話
列データ	病院	香椎丘リハビリ	813-0002	福岡県福岡市	662-3200
	病院	雁の巣病院	811-0206	福岡県福岡市	606-2861
	病院	香椎療養所	813-0011	福岡県福岡市	661-1083
	病院	香椎原病院	813-0011	福岡県福岡市	662-1333
	病院	疋田病院	813-0011	福岡県福岡市	681-3111
	病院	医療法人相	813-0017	福岡県福岡市	662-3001
	病院	杉岡記念病	813-0017	福岡県福岡市	662-3535

この関数から得られた数値がその項目名の述語ベクトルとなる。図2に述語ベクトルの生成方法を視覚的に示す。

また述語ベクトルの生成と同時に、ある項目名のすべての列データが数値、または空のデータである場合はその項目名は述語ベクトルを生成しないようにしている。なぜならそのような列データから取得さ

図2 述語ベクトルの生成方法

れる値が数量なのか、あるいは面積なのか何のデータなのか把握するのは難しいからである。今回の実験では不明な数値のみのデータは除外している。電話番号やFAX番号の- (ハイフン) が含まれているものや緯度・経度等の一般的な表現形式に限り、正規表現を用いて判定されるようにしている。

次にすべての判定条件にヒットせず、述語ベクトルがすべて0である項目名の除去を行う。最後に項目名と述語ベクトルに判定条件を追加したものをCSVファイルとして出力して生成完了である。

3.3.1 行数に応じた重み付け

オープンデータの各データの行数を調べてみると、少ないものは1行から多いものは数千行に及ぶものもあった。述語ベクトルは行数の平均なので、行数の少ないものは1つの行の影響が大きくなるので、述語ベクトルの各要素を行数に応じて式 (1) に示すシグモイド関数で重み付けする。ゲイン a の値は、行数の中央値が約38であったことから、38行程度のデータの場合に1になるように $a = 0.2$ とした。図 3に $a = 0.2$ の場合のグラフを示す。

$$f(x) = \frac{2}{1 + e^{-ax}} + 1 \quad (1)$$

ここで、 x はオープンデータの行数である。

3.4 項目判定関数

述語ベクトルの生成における各項目値を判定するものを項目判定関数と呼ぶ。項目判定関数は、先に抽出した地方公共団体の13,000ファイルの中の語彙を抽出して、MeCab (IPA辞書) を用いて分かち書きを行い、その分かち書きを行った333,220個の語彙から頻度30個以上の語彙15,934個の単語を抽出した。

図 3 行数に応じた重み付けしたシグモイド関数

3.4.1 項目判定方法

項目の判定方法には、Judges1, 2, 3の3つの判定方法を用いた。

Judges1は、該当する単語を含むか含まないかの判断を行う。例えば「施設」と言う単語を含む場合は1を返し、含まない場合は0を返す。このような単語を15,934個準備し、各々を述語ベクトルの軸とした。

Judges2は、あるグループ内の単語を一つでも含むかどうかの判断を行う。例えば「教育施設」という軸に対して「保育園、幼稚園、こども園、小学校、中学校、高校、大学」のいずれかを含めば1を返し、含まなければ0を返す。

Judges3は、正規表現を用いて語彙の種類、性質に応じて判断を行う。例えば経度、緯度を判断するためには以下のような正規表現を用いた。

```
^(1(2[2-9]|[34][0-9]|5[0-3]))¥..*$
^([23][0-9]|4[0-5])¥..*$
```

これは、日本の経度は122度～153度の範囲に入っており、緯度は20度～45度の範囲に入っていると判断している。

述語ベクトル法の具体的な例を別表2に示す。例えば、項目の列に施設名称のデー

タが入っていた場合に判定条件となる単語「鹿児島, 駅, 病院, 公園, 市立, 県立, 小学校, 中学校, 高校」に合致していれば1を返し, 合致していなければ0を返すことになる. その項目列の平均に前述した式(1)の重みを乗じた値が述語ベクトルとなる.

3.5 列間類似度の計算

すべてのオープンデータについて各列の述語ベクトルを計算できたら, それを使ってすべての列の組み合わせに対して列間の類似度を計算する. オープンデータ*i*の第*k*列の述語ベクトルを v_{ik} として, 以下の式で列間類似度 s_{ijkl} を計算した.

$$s_{ijkl} = v_{ik} \cdot v_{jl}$$

ここで「 $\cdot$ 」はベクトルの内積である. 列間類似度をコサイン類似度によって計算する方法も考えられるが, 述語ベクトルでは, 各要素の大きさは項目データの一致個数に比例するのでベクトルの大きさには意味がある. このため大きさを1に正規化する方法は適さないと判断し, 内積を使用することとした.

