Tag Recommendation for Open Government Data by Multi-label Classification and Particular Noun Phrase Extraction

Yasuhiro Yamada[1] and Tetsuya Nakatoh[2]
[1] *Institute of Science and Engineering, Academic Assembly, Shimane University,*
1060 Nishikawatsu-cho, Matsue-shi, Shimane, 690-8504, Japan
[2] *Research Institute for Information Technology, Kyushu University,*
744 Motooka, Nishi-ku, Fukuoka, 819-0395, Japan

Keywords: Open Government Data, E-Government, Tag Recommendation, Multi-label Classification, Metadata.

Abstract: Open government data (OGD) is statistical data made and published by governments. Administrators often give tags to the metadata of OGD. Tags, which are a collection of a single word or multiple words, express the data. Tags are useful to understand the data without actually reading the data and also to search for OGD. However, administrators have to understand the data in detail in order to assign tags. We take two different approaches for giving appropriate tags to OGD. First, we use a multi-label classification technique to give tags to OGD from tags in the training data. Second, we extract particular noun phrases from the metadata of OGD by calculating the difference between the frequency of a noun phrase and the frequencies of single words within the noun phrase. Experiments using 196,587 datasets on Data.gov show that the accuracy of prediction by the multi-label classification method is enough to develop a tag recommendation system. Also, the experiments show that our extraction method of particular noun phrases extracts some infrequent tags of the datasets.

1 INTRODUCTION

Open government data (OGD) is statistical data published by governments on their websites. The categories of the data are various, for example, budget, education, health and finance. One purpose of OGD is to enable anyone to freely access and reuse this data[1]. The U.S. Government publishes OGD on the site "Data.gov[2]". This site had 196,587 datasets on September 12th, 2017. The Japanese government started publishing OGD on the site "Data.go.jp[3]". This site had 18,717 datasets on March 21st, 2017. Some local governments also have published their data on their own sites.

There are three kinds of stakeholders to recognize the benefits of OGD: publishers, re-users, and consumers (Köster and Suárez, 2016). Re-users develop applications using OGD. Consumers obtain useful information from OGD and use the applications developed by re-users.

When government agents publish OGD on the Web, they generally create metadata about their OGD. Examples of metadata of OGD are the id, the title, the description, the tags, and the publish date of the OGD. This paper focuses on the tags, which are descriptive keywords of OGD. The tags are useful for understanding the content of OGD without actually reading dataset files. The tags also help re-users and consumers to search for OGD that they want. A search by tags enables them to find the desired OGD with accuracy because tags are important words about the OGD.

We collected 196,587 datasets from Data.gov. Each dataset has one or more resources (files). In total, the datasets have 1,105,063 resources. The number of datasets with tags is 73,304 (37.3%). The other 123,283 datasets do not have tags. Publishers have to understand the OGD in detail to give the tags. This means that giving appropriate tags to the OGD is difficult and burdensome work. Therefore, a system to recommend tags automatically is needed.

The lack of consistency in tags negates the advantage of tags in searches. For example, different publishers select different tags for the same OGD. It is important to select appropriate tags from a common tag set. Also, when publishers give new tags to an

[1] Open definition 2.1. http://opendefinition.org/od/2.1/en/ (accessed Jan. 23rd, 2018)
[2] https://www.data.gov
[3] http://www.data.go.jp

Yamada, Y. and Nakatoh, T.
Tag Recommendation for Open Government Data by Multi-label Classification and Particular Noun Phrase Extraction.
DOI: 10.5220/0006937800830091
In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) - Volume 3: KMIS*, pages 83-91
ISBN: 978-989-758-330-8

Table 1: Target of this paper.

	frequent tags (words)	infrequent tags (words)
tags labeled manually in an OGD portal site	**multi-label classification**	(It is difficult to predict infrequent tags by multi-label classification techniques.)
tags not used in the above site	(Tags in this cell will be in the above frequent tags.)	**particular tags extraction from the title and the description of a dataset and its resources**

OGD, they again come up with different tags. It is a significant task to extract tags that are not in the common tag set from the OGD.

This paper takes two different approaches to recommend tags for OGD (see Table 1). The first target is to use a multi-label classification technique to predict appropriate tags for a dataset from tags used in an OGD portal site. The multi-label classification technique learns classifiers for tags from datasets which have already been given tags. Then, it predicts tags for a dataset without tags by using the learned classifiers. However, it is difficult to predict tags which are infrequent in the datasets because the amount of data of infrequent tags is too small for learning (Jain et al., 2016).

