Vol.2017-IS-141 No.3
2017/8/25

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

山田 泰寛[1,a)]

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，それぞれが運営する Web サイト上において，ライセンスフリーの形式で積極的に公開している．このようなデータをオープンデータと呼ぶ．オープンデータの内容を理解したり，検索をするためには，それぞれのオープンデータに対して，タグと呼ばれる語を付与することは重要である．本稿では，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ついて考察する．マルチラベル分類とは，一つのデータに対して複数のラベルを付与することである．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 18,717 件のオープンデータに対して，従来提案されているマルチラベル分類手法を用いたところ，10 点交差検定で高い精度でタグが付与できることを確認した．

キーワード：オープンデータ，タグ，マルチラベル分類，機械学習

1. はじめに

近年，政府や地方自治体が保有する統計データを，それぞれが運営する Web サイト上に，ライセンスフリーの形態で公開する取り組みが行われている．このようなデータは，オープンデータと呼ばれる．政府は，平成 25 年 6 月 14 日に世界最先端 IT 国家創造宣言 [1] を閣議決定し，各自治体が保有する統計データなどを積極的に公開することを推進している．市民や企業は，公開されたオープンデータから有益な情報を得たり，オープンデータを利用した新しいアプリケーションを開発し，公開することができる．実積らは，オープンデータが活用されると年間 1,800 億円から 3,500 億円の経済効果があると報告している [2]．

しかし，地方自治体がオープンデータを公開し，市民や企業がオープンデータを利活用するためには，いくつかの問題が存在する．ここでは，自治体がオープンデータを公開する際と，市民や企業がオープンデータを検索する際の問題について議論する．

オープンデータは，アプリケーションに依存せず，機械判読が可能で，他のデータへのリンクが含まれている形で公開されることが望ましい [3]．しかし，これを実現するためには，オープンデータを公開する地方自治体の職員は専門的な知識が必要になり，また，負担の大きい仕事である．

次に，市民や企業が必要とするオープンデータを探すことを考える．現在，自治体がオープンデータの公開のために運営する Web サイトは各自治体ごとに分散している．ユーザが必要とするデータが，複数の自治体にまたがって存在する場合は，Web サイトごとに検索を行う必要がある．複数の Web サイトを統合して検索できる仕組みがあれば，検索の手間が小さくなる．このように，オープンデータの普及のためには，オープンデータの公開と利活用を支援するシステムや技術が必要である．

本稿では，オープンデータを公開する際に，オープンデータに付与されるタグと呼ばれる語に着目する．タグとは，テキストや画像，動画などに付与される，その内容を表わす語である．オープンデータの内容を理解したり，検索をするためには，それぞれのオープンデータに対して，タグを付与することは重要である．タグの役割として，オープンデータの中身を見る前に，タグを見ることで，予め，その内容をある程度理解することができる．また，オープンデータを検索する際に，全文検索を用いる場合は，入力語は出現するが，重要でないオープンデータも検索に該当する．タグを用いた検索では，重要とされる語のみを用いて検索を行うことができるため，精度の良い検索が期待できる．

[1] 島根大学大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Shimane University, Shimane 690-8504, Japan
[a)] yamada@cis.shimane-u.ac.jp

このように，オープンデータに対してタグを付与することは，オープンデータの理解や検索において有用であり，オープンデータの利活用の支援となる．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]” においては，ほとんどのデータセットに対してタグが付与されているが，地方自治体のオープンデータにはタグが付与されていない場合も見うけられる．

地方自治体の職員がオープンデータに対してタグを付与することは，その職員の感性に依存し，一貫性を欠く可能性がある．また，職員が単純にデータだけを Web サイト上で公開することと比べると，タグの付与は余分な仕事である．分散している各自治体のオープンデータに対して，各自治体ごとにタグを付与した場合，統一されていないタグが付与されることになる．これは，複数の自治体にまたがってオープンデータを検索するときの障害となる．よって，各自治体で個別にタグを付けるのではなく，統一されたタグの集合を用意し，その中から自動でタグを付与できれば，オープンデータの公開と利活用の支援となる．

本稿は，初めに，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 18,717 件のオープンデータに関して，タグの頻度などの調査を行った．次に，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ついての実験を行った．マルチラベル分類とは，一つのデータに対して複数のラベルを付与することである．実験では，従来提案されているマルチラベル分類手法を用いてタグの付与を行なったところ，10 点交差検定で 92.8%の精度でタグが付与できることを確認した．

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，オープンデータについて述べる．4 節で，Data.go.jp におけるタグの調査について述べる．5 節で，マルチラベル分類によるタグの付与に関する実験について述べる．最後に，6 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

2.1 オープンデータ

地方自治体がオープンデータを公開する Web サイトは，CKAN[*2] というソフトウェアを用いて，公開していることがある．CKAN では，検索のための API(Application Programming Interface) が公開されており，これを用いて検索を行うことができる．本田は，日本政府が公開している API に関する調査をし，API 公開の事例を紹介している [4].