3.6 オープンデータ間の連携度の計算

オープンデータ間の連携度とは, 類似度とは異なり, どれだけ似ているかの尺度ではない. 2つのオープンデータを連携させてアプリケーションを作成することを考える場合, それら2つのオープンデータのすべての列が似ている必要はなく, いくつかの特定の列に類似度があればよい. 単純な例では, 2つのオープンデータがどちらも緯度と経度の列を持っていれば, それらを両方とも地図上に表示するアプリを作成することが可能となる.

したがって, 2つのオープンデータの連携度を計算するために, すべての列の類似度の平均を用いるのは適切ではなく, 何ら

かの重みを付けて計算する必要があると考えられる.

3.6.1 連携度の計算式

2つのオープンデータ*i, j*間の連携度 $C(i, j)$ を, それぞれのオープンデータの列間類似度の順位に応じて正規分布で重みを付けた類似度の平均として式(2)で定義する.

$$C(i, j) = \frac{\sum_{k=1}^{n_i} \sum_{l=1}^{n_j} G(o_{kl}) s_{ijkl}}{\sum_{k=1}^{n_i} \sum_{l=1}^{n_j} G(o_{kl})} \quad (2)$$

ここで, *i, j*はオープンデータの番号, s_{ijkl} はオープンデータ*i*の*k*列とオープンデータ*j*の*l*列との類似度, n_i, n_j はそれぞれオープンデータ*i, j*の列数, o_{kl} は2つのオープンデータの列間類似度の順位であり, 最も類似度が高いものを0番, 次を1番の順で数える. $G(o_{kl})$ は平均0の正規分布であり, 標準偏差 σ は以下の式(3)で計算した.

$$\sigma = 0.1n_i n_j \quad (3)$$

係数0.1は以下のようにして決定した. もし全く同じ列内容を持つ2つのオープンデータop1とop2が存在する場合, それらの列間類似度は対応する列の部分が大きく, それ以外の列間では小さくなると考えられる. 具体的な例として, 鹿児島市のAED設置場所(市)と鹿児島市のAED設置場所(民間)との列間類似度を図4に示す.

ここでは, 4つの列が存在しているが, 同じ列どうしの類似度は大きくなり, 異なる列間の類似度は小さくなっている. このように, 列数が*n*のオープンデータが連携するとき, 列間類似度の大きい列が*n*個程度存在すると考えられ, その割合は $\frac{n}{n^2} = \frac{1}{n}$ 程度になると考えられる.

今回の実験では, 列数として2以上20以下を使用しているため, $n = 10$ として計算し, 0.1という値を採用した.

図 4 鹿児島市AED設置場所（市）と
鹿児島市AED設置場所（民間）

4 実験と考察

本章では、提案したオープンデータ間の連携度の有効性を検証するため、実際に地方公共団体が公開しているオープンデータを用いて連携度を計算する実験を行った。比較のために、列データではなく項目名だけを対象とした述語ベクトルによる実験も行い、結果を考察した。以下では、列データを用いた類似度を列データ間類似度、項目名を用いた類似度を項目名列間類似度と呼び、両方を意味する場合は単に列間類似度と呼ぶ。

4.1 実験データの準備

政府のポータルサイトDATA.GO.JPから142の地方公共団体のオープンデータ25,000ファイルを抽出した。そのオープンデータの各地方公共団体の占める割合は、新潟市が2,392ファイル、横浜市が2,302ファイル、徳島市が1,945ファイルとオープンデータファイルの多い都市もあるが、越前

市、光市、喜多方市、大津市、新宿区、草加市、天理市、敦賀市、宇部市、調布市、愛媛県など1つのCSVファイルしかない県・都市もあった。そこでファイル数の多い都市に偏りがないようにランダムに各都市10ファイル以内で300のファイルを抽出した。

抽出した300のファイルのデータの内容がどのような分野に属するかを調べるために、各データの内容のチェックを行い、施設、防災、統計、環境、経済、行政、交通、医療福祉、文化、行事、生活、観光の12の分野に分類をおこなった。またそれぞれの分類でも別表3に示すように他に分類の詳細な内容がわかるように付帯情報も追記した。

4.2 実験の流れ

オープンデータ間の連携度を算出する流れは以下のとおりである。

- ① オープンデータのすべての列についての述語ベクトルを計算する。
- ② 2つのオープンデータのすべての列の組み合わせに対して列間類似度を計算する。
- ③ 列間類似度を用いてすべてのオープンデータ間の連携度を計算する。

ただし、今回は2列以上の列を持つデータを対象とし、列数が多いと計算にかなりの時間を要するので、列数が2以上20以下のファイルを対象とし、44,850ペアの連携度の結果を得た（別表4）。