The second target is to extract new tags from the title and the description of the OGD. There are various viewpoints with respect to appropriate tags for the OGD. A tag recommendation system should display candidate tags from various viewpoints. We apply a term weighting method in (Yamada et al., 2018) to extract particular tags in the OGD. The extracted tags are noun phrases. The idea is simple: a noun phrase is considered particular if the nouns within the noun phrase appear only in the phrase. On the other hand, frequent words will appear in tags of multi-label classification because they are commonly used in the datasets.

The contributions of this paper are summarized as follows:

- The first target is tags labeled manually in an OGD portal site. We verify the accuracy of three typical multi-label classification methods for datasets with tags on Data.gov. We use 196,587 datasets on Data.gov in experiments. The result of multi-label classification shows that the accuracy of prediction by the multi-label classification method is enough to develop a tag recommendation system.
- The second target is noun phrases which are not in the above tags. We propose a method for extracting particular noun phrases as tags from the title and description of the datasets and their resources. The experiments of particular noun phrase extraction show that our method extracts some infrequent tags of the datasets.

This paper is organized as follows. Section 2 describes related research. Section 3 shows statistics about the tags of datasets of OGD on Data.gov. Section 4 describes multi-label classification for OGD. Section 5 describes particular tag extraction from OGD. Section 6 shows experiments applying the multi-label classification and the particular tag extraction. Finally, our conclusions are presented in Section 7.

2 RELATED WORK

In this section, we describe two kinds of related research: open government data and tag recommendation.

2.1 Open Government Data

OGD of governments is published on their data catalog sites. The CKAN platform[4] is often utilized to publish open government data (Oliveira et al., 2016). It is desired that OGD is published with machine-readable and non-proprietary data formats such as CSV and XML[5]. The data formats of OGD are varied and include, for example, PDF, XLSX, CSV, XML, and HTML. Oliveira et al. reported that the CSV format is the most used data format in Brazilian OGD portals (Oliveira et al., 2016). Corrêa and Zander reported that about 13% of dataset files in some main open data portals around the world are PDF formats (Corrêa and Zander, 2017). Most OGD on the Japanese government OGD portal site Data.go.jp are PDF or HTML files.

Some other proposed research studies support publishers of OGD (Corrêa and Zander, 2017; Tambouris et al., 2017). Corrêa and Zander investigated methods and tools for extracting tables in PDF files (Corrêa and Zander, 2017). A lot of OGD have tables because they are statistical data. Therefore, it is important to translate tables in PDF files into a non-proprietary open format such as CSV. Linked open data is the most desirable format of OGD. However, it is difficult for publishers serving as government agents to make OGD satisfy the requirement for linked open data because traditionally they do not have

[4] http://ckan.org

[5] 5-star open data. http://5stardata.info/en/ (accessed Jan. 25th, 2018)

the required skills. Tambouris et al. presented tools to help the publishers make linked open data from various file formats(Tambouris et al., 2017). Our paper focuses on tags as metadata of OGD. However, we could not find research to support the publishers in the task of labeling tags of OGD.

As a reuse of OGD, some OGD portal sites of governments have introduced applications using OGD. Some applications using OGD of the Japanese government are introduced on Data.go.jp. For example, the Japan Seismic Hazard Information Station[6] was established by the National Research Institute for Earth Science and Disaster Resilience to help prevent and prepare for earthquake disasters. Vasa and Tamilsevam developed a web application that uses data from the Department of Agriculture in India (Vasa and Tamilselvam, 2014). This application helps users select recipes based on real-time food prices.

2.2 Tag Recommendation

One approach to tag recommendation is multi-label classification. Given a set of training examples each of which consists of a feature vector and a set of labels, the learning of multi-label classification consists of generating a classifier for the labels. Then, using the classifier, the multi-label classification predicts labels from an example without labels. For further information, refer to previous surveys that describe the definition of multi-label classification, algorithms, datasets, and evaluation measures (Herrera et al., 2016; Tsoumakas et al., 2010).

Some research has dealt with a large number of labels (Babbar and Schölkopf, 2017; Jain et al., 2016; Prabhu and Varma, 2014; Xu et al., 2016). Compared with the datasets in (Babbar and Schölkopf, 2017; Jain et al., 2016; Prabhu and Varma, 2014; Xu et al., 2016), the datasets of Data.gov considered in the present paper are also considered to be extreme. We verify the accuracy of three typical multi-label classification methods for the datasets in Section 6.1.

