Vol.2019-IS-149 No.7
2019/8/23

分散表現を用いた政府オープンデータにおけるタグの階層構造の構築

長谷川 誠[1,a)]　山田 泰寛[2,b)]

概要：近年，政府や地方自治体は保有する統計データなどをオープンデータとして公開している．公開の際に，オープンデータに対するメタデータの一つとして，内容を表わす語であるタグが付与される．タグはオープンデータの検索の際に役立つが，タグの付与や検索時のタグの選択は，適切に行なわれるとは限らない．本稿は，オープンデータに対してタグの階層構造を構築することで，タグの付与や選択の補助を行なうことを目的とする．タグの階層構造により，あるタグの上位もしくは下位の意味にあたるタグを知ることできる．提案手法は，入力として与えられるオープンデータの集合から，それぞれのタグの分散表現を獲得する．次に，1 つのオープンデータに共起している 2 つのタグの類似度を計算する．その類似度が，ある閾値の範囲にある時，2 つのタグは関連があるとし，頻度の大きいタグを上位のタグとする．実験では，日本政府のオープンデータカタログサイトである DATA.GO.JP から収集したオープンデータに出現するタグを用いて，階層構造の構築を行った結果について報告する．

キーワード：オープンデータ，メタデータ，タグ，分散表現

1. はじめに

近年，政府や地方自治体は，それぞれの Web サイト上に，保有する統計データなどをライセンスフリーの形態で積極的に公開している．このようなデータは，「オープンデータ」と呼ばれる．オープンデータは，民間や企業にとって有益な情報であり，オープンデータを利用したアプリケーションの制作が行われている．

オープンデータを公開する際には，そのデータの作成者や作成日，タグ，概要などを示すメタデータが付与される．**図 1** は，日本政府のオープンデータカタログサイトである DATA.GO.JP[*1] で公開されている 1 件のオープンデータのメタデータであり，タイトルや概要，タグなどが記載されている．本稿では，メタデータの中でタグに着目する．タグとは，データやテキスト，画像，動画の内容を表す語であり，一つの対象に対して，複数のタグが付与されることが多い．オープンデータにタグを付与する利点として，オープンデータのファイルの中身を見なくても，タグを見ることで，ある程度内容を把握することができ，ユーザにとって必要なデータか不必要なデータかを判断することができる．また，オープンデータを検索する際に，オープンデータの全文を検索対象にするのではなく，重要な語として選択されたタグを利用することで，質の良い検索を行うことができる．

しかし，タグを付与するためには，オープンデータに対して専門的な知識が必要となる．付与者は，オープンデータを理解した上で，適切なタグを付与しなければならない．また，タグ付与の目的である，ユーザがタグを見ることで，オープンデータの内容を大まかに想像できるためには，一つのオープンデータに対して，複数のタグが付与されることが望ましい．付与者が，オープンデータに対して，ある一つのタグを思い浮かべるとき，そのタグに関連するタグを推薦する仕組みがあれば，タグ付与の補助を行うことができる．

次に，オープンデータの検索について考える．オープンデータを検索する際，一つの検索語を入力した検索結果では，必要なオープンデータが含まれているとは限らない．また，検索結果として表示されたオープンデータのリストから，必要なオープンデータを探すことに時間を必要とする．そのため，他の検索語を入力したり，複数の検索語を

[1] 島根大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Shimane University
[2] 島根大学学術研究院理工学系
Institute of Science and Engineering, Academic Assembly, Shimane University
[a)] n18m109@matsu.shimane-u.ac.jp
[b)] yamada@cis.shimane-u.ac.jp
[*1] https://www.data.go.jp

フィールド	値
タイトル	子ども・若者白書_平成26年版
説明	内閣府が発行している子ども・若者白書の平成26年版
公表組織名	内閣府
連絡先	政策統括官（共生社会政策担当）
作成者	政策統括官（共生社会政策担当）
タグ	2000012010019,education,labor,white_paper and ann...,いじめ,ニート,体力,健全育成,労働,地域,子ども,子供,学力,学校,家庭,少年法,引きこもり,教育,文化,白書_年次報告,白書_年次報告書等,社会,職業,若者,青少年,非行防止
リリース日	2014-06-03
作成頻度	1年
公開ウェブページ	https://www8.cao.go.jp/youth/whitepaper/h26honpen/index.html
対象地域	