4.3 列データ間類似度による連携度

列データ間の類似度をもとに各オープンデータ間の連携度の上位1位から7位と1000位の列データ間類似度の内容を調べた。それらをピボットテーブルを用いてグラフ化した。

図 5 品川区の防災情報と江戸川区の防災情報

4.3.1 連携度の第1位(連携度5.91)

第1位は品川区の防災情報（防災無線設置場所）と江戸川区の防災情報（防災無線設置場所）となり，防災情報が一致している（図 5）．項目名においては品川区の種別，緯度，経度に対して，江戸川区の名称，緯度，経度に対応している．

品川区の項目名：「種別」の項目値は，「防災行政無線固定系」であり，江戸川区の項目名：「名称」の項目値は，「行政防災無線」と内容が似通っていることがわかる．

4.3.2 連携度の第2位(連携度4.21)

第2位は半田市の統計情報（小学校の世帯数）と三条市の施設情報（小学校）となった（図 6）．各々の項目名は，半田市が小学校区名称，年月日であり，三条市は，address, category, latitude, longitude, name, phone_number, URLであった．半田市の項目名：「小学校区名称」対して，三条市の項目名「category」「name」の列データ間類似度が高いのは，どちらにも「小学校」という

図 6 半田市の統計情報と三条市の施設情報

単語が含まれていた．また半田市の項目名「年月日」と三条市の項目名「phone_number」の列データ間類似度が高い．これは半田市の「年月日」の列データ値が2017-04-30の様にハイフンで3つに区切ってあることで電話番号と同じ形式に

図 7 宇都市の施設情報と府中市の施設情報

なっていることが原因である。

4.3.3 連携度の第3位(連携度3.87)

第3位は宇部市の施設情報（ふれあいセンター）と府中市の施設情報（介護施設）となった（図 7）。宇部市の項目名は、サマリ、タイトル、メモタイトル1、メモ内容1、大字名、番地2で、府中市の項目名は、名称、郵便番号、所在地、電話番号、FAXである。府中市の「名称」に対して、宇部市のサマリ、タイトルの列データ間類似度が高いのは、どちらもセンターという語彙が多く含まれていた。また府中市の所在地と宇部市のメモ内容1、大字名の列データ間類似度が高いのは、メモ内容と大字名に住所が含まれていることが要因と考えられる。

4.3.4 連携度の第4位(連携度3.74)

江戸川区1の施設情報（子育てひろば）と江戸川区2の施設情報（公園場所）となった（図 8）。江戸川区1の項目名は、名称、電話番号、所在地、緯度、経度、URL、FAXであり、江戸川区2の項目名は、名称、所在地、写真1、写真、緯度、経度、URLである。江

一」という単語が多く合致して列データ間類似度が高くなっていることがわかる。

4.3.6 連携度の第6位(連携度3.71)

永平寺町1の施設情報(教育機関)と永平寺町2の施設情報(避難所)となった(図10)。永平寺町1は、項目名が記載されておらず1行目が列データ値になっていて、永平寺町2の項目名は名称、住所、経度、緯度、ふりがなである。列データ間類似度を見ると永平寺町1の緯度、経度の数値と永平寺町2の「緯度」、「経度」の項目名が合致している。また永平寺町1のふりがなで表示してある学校名が、永平寺町2の「ふりがな」の項目名とも合致している。

図10 永平寺町1の施設情報と永平寺町2の施設情報

4.3.7 連携度の第7位(連携度3.62)

日進市の施設情報(小学校)と三条市の施設情報(小学校)となった(図11)。日進市の項目名は、連絡先名称、名称-カナ、名称、電話番号、住所表記、種別、経度、緯度、Webサイトであり、三条市の項目名は、address, category, latitude, longitude, name, phone_number, URLである。三条市のname

とcategoryに対して、日進市の種別、名称、連絡先名称の項目間類似度が高いのは、「小学校」「中学校」「高校」や学校名、地名がお互いに含まれているのが要因だと考えられる。

図11 日進市の施設情報と三条市の施設情報

4.3.8 連携度の第1000位(連携度3.54)

東久留米市の防災情報(避難所)と千葉市の施設情報(防災情報)となった(図12)。東久留米市の項目名は、電話番号、中項目、大項目、所在地、施設名、経度、緯度であり、千葉市の項目名は、緯度(世界測地系)、

図12 東久留米市の防災情報と千葉市の施設情報

経度（世界測地系）、施設・場所・イベントの名称、住所、説明文である。東久留米市の大項目と千葉市の場所・イベントの名称、説明文との項目間類似度が高いのはお互いに防災、防犯等の共通の語彙が含まれているからである。