オープンデータを用いたアプリケーション開発に関する研究として，観光イベント情報を利用したモバイル端末向けのイベントガイドアプリの開発 [5] や国土数値情報の災害・防災関連データを利用した洪水から身を守るための防災アプリの開発 [6] などがある．他にも，政府や自治体のオープンデータポータルサイトにおいて，活用事例やアプリケーションが紹介されている．Data.go.jp では，公共データ活用事例 [*3] が紹介されており，福井県鯖江市のポータルサイト「データシティ鯖江ポータルサイト [*4]」においても，オープンデータを活用したアプリケーションを紹介している．

2.2 マルチラベル分類

マルチラベル分類とは，機械学習における分類問題の一つで，一つのデータが複数のラベルを持つ．マルチラベル分類は，訓練データを用いて分類器を学習しておき，未知のデータに対して，その分類器を用いて，複数のラベルを付与することである．訓練データを $\{(\boldsymbol{x_1}, \boldsymbol{y_1}), (\boldsymbol{x_2}, \boldsymbol{y_2}), \cdots, (\boldsymbol{x_n}, \boldsymbol{y_n})\}$ と表わす．ここで，$\boldsymbol{x_i} \in \mathbb{R}^d$ は，i 番目の訓練例の特徴ベクトルであり，$\boldsymbol{y_i} \in \{0,1\}^L$ は，i 番目の訓練例に対して，j 番目のラベルが付与されていれば 1，付与されていなければ 0 をとるベクトルである．マルチラベル分類の学習とは，上で定義された訓練データが与えられたとき，$f : \mathbb{R}^d \to \{0,1\}^L$ となる分類器 f を学習する問題である．マルチラベル分類は，ラベルが未知のデータ $\boldsymbol{x} \in \mathbb{R}^d$ が与えられた時，この分類器 f を用いて，ラベルを予測することである．

マルチラベル分類に関する研究としては [7–10] などがある．手法の詳細については，各文献を参照されたい．本稿では，文献 [7] による手法を用いてマルチラベル分類を行った．この手法のプログラムは，著者のサイト [*5] で公開されている．

3. 自治体オープンデータ

アメリカ政府は，“Data.gov[*6]” においてデータを公開しており，2017 年 3 月 21 日時点で，192,322 件のデータが公開されている．日本政府は，2014 年 10 月に，自身のデータカタログサイトである “Data.go.jp” においてデータの公開を初めた．2017 年 3 月 21 日時点で，18,717 件のデータが公開されている．日本政府だけでなく，地方自治体も積極的にデータの公開を行っている．例えば，島根県松江市は，松江市統計情報データベース [*7] においてデータを公開しており，2017 年 1 月 24 日時点で 2,786 件のデータセットを公開している．この他にも，Data.go.jp において，データの公開を行っている地方自治体の一覧を掲載している．

Tim Berners-Lee によると，オープンデータは，アプリケーションに依存せず，機械判読が可能で，他のデータへ

[*1] http://www.data.go.jp
[*2] https://ckan.org
[*3] http://www.data.go.jp/public-data-case-studies/
[*4] http://data.city.sabae.lg.jp
[*5] https://sites.google.com/site/rohitbabbar/code/dismec
[*6] https://www.data.gov
[*7] http://ntoukei.city.matsue.shimane.jp

図 1 Data.go.jp におけるデータセットのページ

のリンクが含まれている形で公開されることが望ましいとしている [3]．しかしながら，公開されているデータ形式は，政府，自治体によって違いがあり，例えば “Data.go.jp” における，ほとんどのデータセットが PDF 形式か HTML 形式である．ブラジルのオープンデータポータルサイトにおけるデータのほとんどは CSV 形式であるという報告がある [11]．松江市統計情報データベースでは，ほとんどのデータが XLS 形式である．

オープンデータの理想的な公開を実現するためは，地方自治体の職員には専門的な知識が必要になる．これを支援する目的で，LinkData.org[*8] では，表形式のデータを RDF 形式のデータに変換する機能を提供するなど，オープンデータの活用を支援するプラットフォームを提供している [12, 13]．