図 1 オープンデータのメタデータ（https://www.data.go.jp/data/dataset/cao_20140904_0304）

入力し AND 検索を利用して，必要なオープンデータを探していく．もし，ユーザが入力する検索語に関連する検索語を知ることができれば，検索の補助となる．また，政府や地方自治体のオープンデータでは専門的な用語が用いられることもあるため，専門的な用語が推薦されれば，ユーザーにとって必要なオープンデータを絞りやすくなる．

本稿では，タグの付与や検索の補助のために，オープンデータに対するタグの階層構造の構築手法を提案する．タグの階層構造は，上位にあるタグほど一般的な意味を表し，下位あるタグほど専門的な意味を表す．タグの付与においては，付与者が，最初に思い付いたタグに対して，タグの階層構造から，そのタグの上位または下位に当たるタグを把握することができる．これによって，上位と下位のタグから必要なタグを選択することができる．

オープンデータの検索においても，ユーザが検索を行う際に，タグの階層構造を利用することで，入力する検索語に関連するタグを知ることができる．例えば，ある語が入力された際，タグの階層構造によってその語の下位語を検索候補として提案することで，必要となるオープンデータを絞り込むことができる．また，検索語より一般的な意味の単語で検索したい場合，その語の上位語を見ることで，検索の補助をすることができる．

提案するタグの階層構造の構築手法は，入力として与えられたタグ集合から，それぞれのタグの分散表現を word2vec を用いて獲得する．次に，一つのオープンデータに共起している二つのタグの類似度をコサイン類似度によって計算する．その類似度が，あらかじめ決めたコサイン類似度の範囲にあるとき，二つのタグは関連があるとし，頻度の大きいタグを上位のタグとする．

本稿では，DATA.GO.JP から収集した 18,675 件のオープンデータに対して，提案手法によってタグの階層構造の構築を行った．実験では，二つのタグの類似度を測るコサイン類似度について，二つの異なるコサイン類似度の範囲を設定し，タグの階層構造をそれぞれ構築した．コサイン類似度の範囲が 0.7〜1.0 のとき，同データに付与されるタグ集合の中で，同義語となるタグが同階層ではなく，上位または下位のタグとして抽出された．コサイン類似度の範囲が 0.4〜0.7 のときは，範囲 0.7〜1.0 に比べて同義語となるタグが同階層に位置することが多く，階層も深くなった．しかし，コサイン類似度の範囲を下げたことで，上位と下位のタグとして意味の繋がりのないものが抽出された．

本稿は，以下のように構成されている．2 節では，関連研究について述べる．3 節で，タグの階層構造構築手法について提案し，4 節で，提案手法に関する実験について述べる．最後に，5 節において，まとめと今後の課題について述べる．

2. 関連研究

山田らは，機械学習を用いて，政府オープンデータに対して付与すべきタグの推定に関する実験をしている [1, 2]．あらかじめ，オープンデータに付与されたタグを学習しておき，タグが付与されていないオープンデータに対してタグの推定を行った．しかし，推定されたタグはそれぞれ個別のものであり，タグの間に同義の関係や上位下位の関係があるかまでは判断できない．

タグの共起数と頻度から階層構造を作り出すための手法が Chen らによって提案されている [3]．階層構造のパスに沿ったノードの意味が一貫していないという問題と，木構造の一部で非常に深いパスや多くの分岐が発生する問題に対して，Chen らは，ノードの兄弟間に多様性を持たせ，子ノードが均等になるようなアルゴリズムを提案している．しかし，ここで提案されたタグの階層構造は，各タグの上位タグが一つに限定される．政府オープンデータのタグを考えたとき，あるタグの上位にあたるタグは一つであるとは限らない．

分散表現を利用して，タグの上位下位関係を抽出する方

表 1　word2vec の学習データとオプション

学習コーパス	Wikipedia
単語の種類数	554,897
単語の出現数	576,253,636
次元数	300
ウィンドウサイズ	8

法が鈴木ら [4] によって提案されている．用意された is-a 関係のデータセットを word2vec でベクトル化し，それぞれの is-a 関係を示す差ベクトルを獲得する．それを学習データとして，X-means 法によりクラスタリングし，上位下位関係の判別機を作成する．この判別機を用いて，テストデータに対して上位語の予測を行う．しかし，この手法では，あらかじめ is-a 関係のデータセットを用意しておく必要がある．本稿で対象とする政府オープンデータのタグでは，上位や下位の関係は定義されていないため，鈴木らの手法を用いることはできない．