4.3.9 列データ間類似度を用いた実験のまとめ

列データ間類似度を用いて計算したオープンデータの連携度においては、片方のオープンデータの一つの列について、もう片方のオープンデータの1ないし数個の列に高い類似度があるという特徴がみられる。これは、2つのオープンデータを連携させるときに有用な特徴であると考えられる。

また、項目名の単語が合致しなくても、列データ値が類似していれば類似度が高い結果が得られる。

一方で、数値のみの列のデータについては、電話番号、緯度・経度等のように正規分布で判読されるもの以外は、類似度を測ることが難しいという問題もある。

4.4 項目名間類似度による連携度

項目名間類似度も列データ間類似度の計算の同様の300のオープンデータ間の各々の項目名間類似度を約16,000次元の軸を用いて連携度の計算を行った。ここでも列データ間類似度と同様に列数が2以上20以下のファイルの計算を行い、44,850ペアの類似度の結果を得た（別表5）。

項目名列間類似度の上位1位から4位の1000位のオープンデータのペアの項目名同士の類似度を調べてピボットテーブルを用いて、項目名列間の類似度をグラフ化した。

4.4.1 連携度の第1位（連携度7.64）

防府市1の行政情報（26年と27年度予算額）と防府市2の行政情報（27年と28年度予算額）となった（図 13）。同じ自治体の掲載年度だけが違う同じ内容のオープンデータであり、項目名もすべて合致していることがわかる。

図 13 防府市1の行政情報と防府市2の行政情報

4.4.2 連携度の第2位（連携度7.35）

川崎市の施設情報（美容室）と奈良市の施設情報（旅館）となった（図 14）。お互いのオープンデータの内容は、行政による施設の許認可のデータである。項目名に共通の単語（営業、所在地、氏名、名、番号など）が多く含まれているので項目名列間類似度が高い結果になっている。

図 14 川崎市の施設情報と奈良市の施設情報

4.4.3 連携度の第3位(連携度6.83)

三重県1の行政情報と三重県2の行政情報となった(図 15). 三重県1は、営業所の許可の施設情報であり、三重県2は営業所の屋号の施設情報で、項目名に営業、氏名、番号と言う語彙が多く含まれることで

図 15 三重県1の行政情報と三重県2の行政情報

項目名列間類似度が高くなっている。

4.4.4 連携度の第4位(連携度6.32)

川崎市の施設情報(美容室)と三重県の行政情報(業種別認可)となった(図 16). 項目名同士の共通の語彙(営業、所、名、業)などの多いので項目名列間類似度が高くなっている。

図 16 川崎市の施設情報と三重県の行政情報

4.4.5 連携度の第1000位(連携度2.40)

川崎市の施設情報(美容室)と志摩市の防災情報(避難種別施設)でとなった(図 17).

項目名の各単語意味はお互いに全く違うものであるが、部分的に同じ漢字や数字が使われているので類似度が高いものがある。

4.4.6 項目名間類似度を用いた実験のまとめ

項目名間類似度を用いて計算した連携度においては、片方のオープンデータの一つの列に対して、もう片方のオープンデー

図 17 川崎市の施設情報と志摩市の防災情報

タの多く列の類似度が高くなる傾向が見られた。これは、項目名として同じ単語が複数の列に使用されているケースが多くあることが原因である。例えば、連携度第2位の川崎市の施設情報（美容室）と奈良市の施設情報（旅館）の場合は、「営業～」という項目名が多く表れた結果、ほとんどの列に対して類似度が高くなっている。このような場合は、2つのオープンデータを連携させることが難しいと考えられる。

一方、データが数字のみの場合でも、項目名がそのデータの意味を表していれば、正しく連携されるという特徴もある。ただし、現状で公開されているオープンデータの項目名は列の内容を正しく表していないケースも多く、項目名に頼って内容を判断するだけではよい結果は得られなかった。

4.5 考察

列データ間類似度を用いて計算した連携度（以下連携度I）と、項目名間類似度を用いて計算した連携度（以下連携度II）にどのような違いがあるかについて考察する。

連携度Iの第1位の組み合わせである品川区の防災情報（防災無線）と江戸川区の防災情報（無線設置場所）は、連携度IIでは第748位となった。連携度Iのグラフを図 5に、連携度IIのグラフを図 18に示す。これらを比較すると、緯度・経度についてはいずれも類似度が高い結果となっているが、種別と名称の間の類似度が、連携度Iでは高いが連携度IIでは0になっている。これは、連携度IIでは項目名の「種別」と「名称」に間に共通する単語がないため、類似度を算出できていないためである。一方、連携度Iでは、「種別」と「名称」の各データの中身を見て、それが類似していることを正しく示すことができる。

