１　はじめに

陳博\*　泊大貴\*　程芳 \* 渕田孝康\*

（鹿児島大学　\*大学院理工学研究科)

オープンデータの述語サジェストのための項目名のクラスタリングの研究

A study on the clustering of item names for predicate suggestion of Open Data

近年，新型コロナウイルスにより，台湾のオープンデータを活用したマスク在庫管理マップアプリが話題となり，世界的にオープンデータへの関心が高まりつつある。日本でもオープンデータの活用が推進され，国や地方自治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。鹿児島市でも2016年7月にオープンデータ[1]がCSV等形式で公開されているが，このようなデータ形式・フォーマットの違いにより開示されても積極的な活用まで至っていない。

本研究では，RDFに焦点を当てて，適切な述語をサジェストするために，公開したオープンデータの項目名を抽出し，項目名の単語をWord2Vec[２] によって50次元のベクトルで表現し，それらをを使用し，既存の階層的クラスタリングを行い，その実験結果を示す。Word2Vecは単一の単語をベクトル化するが,複合語は単語ベクトル空間で表現できる．例えば, 図1の単語ベクトルの計算イメージに示すように「鹿児島」と「県」の和で「鹿児島県」で表現できる．また，新しいクラスタリング手法を提案し，既存の実験結果を比較する。

２　外部状況

現在RDFとして公開されているデータも名前空間の共通化についてほとんどなされていない。最近，様々な方法で名前空間が提案され，それらの名前空間の中で定義されているクラスやプロパティを述語として利用されている。しかし，実際に市民や民間企業側にはそのような形の名前空間理解し，述語を使ってRDFを作成するのは非常に困難である。また，全国のデータを複数利用し,データの整備を整えるシステムが開発されていないのが現状である[3]。

３　実験・考察

* 1. 実験１

階層的クラスタリングの流れとして下記の手順である。

1. 自治体が公開しているオープンデータのcsvファイルを収集する。
2. 項目データと項目名を抽出し,分かち書きする．日本語の分かち書きとは文を語に分解し,それぞれをスペースによって区切る書き方である。Janome[4]というPythonの形態素解析エンジンを使用する。
3. Word2Vecで項目データと項目名のベクトルを取得し,図３に示すようにベクトルを合成する。
4. 項目名のベクトルの階層的グラスタリグを行う。 図2に階層的クラスタリングによって得られたデンドログラムを示す。

クラスタ数として,「50」を設定した時クラスタリングした結果について解析した。一部のクラスタ内の類似度の分布は図３に示す。異なるクラスタ間の類似度の分布は図4に示す。図３と図４の分布図より，同じクラスタ内でも低い類似度のものがある一方で,異なるクラスタ間でも高い類似度のものもあることが分かった。原因として，クラスタ数を「５０」に設定して，強制的に合ってないクラスタに入れられてしまっている可能性があると考えられる。

図1 ベクトル空間で単語の合成ベクトル

図2階層的クラスタリングによって得られたデンドログラム

図３クラスタ内のcos類似度の分布

図4クラスタ間のcos類似度の分布

図5 項目名のベクトルは次元を落として可視化

図6 クラスタ間の類似度算出結果

* 1. 実験2

実験１の結果より,階層的クラスタリングではうまくいかなかったことがわかった。項目名のベクトルは次元を落として可視化は図5に示す。高次元空間における50次元の場合は異常に広い空間であり，ほとんどの部分には点がない状態にあるという特徴がある。そのため，新しいクラスタリング手法を提案する。流れとして下記の手順である。

1. 類似度の閾値sh(0.6,0.7,0.8,0.9)を適当に決める。
2. すべての項目名について,他の項目名との類似度を計算し,sh以上のものをクラスタ候補とする。
3. すべての項目名のクラスタ候補の中で重なっている率を計算する。クラスタ間の類似度を計算し，どの程度単語の分類のされ方に違いところを定量的に求めた。クラスタに含まれる単語数をn，ほかのクラスタに含まれる単語数をm，クラスタの間に含まれる共通の単語数をsとすると,類似度rは，

で表す。式により，二つのクラスタ間に共通の単語が一つもない場合は類似度が0となり，クラスタ間の単語が

図7 クラスタ内のcos類似度の分布

図8 クラスタ間のcos類似度の分布

すべて一致した場合は類似度が1となる。

1. クラスタの重なり率を表にする。SH値=0.7の時クラスタ間の重なり率表は図6に示す。図6に重なり率が範囲[0.9,1]となる場合は赤色となり,[0.8,0.9]となる場合は黄色となり,[0.7,0.8] となる場合は緑色となる。

提案手法ではクラスタ数の設定が不要となり, 一部のクラスタ内の類似度の分布は図7に示す。クラスタ間の類似度の分布は図8に示す。階層的グラスタリグを比較し,クラスタに含まれる項目名が減っているが,分類したクラスタ数が増えている。

４　まとめ

オープンデータのCSVファイルの項目名を抽出し，Word2Vecで得られた単語ベクトルを用い，クラスタリングを行った。さらに，クラスタリングの条件を変えることで，生成されるクラスタの適切さをDCSEで定量的に評価した。また，クラスタの適切さにほかの評価手法も検討していく。今後，割り当てられたカテゴリを教師信号として活用し，Deep Learningで学習し，述語のサジェストを行い，RDF形式に変換していく方法を検討する。

参考文献

1. 鹿児島市オープンデータhttps://www.city.kagoshima.lg.jp/jousys/opendata.html
2. 西尾泰和，word2vec による自然言語処理，株式会社 オライリー・ジャパン，pp.32-41，2014
3. 著者 庄司 昌彦 “国内における活用環境整備”, 情報処理学会論文誌, vol.54, no-12, pp. 1244 - 1247
4. Janome

<https://github.com/mocobeta/janome>