3. 分散表現を用いたタグの階層構造構築手法

本稿では，日本政府のデータカタログサイト DATA.GO.JP 上で公開されているデータセットを扱う．実験では，このサイトで公開されている 18,675 件のデータセットを利用した．

3.1 分散表現

分散表現とは，単語を高次元のベクトルで表現することで，本稿では，word2vec を用いて単語をベクトル化した．word2vec は，隠れ層と出力層の 2 層からなるニューラルネットワークで，これを用いて単語の分散表現を学習する．学習方法には，CBoW(Countinues Bag-of-Words) モデル [5] と Skip-gram モデル [5,6] の 2 種類がある．実験では，後者の Skip-gram モデルを用いて単語の分散表現を学習した．学習に用いたデータ及びオプションを，**表 1** に示す．学習コーパスは政府オープンデータのメタデータもしくはオープンデータのファイルを用いるべきと考える．しかし，メタデータでは学習コーパスとしてデータが小さすぎることと，全てのファイルを取得するには時間と手間を必要とすることから，今回は Wikipedia を学習コーパスとした．

3.2 タグの階層構造構築手法

本節では，日本政府のデータカタログサイトである DATA.GO.JP におけるタグの階層構造構築手法について提案する．

一つのオープンデータに付与されているタグ集合は，そのデータに関する内容で重要となる語が付与されることから，同データに付与されているタグは，同義語や上位・下位語の関係があると仮定した．提案手法は，各オープンデータに付与されているタグ集合から上位・下位の関係を抽出し，全てのデータにおける上位・下位タグの関係から一つのタグの階層構造を構築する．タグの階層構造は，上位にあるタグほど一般的な意味を表し，下位あるタグほど専門的な意味を表す．

Algorithm 1 タグの階層構造構築手法

Input: $T_1, T_2, \ldots, T_n$ (ただし，T_i は 1 オープンデータ中のタグ集合), α, β (ただし，$-1 \leq \alpha \leq \beta \leq 1$)
Output: 階層構造 (有向グラフ) $G = (V, A)$

$V = \bigcup_i T_i$
for $i = 1$ to n **do**
  word2vec を用いて，T_i 中の全てのタグをベクトル化
  for T_i 中の全てのタグの組み合わせ t_{i_a}, t_{i_b} に対して **do**
    if t_{i_a} と t_{i_b} のコサイン類似度が α 以上かつ β 以下 **then**
      if $freq(t_{i_a}) > freq(t_{i_b})$ **then**
        弧 (t_{i_a}, t_{i_b}) を A に追加
      else if $freq(t_{i_a}) < freq(t_{i_b})$ **then**
        弧 (t_{i_b}, t_{i_a}) を A に追加
      else
        if $length(t_{i_a}) < length(t_{i_b})$ **then**
          弧 (t_{i_a}, t_{i_b}) を A に追加
        else if $length(t_{i_a}) > length(t_{i_b})$ **then**
          弧 (t_{i_b}, t_{i_a}) を A に追加
        else
          弧 (t_{i_a}, t_{i_b}) を A に追加
        end if
      end if
    end if
  end for
end for
V 中で，弧を持たないノード (タグ) を V から削除
弧 $(t_a, t_b) \in A$ について，他に t_a から t_b への歩道が存在するならば，(t_a, t_b) を A から削除

タグの階層構造構築手法を Algorithm 1 に記載する．提案手法は，入力として，n 個のオープンデータのタグ集合 $T_1, T_2, \ldots, T_n$ と，コサイン類似度の閾値 α, β(ただし，$-1 \leq \alpha \leq \beta \leq 1$) が与えられる．$T_i$ は，i 番目のオープンデータにおけるタグ集合を表わし，$T_i = \{t_{i_1}, t_{i_2}, \ldots, t_{i_k}\}(t_{i_j}$ はタグ) である．出力は，有向グラフ G で，V がタグ集合，A は二つのタグを繋ぐ弧の集合である．$freq(t)$ は，タグ t の $T_1, T_2, \ldots, T_n$ における頻度を表わし，$length(t)$ は，タグ t の文字数を表わす．