連携度Iの第2位である三条市の施設情報（小学校）ローマ字表記と半田市の統計情報（小学校世帯数）は、連携度IIでは第27088位となった。この場合の連携度Iのグラフを

図 18 連携度IIの第748位

図 19 連携度IIの第27088位

図 6に、連携度IIのグラフを図 19に示す。これらのオープンデータはいずれも小学校に関するデータであるが、片方の項目名は日本語であり、もう片方は英語となっているため、連携度IIの類似度は全く一致せず、すべて0になっている。一方、連携度Iの類似度は列データの中身を見て判断しているため、それらが似たデータであることを正しく判定できていることがわかる。

5 まとめと今後の課題

最後に、本研究の内容を総括し、今後の課題について述べる。

5.1 まとめ

本研究では、全国の地方公共団体のオープンデータを用いてCSVファイルの列データから連携度の高いファイルを抽出する手法を提案した。政府の推進するオープンデータの利活用を行うためには、データ間の連携が重要であり、かつ大量のデータから連携を図るためには、機械判読で連携可

能なデータを抽出できることが重要である。

各地方公共団体のオープンデータのCSVファイルの内容をみると、人間の目で見やすい形式にしてあるものや、項目名と項目値だけでなく、メタデータとしてファイルの説明文が1行目にあるもの、さらに1列目に項目名があるものもあった。また各地方公共団体によって同じ意味であっても項目名として異なった単語を用いられていることが多く、統一されていないことがわかった。したがってデータの連携を図るために項目名だけを利用したのでは不十分であり、列データ同士に共通する単語をチェックすることが必要である。本研究ではこれらの課題を解決するために述語ベクトル法を用いて列間類似度を計算し、さらにこれを用いてオープンデータの連携度を算出した。

連携度が大きければ特定の列の類似度が高いものが含まれているが、必ずしも連携がとれるとは限らない。なぜなら現状のオープンデータのCSVファイルの中には項目名が縦に書いてあるなど、フォーマットが不適切なものも含まれているからである。そこで、連携度の値について以下に考察してみる。

もし同じような内容を表す二つのデータがあった場合、それらの対応する列同士の類似度が高くなり、それ以外の列同士の類似度は低くなる。今、代表的なデータとして列数が10である二つのデータを考えると、列の組み合わせの総数は100であり、そのうち類似度の高い列の数は10となる。このときの二つのデータの連携度は、式(2)より、

$$C(i, j) = \frac{\sum_{k=0}^9 \sum_{l=0}^9 G(o_{kl}) s_{ikjl}}{\sum_{k=0}^9 \sum_{l=0}^9 G(o_{kl})}$$

となる。ここで、正規分布の標準偏差は式(3)より10.0となるので、高い類似度を1.0、低い類似度を0.0と仮定すると、

$$\begin{aligned}
C(i,j) &= \frac{G(0) + G(1) + \dots + G(9) + 0 + \dots + 0}{G(0) + G(1) + \dots + G(99)} \\
&= \frac{0.071 + 0.070 + \dots + 0.047}{0.071 + 0.070 + \dots + 3.69 \times 10^{-23}} \\
&\approx 0.67
\end{aligned}$$

となる。したがって、今回の実験においては連携度が0.67程度以上が、二つのオープンデータの連携の目安となると考えられる。連携度0.67となる二つのオープンデータは9205位であった。

図 20 オープンデータの順位による分類の一致数

図 20に今回、実験に使用した300個のオープンデータの連携度の順位と、人間が二つのデータの中身を見て判断した分類の一致数のグラフを示す。分類は施設、防災、統計など12種類を使用した。別表3にその一部を示す。

このグラフを見ると、連携度による順位と分類との間には相関があり、順位が10000前後で一致数の傾きに変化がみられることがわかる。10000位の前後で分類の一致数が増えていることから、前述した連携度0.67程度がデータ連携の目安になるのではないかと考えられる。

連携度Ⅰの結果より、2つのオープンデータ間の列データの類似度の特徴は、項目名が違って列データの内容が似通っていれば類似度が高くなり、それらの列を共通化することにより、それらのオープンデータの連携が可能になる。一方で連携度Ⅱの結果より、項目名だけの述語ベクトルを用いた場合は、項目名の似通ったもの同士が抽出されてデータの連携にはあまり適していないことがわかった。

5.2 今後の課題

述語ベクトル法における項目データの判定に項目判定関数を利用するが、今回の実験では約16,000個の軸を利用した。軸が多いと述語ベクトルの計算に時間がかかるようになる。今回は、ランダムに抽出した地方公共団体の300ファイルの44,850ペアの連携度を計算するのに約30時間を要したので、オープンデータ間の連携度を高めつつ、この項目判定関数の軸の精査し、削減を行う必要がある。