4. Data.go.jp におけるタグ

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” におけるタグについて調査した．このサイトにおいて，2017 年 3 月 21 日時点で，18,717 件のデータセットが公開されている．図 1 は，Data.go.jp における 1 件のデータセットのページ [*9] である．データセットのタイトルやタグの一覧が表示されている．18,717 件のデータセットのタイトルやタグに関する情報は，Data.go.jp の開発者向け情報のページ [*10] から入手した．

図 2 は，各タグが付与されているデータセット数を示している．横軸はデータセット数でソートしたときのタグの順位，縦軸はデータセット数を表す．縦軸は対数軸としている．タグは 2,489 種類あるが，頻度が高いタグは少数で，ほとんどのタグは頻度が低いことが分かる．1,000 個以上のデータセットに付与されているタグは 28 種類，100 回以上のデータセットに付与されているタグは 144 種類，10 回未満は 2,092 種類であった．

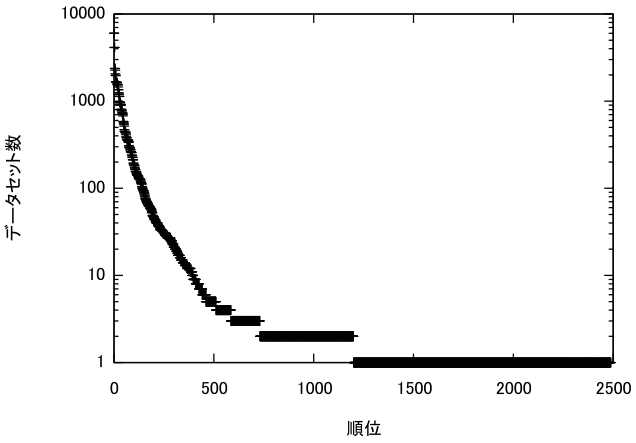

図 2 各タグが付与されているデータセット数

次に，表 1 は，1 個のデータセットに付与されているタグの数を示している．2 個から 10 個のタグが付与されているデータセットが，全体の 87.8%を占めている．385 件のデータセットにはタグが付与されていない．平均で，1 データセット当たり 5.95 個のタグが付与されている．

表 2 は，頻度の上位 20 種類のタグを示している．Data.go.jp ではタグを付与する際に，日本語のタグと英語のタグを同時に付与することが多いため，同じ意味の日本語と英語のタグが同頻度出現している．また，多くのデータに関連ありそうな一般的なタグが使われれいること

[*8] http://ja.linkdata.org
[*9] http://www.data.go.jp/data/dataset/npa_20170313_0007
[*10] http://www.data.go.jp/for-developer/download-of-metadata/

表 1 各データセットに付与されているタグの種類数

タグ数	データセット数
0	385
1	477
2	2,994
3	2,245
4	3,188
5	1,021
6	1,601
7	1,014
8	1,527
9	1,562
10	1,284
11	190
12	120
13	115
14	276
15 以上	718

表 2 Data.go.jp における頻度の上位 20 件のタグ

タグ	データセット数
statistics	6,021
統計	6,018
budgets and final accounts and procurement	4,128
予算_決算_調達関連情報	4,128
statistics survey result	2,370
統計調査結果	2,370
交通	2,268
財政	2,057
white paper and annual report	1,978
白書_年次報告	1,978
budgets	1,667
予算	1,667
security	1,645
budgets and account settlement	1,634
予算及び決算の概要	1,634
安全	1,633
disaster prevention and mitigation	1,559
防災_減災関連情報	1,559
government except elsewhere classified	1,539
公務 他に分類されるものを除く	1,539

が分かる．

表 3 は，1,289 種類ある出現頻度が 1 のタグから 10 個選んだものである．表 2 と比較して，具体的なタグが出現している．

5. マルチラベル分類によるタグの付与に関する実験

本節では，マルチラベル分類によるタグの付与に関する実験について述べる．実験では，2017 年 3 月 21 日に収集した Data.go.jp における 18,717 件のデータセットを用いた．マルチラベル分類には，[7] による手法を用い，[7] の著者のサイトから入手したプログラムを使用した．2.2 節の定義において，テストデータ中の 1 件のデータに対して予測されるラベルの集合は，0 もしくは 1 のベクトルで表現される．そして，1 に対応するラベルを付与する．[7] では，予測されるラベルの集合は実数のベクトルで出力され，付与されるべきラベルほど大きい値を取る．実験では，この上位のラベルを用いて，ラベルの付与を行った．