初めに，入力として与えれらた $T_1, T_2, \ldots, T_n$ 中の全てのタグを V に追加する．それぞれの $T_i(1 \leq i \leq n)$ に対して，T_i 中の全てのタグをベクトル化する．本稿では，3.1 節で述べた word2vec によってタグの分散表現を獲得した．しかし，DATA.GO.JP に含まれるタグは，単語だけでなく，複合語も含まれている．本稿では，word2vec は形態素解析された単語を学習しているため，複合語から分散表現を獲得することができない．複合語であるタグの分散表現は，複合語を構成する名詞の平均ベクトルとして計算した．

次に，T_i 中のタグに上位下位関係が存在するかを調べ，上位下位関係があると判断した場合，頻度と文字数に従って

表 2　階層構造の比較

コサイン類似度の範囲	ノード数	弧の数	深さの最大
(i) 0.4-0.7	2,123	2,313	55
(ii) 0.7-1.0	670	504	14

二つのタグを弧で結ぶ．二つのタグ t_{i_a} と t_{i_b} は word2vec によりベクトルとして表現されており，二つのベクトルの類似度をコサイン類似度によって計算する．その類似度が α 以上 β 以下である時，二つのタグの間には，上位下位関係があると判断する．二つのタグのコサイン類似度は，二つのタグの近さを表わしており，1 に近いほど類似している．

二つのタグのコサイン類似度が α 以上かつ β 以下であったとき，全オープンデータ中で，頻度の高いタグを上位のタグとし，A に弧を追加する．多くのオープンデータに付与されているタグの方が一般的な意味を持っていると仮定し，頻度の高いタグを上位のタグとした．二つのタグの頻度が同じであった場合は，文字数が小さいタグを上位のタグとする．DATA.GO.JP においてデータセットに付与されるタグを確認したところ，「産業」と「福祉産業」のように，片方のタグ t_{i_a} の文字列が，もう一つのタグ t_{i_b} に含まれることが見られた．このような場合，t_{i_a} は上位タグであり，t_{i_b} は下位タグである傾向があった．

最後に，V から弧を持たないノード (タグ) を V から削除する．また，弧 $(t_a, t_b) \in A$ について，他に t_a から t_b への歩道が存在するならば，(t_a, t_b) を A から削除する．

4. 実験

実験では，二つのタグの類似度を測るコサイン類似度の範囲を 2 種類設定し，タグの階層構造をそれぞれ構築して，比較と考察を行う．コサイン類似度の範囲は，(i)0.4〜0.7 と (ii)0.7〜1.0 に設定した．範囲 (i) については，コサイン類似度の範囲が高すぎると上位下位の関係よりも，類似した単語や同義語に弧が引かれると考え，中程度類似しているタグを想定して設定した．範囲 (ii) については，より類似しているものに弧を引くことを想定して設定した．

使用した実験データは，2017 年 3 月 21 日に収集した DATA.GO.JP における 18,675 件のデータセットを用いた．データセットには，日本語のタグに対応する英語のタグがいくつか存在するが，それらは除外した．提案手法では，word2vec を用いてタグの分散表現を獲得する．しかし，タグを構成する単語全てが学習コーパスに存在しない場合は，タグをベクトル化できないため，そのタグについては除外した．また，タグの形態素解析後，タグは名詞で構成されることが多いため，名詞以外の単語を除外して実験を行なった．

4.1 実験結果

表 2 は，コサイン類似度の範囲を (i)0.4〜0.7 と (ii)0.7〜1.0 に設定し，ノード数と弧の数，最大の深さを比較した結果である．タグの階層構造におけるノード数は，(i) が (ii) より約 3 倍多く，深さは約 4 倍深かった．(i) で構築したタグの階層構造を，**図 2** に示す．

(i) と (ii) の階層構造を比較すると，(ii) は (i) と比べて，階層構造におけるタグが少なく，一つのタグに対して一つの下位タグのみ持つものが多く見られた．そのため，(ii) の階層構造は，タグの付与や検索の観点からは，ユーザーが思い付いた一つのタグを選択しても，そのタグに関連する少数のタグ候補しか挙げることができない．