今後はオープンデータ間のより良い連携を図るために、類似度や連携度の計算式に用いるパラメータの調整も必要である。

列間類似度の計算を行う前に項目名や列データの分かち書きをすれば、列間類似度の精度が高まるものと思われる。

今回の実験では、列データが数値のみの場合は、郵便番号、電話番号、緯度・経度のデータしか計算をしなかったが、項目名と数値を照らし合わせて列間の類似度を計算できる仕組みを考えたい。

またオープンデータ間の連携度の高いものについては、そのデータを利用したアプリケーションの作成も視野にしていきたい。

謝辞

本研究はJSPS科研費 JP16K00421の助成を受けたものです。

参考文献

- [1] 電子行政オープンデータ戦略
(2012) 高度情報通信ネットワーク社会
推進戦略本部
https://www.kantei.go.jp/jp/singi/it2/pdf/120704_siryu2.pdf (2020/01/07)
- [2] DATA.GO.JPサイト (2014) 内閣官房
情報通信技術 (IT) 総合戦略室
<http://www.data.go.jp/> (2020/01/07)
- [3] オープンデータ2.0 (2016) 官民一体
となったデータ流通の促進～課題解決の
ためのオープンデータの「実現」～高度
情報通信ネットワーク社会推進戦略本部
https://www.kantei.go.jp/jp/singi/it2/densi/opendata2/data_sokushin.pdf (2020/01/07)
- [4] Tim Berners-Lee : Open Dataのための5
つ段階 (2012) EC FP7 Support Action
LOD-Around-The-Clock (LATC)
<http://5stardata.info/ja/> (2020/01/07)
- [5] オープンデータ研修ポータル
(2018) 総務省情報流通行政局地域通信
振興課地方情報化推進室

<https://www.opendata-training.org/>
(2020/01/07)

- [6] オープンデータ憲章 (2013) 外務省
<https://www.kantei.go.jp/jp/singi/it2/densi/dai4/sankou8.pdf> (2020/01/07)
- [7] 世界最先端IT国家創造宣言 (2013)
高度情報通信ネットワーク社会推進戦略
本部
<https://www.kantei.go.jp/jp/singi/it2/kettei/pdf/20130614/siryu5.pdf> (2020/01/07)
- [8] 共通語彙基盤サイトIMI (2013) 独
立行政法人情報処理推進機構 (IPA)
<https://imi.go.jp/goi/> (2020/01/07)
- [9] 浦田真由, 荻島和真, 中條裕基, 遠
藤守, 安田孝美 : 地域防災情報における
自治体オープンデータ推進の実践 社会
情報学会 2018年7巻1号 p.1-17
- [10] 吉賀 夏子, 堀 良彰, 牛島 清豪 : 防
災のためのLinked Dataを用いた河川水位情
報の蓄積配信システムの検討 2019年度
電気・情報関係学会九州支部連合大会発
表論文誌 11-2P-03 p.591
- [11] 菱田隆彰 : (公財) 電気通信普及財
団 研究調査助成報告書 No.31 2016
(2019年10月18日 受付)
(2020年 1月16日 採択)
(2020年 2月14日 J-STAGE早期公開)

別表1 広島市のオープンデータ＞文化施設＞区民文化センター

名称	X座標	Y座標	所在地	TEL	FAX	URL	開館時間	休館日
広島市まちづくり市民交流プラザ	132.4582215	34.39130456	広島市中区袋町6番36号	082-545-3911	082-545-3838	http://www.cf.city	9時30分～22時	月の第3月曜日、年末年始
アステールプラザ (文化創造)	132.4480662	34.38781836	広島市中区加古町4番17号	082-244-8000	082-246-5808	http://www.cf.city	9時～21時	年末年始 (12月29日～)
文化創造センター	132.4480662	34.38781836	広島市中区加古町4番17号	082-244-8000	082-246-5808	http://www.cf.city	9時～21時	年末年始 (12月29日～)
中区民文化センター	132.4480662	34.38781836	広島市中区加古町4番17号	082-244-8000	082-246-5808	http://www.cf.city	9時～21時	月曜日 (祝日及び8月)
国際青年会館	132.4480662	34.38781836	広島市中区加古町4番17号	082-244-8700	082-246-5808	http://hiyh.pr.aren	9時～21時	年末年始 (12月29日～)
東区民文化センター	132.4831885	34.39568779	広島市東区東屋町10番31号	082-264-5551	082-264-5774	http://www.cf.city	9時～21時	月曜日 (祝日及び8月)
南区民文化センター	132.468576	34.38102384	広島市南区比治山本町16番27号	082-251-4120	082-256-8811	http://www.cf.city	9時～21時	月曜日 (祝日及び8月)
西区民文化センター	132.4486264	34.40876247	広島市西区横川新町6番1号	082-234-1960	082-293-1860	http://www.cf.city	9時～21時	月曜日 (祝日及び8月)
安佐南区民文化センター	132.4744285	34.45171263	広島市安佐南区中筋一丁目22番17号	082-879-3060	082-879-3061	http://www.cf.city	9時～21時	月曜日 (祝日及び8月)
安佐北区民文化センター	132.5077138	34.5260475	広島市安佐北区可部七丁目28番25号	082-814-0370	082-814-0770	http://www.cf.city	9時～21時	月曜日 (祝日及び8月)
安芸区民文化センター	132.5266214	34.37217711	広島市安芸区船越南三丁目2番16号	082-824-1330	082-824-1337	http://www.cf.city	9時～21時	月曜日 (祝日及び8月)
佐伯区民文化センター	132.3595789	34.37964654	広島市佐伯区五日市中央六丁目1番10号	082-921-7550	082-921-9898	http://www.cf.city	9時～21時	月曜日 (祝日及び8月)

