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　６ 年　 ６ 月 １９ 日現在

機関番号：１５２０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2023
課題番号：１９Ｋ１２７１５

研究課題名（和文）政府・自治体オープンデータの公開と検索の支援を目的としたタグ付与に関する研究

研究課題名（英文）Tag Recommendation to Support the Release and Retrieval of Open Government Data

研究代表者
山田　泰寛（Yamada, Yasuhiro）

島根大学・学術研究院理工学系・助教

研究者番号：５０５２９６０９

交付決定額（研究期間全体）：（直接経費）　　1,900,000円

研究成果の概要（和文）：本研究は，政府がWeb上で公開している統計データ(オープンデータ呼ぶ)に対して，データセットの内容を表わす語であるタグ(ラベル)を自動で付与することを目的としている．1個のデータセットに対して複数のラベルを付与する手法であるマルチラベル分類を用いて，特に学習データにおいて出現回数の少ないラベルに着目し，それを付与することを目指している．
1個のデータセットにおいて，複数のラベルが同時に出現することを利用して，出現回数の少ないラベルの学習データを増やすオーバーサンプリング手法を提案した．また，オープンデータのタイトルや説明を入力として与えたとき，付与すべきタグを推薦するシステムの開発を行なった．

研究成果の学術的意義や社会的意義
学習データにおいて出現回数の少ないタグは推薦されにくいという問題に対して，疑似的にそれらの学習データを増やす手法を開発した．また，オープンデータのタイトルや説明を入力したとき，そのオープンデータに対して付与すべきタグを推薦するシステムを開発した．オープンデータを公開する際に，ふさわしいタグを付与することの一助となることが期待できる．また，付与されたタグがオープンデータの検索の際にも役立つことが期待できる．

研究成果の概要（英文）：The purpose of this research is to automatically assign tags (labels) to statistical data published on the Web by the government, which is called open government data. We use multi-label classification, a method that assigns multiple labels to a single dataset. We are particularly interested in infrequent labels in training data and aim to assign them.
Focusing on the simultaneous occurrence of multiple labels in a single dataset, we proposed an oversampling method to increase the training data for labels that appear infrequently. Also, we have developed a system that recommends tags to be assigned to a single dataset when the title or description of the dataset is given as input.

研究分野：図書館情報学および人文社会情報学関連

キーワード：オープンデータ　テキストマイニング　タグ推薦

１．研究開始当初の背景

(1) 政府や地方自治体が保有する統計データなどをライセンスフリーの形で公開する動きが広がっている. このようなデータはオープンデータと呼ばれる. オープンデータを公開する際には, メタデータが付与されるが, その中の一つとして, データの内容を表わす語であるタグが付与される. しかし, 政府や自治体の職員がオープンデータに対して手動でタグを付与するには, データそのものを熟知している必要がある.

機械学習において, 過去のデータセットの集合からラベルを学習し, ラベルが未知のデータセットに対して複数のラベルを推定する手法を, マルチラベル分類という. オープンデータに対して自動でタグを付与するために, マルチラベル分類を使うことができる. しかし, マルチラベル分類における問題の 1 つとして, 学習データにおいて出現回数の少ないラベルは予測しにくいという問題がある.

２．研究の目的

(1) 政府や自治体がオープンデータを公開する際の支援と, 利用者によるオープンデータの検索の支援を目的として, オープンデータに対してタグを自動付与する手法とシステムの開発を行う. 特に, 学習データにおいて頻度の低いラベルに着目し, それを推定することを目指す. 頻度の低いタグは, オープンデータの内容について具体的で専門的な意味を持つ語であり, 特定性の高い語である. このため, データセットの中身を見る前にその内容を大まかに理解すること, また, 検索結果を絞り込むことに役立つ.

３．研究の方法

(1) 1 データセットにおいて同時に出現するタグを, そのデータセットにおいてタグが共起していると言う. 頻度の低いタグと共起するタグの特徴ベクトルを利用し, 頻度の低いタグのデータ数を増やした後, 分類器の学習を行うことで, 頻度の低いタグの推薦の出現回数と精度を上げる.

４．研究成果

(1) 頻度の低いタグの推薦

引用文献[4, 5]の手法をベースに, データセットにおけるラベルの共起を利用して, 出現回数の少ないラベルの学習データを増やすオーバーサンプリング手法を提案した(引用文献[1, 2]).