次に，(i) で構築されたタグの階層構造について，部分的な階層構造を抽出し考察する．DATA.GO.JP のオープンデータに付与されるタグ集合には，上位・下位の意味にあるタグだけでなく，同義語のタグが存在する．例えば，「平成 x 年度」や「補正 x 号」というタグがある (x には具体的な整数が入る)．(i) のタグの階層構造の部分木である**図 3** では，タグ「平成 x 年度」がタグ「旅客県間流動調査」の下位タグとして同階層に出ている．図 3 に出現する「平成 x 年度」の他にも「平成 x 年度」は存在するが，「旅客県間流動調査」と同じオープンデータに付与されていないため，弧で結ばれなかった．また，「補正 x 号」についても，「平成 x 年度」と同じような結果になることを確認した．これは，(i) は類似度が高い単語には弧が結ばれない上，「平成 x 年度」タグや「補正 x 号」タグなどに対して，「旅客県間流動調査」のようなタグが同じオープンデータに付与されているためだと考えられる．

DATA.GO.JP に付与されているタグの中で，他に同階層に位置すべきタグとして，国名がある．しかし，(i) の階層構造では，**図 4** の結果が出力された．上位から下位にかけて「英国」「ドイツ」「フランス」「オランダ」「アイルランド」と複数の国名のタグが繋がっている．これらの国名の中で，二つの国名のコサイン類似度は，(i) の範囲にあり，かつ，これらの国名は同データに付与されているため，弧で繋がっている．この結果は「平成 x 年度」や「補正 x 号」のタグとは違い，文字列として類似していないが，国名という意味で類似しているため，(i) において弧で繋がっていると考えられる．

図 2 より，(i) では，ノード数が 2,123，エッジ数が 2,313，深さが 55 の階層構造を構築している．そのため，(i) の階層構造は，タグの付与や検索の観点から見ると，ユーザが最初に思いついたタグについて，タグの階層構造から周辺のタグを選択するのに十分なタグを提示することができる．

図 5 から**図 7** は，図 2 の一部を取り出したものである．図 5 より，(i) はコサイン類似度の範囲を高く設定していないため「外食」と「美容」という意味が類似しないタグ

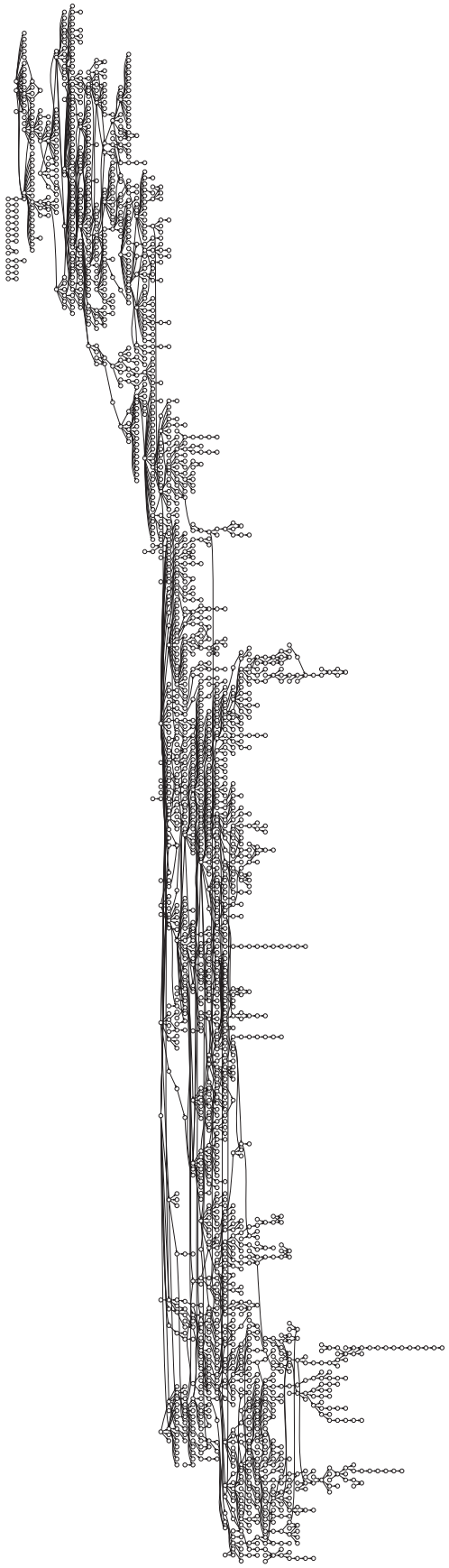

図 2　コサイン類似度の範囲を 0.4〜0.7 に設定したときのタグの階層構造 (左側が上位のノード)