別表2 述語ベクトルの具体的な例

項目名	施設名	項目判定関数（軸）								
		鹿児島	駅	病院	公園	市立	県立	小学校	中学校	高校
項目値	鹿児島市立病院	1	0	1	0	1	0	0	0	0
	鹿児島中央駅	1	1	0	0	0	0	0	0	0
	天文館公園	0	0	0	1	0	0	0	0	0
	鹿児島県立吉野公園	1	0	0	1	0	1	0	0	0
	鹿児島市立山下小学校	1	0	0	0	1	0	1	0	0
	維新ふるさと館	0	0	0	0	0	0	0	0	0
	鹿児島市立健康の森公園	1	0	0	1	1	0	0	0	0
	鹿児島県立鹿児島中央高校	1	0	0	0	0	1	0	0	1
	鹿児島駅	1	1	0	0	0	0	0	0	0
	平田公園	0	0	0	1	0	0	0	0	0
述語ベクトル		0.7	0.2	0.1	0.4	0.3	0.2	0.1	0	0.1

別表3 オープンデータの分類

オープンデータファイル名	オープンデータの分類（付帯情報）
www.city.tosu.lg.jp/Material/52444.csv	鳥栖市の統計情報（年齢別人口）
www.city.matsuyama.ehime.jp/shisei/opendata/metadata/syu_kizitsu.files/h26syusen2_ki	松山市の行政情報（投票者数）
www.city.hofu.yamaguchi.jp/uploaded/attachment/64440.csv	防府市の行政情報（予算額）
www.city.kagoshima.lg.jp/jousys/documents/4-2_jisinnjinotaihibasyo.csv	鹿児島市の施設情報（公園）
www.city.nagaoka.niigata.jp/shisei/cate10/toukei/file/toukei_27.csv	長岡市の行政情報（予算額）
www.city.maebashi.gunma.jp/sisei/499/509/p012146_d/fil/siminservicecenter.csv	前橋市の施設情報（サービスセンター）
opendata.pref.miyazaki.lg.jp/dataset/908/resource/2805/08_ibento.csv	宮崎県の行事情報（担当課別）
www.city.niigata.lg.jp/shisei/seisaku/it/open-data/opendata-kenkoiryo/od-kenkohigai.files	新潟市の福祉情報（検診受診者数）
www.pref.gunma.jp/contents/000363787.csv	群馬県の行政情報（適合施設）
www.city.nisshin.lg.jp/dbps_data/_material/_files/000/000/021/926/23230_kyouikukikan.	日進市の施設情報（小学校）
www.city.nara.lg.jp/www/contents/1147846435189/files/bunkan_csv.csv	奈良市の施設情報（公民館分館）
www.city.yamato.lg.jp/web/content/000122731.csv	大和市の行政情報（犬の登録等）
www.city.yamagata-yamagata.lg.jp/kakuka/kankyo/gomigen/sogo/gazoufile/gomishusekijo	山形市の環境情報（ごみ収集場所）
www.city.sumida.lg.jp/kuseijoho/sumida_info/opendata/opendata_ichiran/gyoseikisosiroyo	墨田区の統計情報（教育機関障害者数）
www.city.aizuwakamatsu.fukushima.jp/docs/2007091400823/files/odei0823.csv	会津若松市の環境情報（浄水場調査）
www.city.ikoma.lg.jp/cmsfiles/contents/0000013/13459/292095_population_20180401.csv	生駒市の統計情報（人口）
www.city.utsunomiya.tochigi.jp/_res/projects/default_project/_page_/001/010/185/syokuty	宇都宮市の福祉情報（食中毒）
www.city.toyohashi.lg.jp/secure/16259/21daichou.csv	豊橋市の統計情報（年月日別世帯人口）
www.city.shima.mie.jp/ikkrwebBrowse/material/files/group/2/2016ago_gomicalendar.csv	志摩市の環境情報（ごみ処理）
www.city.sumida.lg.jp/kuseijoho/sumida_info/opendata/opendata_ichiran/gyoseikisosiroyo	墨田区の統計情報（施設利用数）