提案手法は, 初めに, 学習データとして与えられた各データセットに対して, そのデータセットのタイトルと説明から単語の出現頻度を表す特徴ベクトルと, ラベルの出現を表すラベルベクトルの組を作成する. 全データセットにおいて, 出現回数の少ないラベル(少数ラベルと呼ぶ)の特徴ベクトルを 1 つ選択し, 中心ベクトルとする. 中心ベクトルとしてのデータセットに出現する複数のラベルに対して, 少なくとも 1 つラベルが共通し, かつ, ユークリッド距離が近い特徴ベクトルを n 個選択する。n 個の特徴ベクトルに重み付けし, 重みを用いた確率で, n 個の特徴ベクトルから 1 個選択する. 選択した近傍の特徴ベクトルと中心ベクトルとの間にランダムで 1 つ点を生成し, 新しい特徴ベクトルとする. 新しい特徴ベクトルに対応する少数ラベルを付与し, 学習データに追加する. 以上の手順を, 学習データにおいて少数ラベルの出現回数が閾値に達するまで繰り返す.

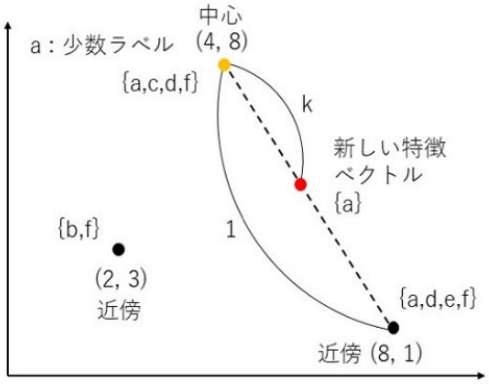

図 1. 新しい特徴ベクトルの作成(引用文献[1])

図 1 は, 特徴ベクトルが 2 次元であるとしたとき, 3 個のデータセットの特徴ベクトル(黄点と

黒点)と，そのデータセットに付与されているラベルを表している．少数ラベルを a とし，このラベルの特徴ベクトルを疑似的に増やすことを考える．中心の特徴ベクトル x_c=(4, 8)(黄点)とする．中心ベクトルのデータセットに出現するラベル a，c，d，f の内，少なくとも一つのラベルが出現する近傍の特徴ベクトルを x_1=(8, 1)，x_2=(2, 3)(黒点)とする．この 2 個の近傍から x_1 が選ばれたとすると，中心ベクトルと x_1 の間に，ラベル a を持つ特徴ベクトル(赤点)を新たに作成する．この特徴ベクトルと，ラベル a だけからなるラベルベクトルを学習データに追加する．

実験では，日本政府のデータカタログサイトである Data.go.jp から収集した 27,169 件のデータセットを使用した．scikit-learn のナイーブベイズ法(MultinomialNB)を用いて学習を行なった．分類器はそれぞれのラベルに対して，そのラベルと他のラベルを分類するための分類器を作成した(OneVsRestClassifier)．表 1 は，少数ラベルのみを対象とした再現率，精度，F 値の macro 平均と micro 平均とラベルの予測回数である．

表 1. 少数ラベルに対する oversampling 前後の実験結果(引用文献[1])

	oversampling なし	提案手法
学習データ数	21,735	33,059
予測回数	74,738	78,155
Macro recall	0.0620	0.2292
Macro precision	0.0014	0.0047
Macro f1	0.0027	0.0092
Micro recall	0.0671	0.2215
Micro precision	0.0016	0.0049
Micro f1	0.0030	0.0096

実験結果より，学習手法としてナイーブベイズ法を用いたときに，少数ラベルについては予測回数が増えることと，再現率が改善されることを確認した．SMOTE(引用文献[4])など他の手法との精度の比較や，ナイーブベイズ法以外の学習手法における精度の調査は今後の課題である．

(2) タグ推薦システムの構築

オープンデータのタイトルや説明を入力として与えたとき，付与すべきタグを推薦するシステムの開発を行なった(引用文献[3])．分類器の作成において，オーバーサンプリング手法として SMOTE(引用文献[4])の手法を利用した．(1)で提案した手法の組み込みやシステムの Web 上への公開は今後の課題である．