表 3 Data.go.jp における頻度が 1 であるタグの例

理工系
シミュレーションモデル
特殊容器
破産手続
2020 年東京オリンピックパラリンピック競技大会
化粧品
著作権
地球一個分の暮らし
環境配慮行動
伝統産業

10 点交差検定を用いることにより，マルチラベル分類の精度を調べた．初めに，データをランダムに 10 個のグループに分け，9 個のグループに属するデータを訓練データとして学習を行う．次に，残りの 1 個のグループに属するデータをテストデータとしてタグの予測を行う．各グループをテストデータ，残りのグループを訓練データとして 10 回の実験を行い，精度を調べた．

訓練データとテストデータ共に，データセットのタイトルから名詞のみを抽出し，tf-idf [14] によって数値化した．形態素解析には，MeCab [15] を使用した．

評価に用いる指標として，精度と，参考文献 [7] での $p@k$ を使用する．精度は，テストデータ中の 1 件のデータに付与されているラベルと同数のラベルを予測したときに，正解と一致する割合である．$p@k$ は，予測の上位 k 位までの精度であり，以下のように定義される．

$$p@k = \frac{1}{k} \sum_{l \in rank_k(\hat{\boldsymbol{y}})} \boldsymbol{y_l}$$

k は予測する個数，$rank_k(\hat{\boldsymbol{y}})$ は予測された上位 k 個のラベルのインデックスを表す．また，y_l は正解のラベルベクトルにおいて，l 番目のラベルが付与されていれば 1，付与されていなければ 0 をとる．つまり，$p@k$ は，テストデータに付与されているラベルを予測できた時，$1/k$ を加算する．両指標とも，0 から 1 までの値を取り，1 に近いほど正しく予測が行われていることを示す．

5.1 結果

図 3 は，予測の上位 1 位，3 位，5 位までの $p@k$ の値であり，$p@1$ は 0.959，$p@3$ は 0.892，$p@5$ は 0.764 であった．また，精度は 0.928 であった．このことから，高い精度で

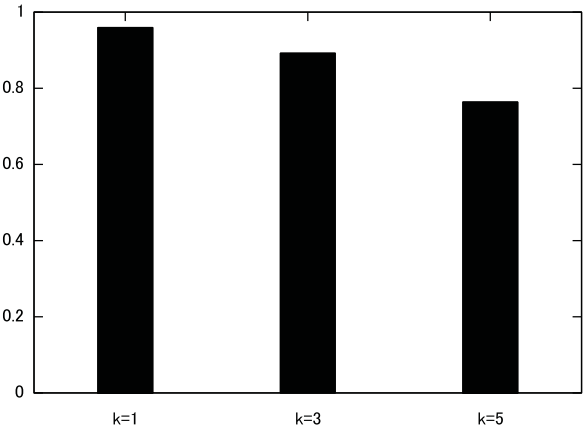

図 3 p@1，p@3，p@5

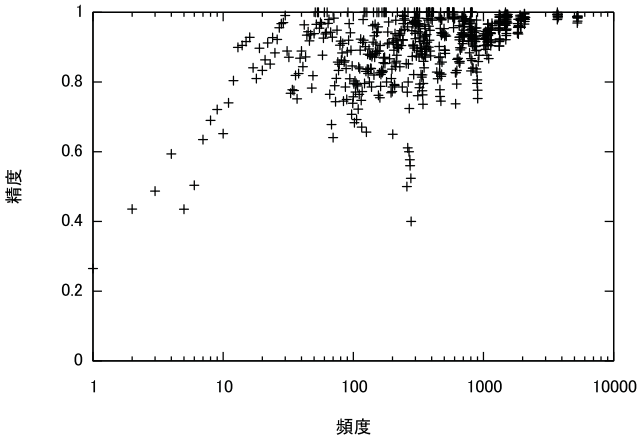

図 4 訓練データにおけるタグの頻度とテストデータにおける精度

タグの付与が行えることを確認できた．

k の値が増えるほど，$p@k$ の値が低くなるため，予測の順位の低いタグほど正しく予測されていないことを示している．また，テストデータに含まれるタグの数が，1 データにつき 3 個未満，もしくは，5 個未満である場合，$p@3$ と $p@5$ が低くなる．

次に，正しく予測できなかったタグについて考察する．図 4 は，訓練データにおけるタグの頻度と，テストデータにおけるそのタグの予測の精度を示したものである．横軸は対数軸としている．図 4 より，訓練データにおいて頻度の低いタグは，テストデータにおけるタグの予測の精度が低いことが分かる．