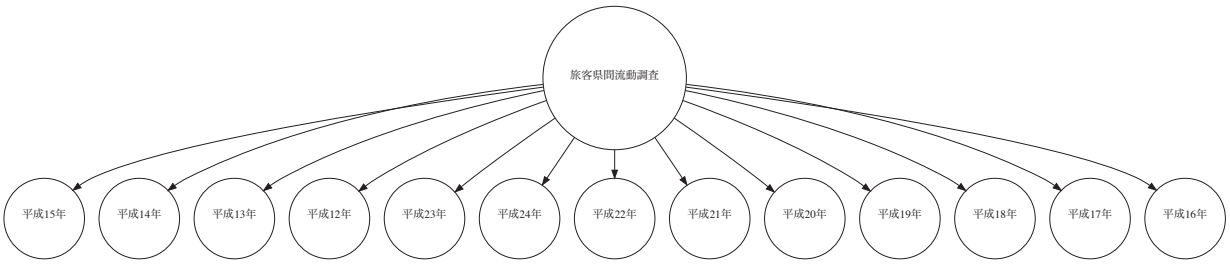

図 3 図 2 における同義語の例

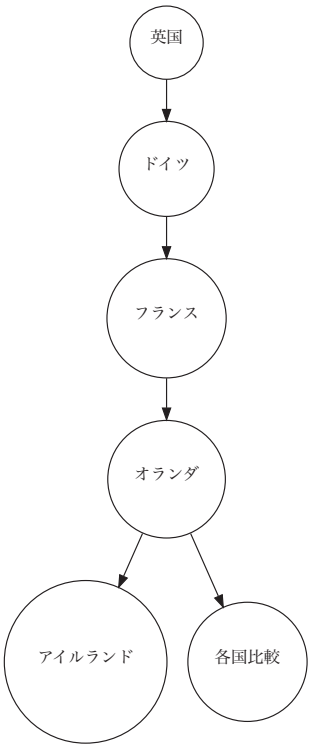

図 4 図 2 における国名の出現する箇所

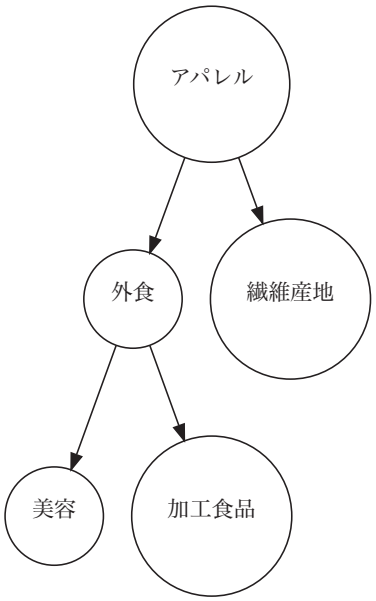

図 5 図 2 の一部 (1)

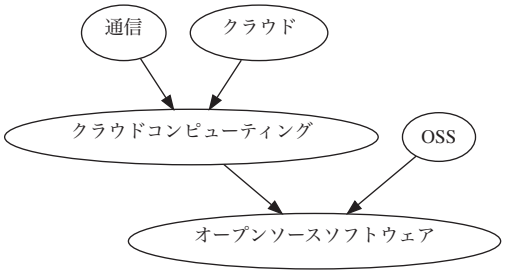

図 6 図 2 の一部 (2)

であっても，関連があると判定された．**図 6** や図 7 のように，意味としてのタグの階層構造が構築された例も存在する．本稿におけるタグの階層構造を構築する目的は，オープンデータに対して人手によるタグ付与と検索の支援である．意味が大きく異なる上位下位関係が構築されている図 5 のような場合であっても，例えば「海外展開を目指すコンテンツ」に関するオープンデータが存在した時，「アパレル」「外食」「美容」それぞれがタグがオープンデータの内容に関連して，図 5 の階層構造を構築されていれば，オープンデータを表現している階層構造が構築できていると考える．

本稿においては，部分的な階層構造を取り出して特徴を説明していたが，定量評価をしていないため，今後の課題として考える．また，定量評価により，オープンデータに対するタグ付与や検索の補助のために，ふさわしい階層構造を構築するための，コサイン類似度の範囲を決定することも今後の課題である．