別表4 列データを用いたオープンデータ間連携度の上位20位まで

順位	オープンデータ1	オープンデータ2	連携度
1	品川区の防災情報（防災無線）	江戸川区の防災情報（無線設置場所）	5.911650
2	三条市の施設情報（小学校）	半田市の統計情報（小学世帯数）	4.207914
3	宇部市の施設情報（ふれあいセンター）	府中市の施設情報（介護施設）	3.868272
4	江戸川区の施設情報（子育てひろば）	江戸川区の施設情報（公園場所）	3.736977
5	墨田区の統計情報（施設利用数）	宇部市の施設情報（ふれあいセンター）	3.734778
6	永平寺町の施設情報（教育機関）	永平寺町の施設情報（避難所）	3.712775
7	日進市の施設情報（小学校）	三条市の施設情報（小学校）	3.622607
8	日進市の施設情報（小学校）	半田市の統計情報（小学世帯数）	3.537817
9	江戸川区の施設情報（小学校住所）	江戸川区の施設情報（公園場所）	3.454595
10	板橋区の施設情報（地域センター）	宇部市の施設情報（ふれあいセンター）	3.446024
11	豊川市の施設情報（公園）	豊川市の施設情報（公園）	3.426981
12	日進市の施設情報（小学校）	秦野市の生活情報（小学校情報）	3.342528
13	江戸川区の施設情報（子育てひろば）	江戸川区の施設情報（小学校住所）	3.337789
14	日進市の施設情報（小学校）	長久手市の施設情報（教育機関）	3.289944
15	三条市の施設情報（小学校）	秦野市の生活情報（小学校情報）	3.241530
16	宇部市の施設情報（ふれあいセンター）	福島県の施設情報（子育て支援センター）	3.192532
17	宇部市の施設情報（ふれあいセンター）	浦添市の環境情報（ごみ収集日）	3.164680
18	板橋区の施設情報（地域センター）	府中市の施設情報（介護施設）	3.157339
19	綾瀬市の行政情報（各施設紹介）	坂井市の観光情報（施設紹介）	3.151521
20	半田市の統計情報（小学世帯数）	笠間市の統計情報（小学校学年別人数）	3.131830

別表5 項目名のみを用いたオープンデータ間連携度の上位20位まで

順位	オープンデータ1	オープンデータ2	連携度
1	防府市の行政情報（予算額）	防府市の行政情報（予算額）	7.644626
2	川崎市の施設情報（美容室）	奈良市の施設情報（旅館）	7.346210
3	三重県の行政情報（業種別認可）	三重県の行政情報（業種別認可）	6.829889
4	川崎市の施設情報（美容室）	三重県の行政情報（業種別認可）	6.316383
5	川崎市の施設情報（美容室）	三重県の行政情報（業種別認可）	6.245795
6	川崎市の施設情報（美容室）	奈良市の行政情報（食品店認可）	6.001161
7	さいたま市の統計情報（地区別人口）	さいたま市の統計情報（地区別人口）	5.855432
8	長岡市の行政情報（投票者数）	知多市の行政情報（選挙投票者数内訳）	5.815247
9	川崎市の施設情報（美容室）	宇都宮市の施設情報（クリーニング店）	5.710058
10	奈良市の行政情報（食品店認可）	奈良市の施設情報（旅館）	5.437359
11	三重県の行政情報（業種別認可）	奈良市の施設情報（旅館）	5.324943
12	三重県の行政情報（業種別認可）	奈良市の施設情報（旅館）	5.252509
13	奈良市の施設情報（クリーニング店）	奈良市の施設情報（旅館）	5.171285
14	徳島市の統計情報（産業別雇用者数）	函館市の統計情報（労働者雇用）	5.098786
15	福島県の行政情報（議会情報）	福島県の行政情報（三役）	5.047107
16	千葉市の施設情報（イベント名称）	旭川市の施設情報（介護施設）	4.990488
17	千葉市の施設情報（イベント名称）	豊田市の施設情報（こども園）	4.950444
18	松山市の行政情報（投票者数）	知多市の行政情報（選挙投票者数内訳）	4.945077
19	奈良市の行政情報（食品店認可）	三重県の行政情報（業種別認可）	4.751153
20	川崎市の施設情報（美容室）	奈良市の施設情報（クリーニング店）	4.726888