単語の一般性と特殊性はその単語の頻度と関係があり，一般的な単語は頻度が高く，特殊な単語は頻度が低くなる傾向にある．実験より，頻度の低いタグは正しく付与されにくい．しかし，頻度の低い特殊なタグも，そのオープンデータを理解するためには有用なタグであると考えられるため，頻度の低いタグの付与は重要である．

6. おわりに

本稿では，地方自治体が公開するオープンデータに対して，統一的なタグを付与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を付与することを考察した．実験では，従来提案されているマルチラベル分類手法で十分な精度が得られることが確認できた．一方で，訓練データにおいて出現頻度の低いタグについては，テストデータにおけるタグの予測の精度が低いことが確認できた．

本稿では，統一的なタグの中から，オープンデータにふさわしいタグを付与することを考えた．一方で，1 個のオープンデータに特有のタグを付与することも重要である．例えば，自治体のある地域特有の重要語などはタグとして付与されるべきである．よって，オープンデータそのものから重要語を抽出することで，そのオープンデータに特有のタグを付与することが今後の課題である．

また，訓練データにおいて頻度の低いタグを，精度良く付与する手法の開発も今後の課題として挙げられる．頻度の低いタグは，意味を限定するタグであり，オープンデータを説明する上で有用である．

以上の課題を解決した後，オープンデータのタイトルを入力するとタグを推薦する Web システムの開発を目指したい．

参考文献

[1] 世界最先端 IT 国家創造宣言について，平成 25 年 6 月 14 日．http://www.kantei.go.jp/jp/singi/it2/kettei/pdf/20130614/siryou1.pdf (2017.7.26).
[2] 実積寿也，八田真行，野田哲夫，渡辺智暁：Innocation Nippon 研究会報告書 オープンデータの経済効果推計，(2013)．http://innovation-nippon.jp/reports/2013StudyReport_OpenData.pdf (2017.7.26).
[3] 5-star Open Data. http://5stardata.info/en/ (2017.7.26).
[4] 本田正美：日本政府による API 公開の現状と課題，情報処理学会研究報告，Vol. 2017-IS-139，No. 2，pp. 1–4 (2017).
[5] 荻島和真，福安真奈，浦田真由，遠藤守，安田孝美：観光イベント情報を活用したオープンデータ化の試行と実践，社会情報学，Vol. 4，No. 2，pp. 1–16 (2016).
[6] 天野貴文：オープンデータ・国土数値情報を活用したスマートフォン向け洪水ハザードマップアプリの開発，GIS-理論と応用，Vol. 23，No. 2，pp. 37–42 (2015).
[7] Babbar, R. and Schölkopf, B.: *DiSMEC: Distributed Sparse Machines for Extreme Multi-label Classification*, Proc. of the 10th ACM International Conference on Web Search and Data Mining, pp. 721–729 (2017).
[8] Prabhu, Y. and Varma, M.: *FastXML: A Fast, Accurate and Stable Tree-classifier for eXtreme Multi-label Learning*, Proc. of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 263–272 (2014).
[9] Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y. and Yan, S.: *CNN: Single-label to Multi-label*, CoRR, Vol. abs/1406.5726 (2014). http://arxiv.org/abs/1406.5726 (2017.7.27).
[10] Xu, C., Tao, D. and Xu, C.: *Robust Extreme Multi-label Learning*, Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1275–1284 (2016).

[11] Oliveira, M. I. S., Oliveira, H. R. de, Oliveira, L. A. and Lóscio, B. F.: *Open Government Data Portals Analysis: The Brazilian Case*, Proc. of the 17th International Digital Government Research Conference on Digital Government Research, pp. 415–424 (2016).
[12] 下山紗代子, 西方公郎, 吉田有子, 豊田哲郎：LinkData.orgを使ったRDF教育とデータ公開化運動の推進，人工知能学会全国大会論文集，Vol. 26，3C2-OS-13b-2，pp. 1–3 (2012).
[13] 下山紗代子，豊田哲郎：行政と市民によるオープンデータ共創支援プラットフォーム，人工知能学会全国大会論文集，Vol. 28，1G5-OS-19b-6in，pp. 1–4 (2014).
[14] Salton, G. and McGill, J. M.: *Introduction to Modern Information Retrieval*, McGraw Hill New York (1983).
[15] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. `http://taku910.github.io/mecab/` (2017.7.26).