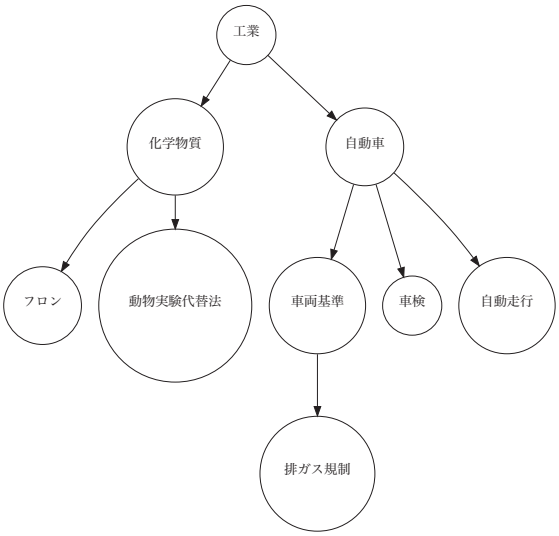

図 7 図 2 の一部 (3)

5. おわりに

本稿では，政府や地方自治体が公開しているオープンデータに対し，タグの付与や検索を支援することを目的として，日本政府のオープンデータカタログサイトである DATA.GO.JP で用いられているタグの階層構造の構築に関する手法を提案した．タグの階層構造を用いることで，あるタグの上位もしくは下位にあたるタグを知ることができる．オープンデータに対してタグを付与する時には，付与すべき一つのタグに対して，タグの階層構造を用いて近くのタグを選択することで，タグ付与に対する労力の負担を軽減することができる．また，検索においても，階層構造を用いてタグを選択することで，質の良い検索結果を望める．

提案手法は，入力として与えられるタグ集合から，それぞれのタグの分散表現を獲得し，コサイン類似度によって，二つのタグに関連があるか判定する．また，タグの頻度と文字数によりタグの上下関係を決定する．

実験では，提案手法により，DATA.GO.JP のオープンデータのタグ集合からタグの階層構造を構築し，考察を行った．コサイン類似度について，高い範囲を設定した時，深さの浅い階層構造が構築され，中程度の範囲を設定した時，深さの深い大規模な階層構造が構築されることを確認した．中程度の範囲では，同義語となるタグ同士が同階層に位置することが確認できたが，国名など上位下位が存在しない語に関しては同階層に位置せず，弧で結ばれた．また，上位タグ下位タグの意味が大きく異なっていてもコサイン類似度の範囲が低いことから，上位下位関係が構築されることを確認した．ただし，語の意味として関連のなさそうな上位下位関係が構築されていても，データセットにおけるタグの共起を表現しているため，必ずしもタグの階層構造として間違っているとは限らないと考える．

本稿では，構築したタグの階層構造に対して，部分的な階層構造の評価しか行っていないため，提案手法により作成されたタグの階層構造の定量的な評価が今後の課題である．

謝辞 本研究は JSPS 科研費 JP19K12715 の助成を受けたものです．

参考文献

[1] 山田泰寛，マルチラベル分類を利用した自治体オープンデータへのタグの付与に関する考察，第 141 回情報システムと社会環境研究発表会，情報処理学会研究報告，Vol. 2017-IS-141，No. 3，pp. 1–6，2017.
[2] Y. Yamada and T. Nakatoh, Tag Recommendation for Open Government Data by Multi-label Classification and Particular Noun Phrase Extraction, Proc. of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) - Volume 3: KMIS, pp. 83–91, 2018.
[3] C. Chen and P. Luo, Enhancing Navigability: An Algorithm for Constructing Tag Trees, Journal of Data and Information Science，Vol. 2, No. 2, pp. 56–75, 2017.
[4] 鈴木宏明，尾崎知伸，分散表現を利用したタグ集合の階層化，人工知能学会全国大会論文集，3A2-3，pp. 1–3，2017.
[5] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient Estimation of Word Representations in Vector Space, Proc. of the International Conference on Learning Representations, arXiv:1301.3781, 2013.
[6] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, Distributed Representations of Words and Phrases and their Compositionality, Proc. of the 26th International Conference on Neural Information Processing Systems, Vol. 2, pp. 3111–3119, 2013.