Data.go.jp では，タグについて次のようにカテゴリが分けられている：(1) G8 の重要データカテゴリ，(2) 「行政情報の電子的提供に関する基本的考え方（指針）」における「共通のカテゴリー」，(3) 電子行政アクションプランにおける業務分類，(4) 「電子行政オープンデータ推進のためのロードマップ」における重点分野，(5)その他のタグ．開発したタグ推薦システムでは，(5)その他のタグについてのみオーバーサンプリング手法を適応した．タグ推薦システムは，5 個のカテゴリそれぞれに対して，タグを推薦するための分類器をあらかじめ学習しておく．

図 2. タグ推薦システムの入力(左図)と出力(右図) (引用文献[3])

タグ推薦システムは，オープンデータのタイトルや説明と上の 5 つのカテゴリから一つ選択したものを入力し(図 2 左)，事前に学習した分類器を用いて，そのオープンデータにふさわしいタグを出力する(図 2 右)．図 2 は，「気象予報_天気予報・台風の資料」を入力したときに，出力されたタグである．

開発したタグ推薦システムにより，オープンデータを公開する際に，ふさわしいタグを付与することの一助となることが期待できる．また，検索の際にも役立つことが期待できる．この研究成果について，国際会議 KMIS2023 において発表を行なった．

(3) 今後の課題

① 新しいタグの抽出

マルチラベル分類では，過去のデータセットで既に付与されているタグしか推薦できない．申請時においては，オープンデータそのものから新しいタグを抽出することを目的の一つとしていた．しかし，頻度の低いタグの推薦手法の開発とタグ推薦システムの開発に多くの時間を割り当てたため，この問題については今後の課題である．

② 新しい日本政府オープンデータサイト「e-Gov Data Portal」

本研究において実験で使用したデータのある日本政府のオープンデータサイト Data.go.jp は，2023 年 3 月に新しいサイト e-Gov Data Portal に移行された．新しいサイトでは，タグ付与に関する方針が変更されているため，新しいサイトのデータに対するシステムの構築が今後の課題である．

③ 大規模言語モデルの利用

研究期間中に，テキストに対する分析技術として，BERT などのディープラーニング技術や，大規模言語モデルが急速に開発されてきた．このような技術を用いることにより，低頻度ラベルの推定の調査と，それらを用いた低頻度ラベルの推薦手法の開発も課題である．

＜引用文献＞

① 河野湧芽，山田泰寛，政府オープンデータにおける少数ラベルの推定，2022 年度(第 73 回)電気・情報関連学会中国支部連合大会，2022.
② 河野湧芽，政府オープンデータにおける低頻度なラベルの予測，島根大学大学院自然科学研究科理工学専攻知能情報デザイン学コース，学位論文（修士論文），2023.
③ Y. Yamada, Tag Recommendation System for Data Catalog Site of Japanese Government, Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023), pp. 325-331, 2023.
④ N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, Vol. 16, No. 1, pp. 321-357, 2002.
⑤ F. Charte, A. J. Rivera, M. J. del Jesus and F. Herrera, MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation, Knowledge-Based Systems, Vol. 89, pp. 385-397, 2015.

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Yamada Yasuhiro	４．巻 KMIS
２．論文標題 Tag Recommendation System for Data Catalog Site of Japanese Government	５．発行年 2023年
３．雑誌名 Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023)	６．最初と最後の頁 325-331
掲載論文のDOI（デジタルオブジェクト識別子） 10.5220/0012260000003598	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 河野湧芽	４．巻 －
２．論文標題 政府オープンデータにおける低頻度なラベルの予測	５．発行年 2023年
３．雑誌名 島根大学大学院自然科学研究科理工学専攻知能情報デザイン学コース，学位論文（修士論文）	６．最初と最後の頁 －
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 河野湧芽，山田泰寛	４．巻 －
２．論文標題 政府オープンデータにおける少数ラベルの推定	５．発行年 2022年
３．雑誌名 2022年度(第73回)電気・情報関連学会中国支部連合大会	６．最初と最後の頁 －
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 Yamada Yasuhiro
２．発表標題 Tag Recommendation System for Data Catalog Site of Japanese Government
３．学会等名 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023)（国際学会）
４．発表年 2023年

１．発表者名
河野湧芽

２．発表標題
政府オープンデータにおける少数ラベルの推定

３．学会等名
2022 年度(第73 回)電気・情報関連学会中国支部連合大会

４．発表年
2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

政府・自治体オープンデータ・タグ推薦システム
http://buti.cis.shimane-u.ac.jp/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関