Another approach is to extract a candidate term, which is a single word or multiple words, as an appropriate tag from texts other than tags (Martins et al., 2016; Ribeiro et al., 2015). Ribeiro et al. extracted candidate terms from the publication metadata of researchers (Ribeiro et al., 2015). Martins et al. dealt with the title and descriptions of a target object (Martins et al., 2016). In the present study, the collective target from which to extract tags is the title and the description of a dataset and its resources on Data.gov. Candidate terms in the present paper are noun phrases.

Some research has proposed metrics for calculating the relevance of a candidate term as tag recommendation for an object. Venetis et al. and Ribeiro et al. used three kinds of metrics: term frequency, the tf-idf of a term, and the coverage of terms (Ribeiro et al., 2015; Venetis et al., 2011). The present paper focuses on the discriminative power of a tag. In contrast to popularity, which means that a tag is assigned to numerous objects, a tag with discriminative power distinguishes a small number of specific objects from other objects. For example, metrics for the discriminability are the Inverse Feature Frequency (Figueiredo et al., 2013; Martins et al., 2016) and the document frequency of a term. When limited to noun phrases as candidate terms, we propose a new method to calculate the discriminative power of a noun phrase in Section 5.

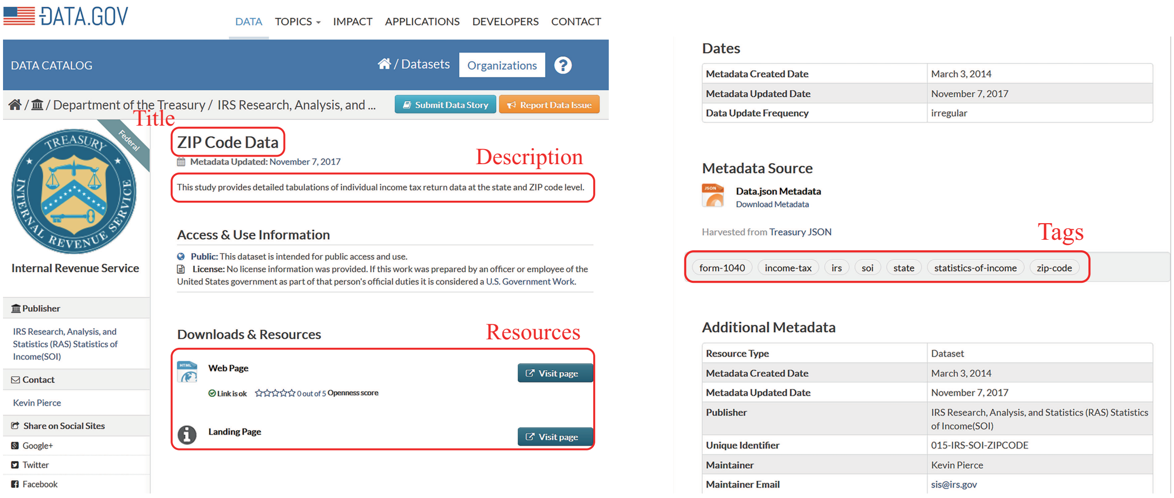
3 STATISTICS OF TAGS IN DATA.GOV

This section describes the statistics of tags in Data.gov, which is the OGD portal site of the U.S. Government. Figure 1 is a Web page of OGD on Data.gov. The left side of the figure is the top of the page, and the right side is the rest of the page. The left side describes the title, the description of a dataset, and the right side shows the date, tags, and other information. A dataset has one or more resource files. The metadata of a dataset includes the title, the description, and the tags of the dataset.

We collected 196,587 datasets of Data.gov on September 12, 2017. The number of all tags in the datasets is 57,430. Figure 2 shows the number of datasets that each tag appears. The tags are ranked according to the number of datasets in which they appear. The vertical axis is in log scale. We see that most of the tags are infrequent. The number of tags appearing once in the datasets is 31,332 (54.6%). The number of tags whose frequency is less than or equal to 10 is 52,435 (91.3%). On the other hand, the number of tags whose frequency is greater than or equal to 1,000 is 41 (0.0007%). Table 2 shows the top 20 most frequent tags. Table 3 lists examples of tags appearing once.

Figure 3 shows the number of tags in a dataset. Both axes are in log scale. The number of datasets with tags is 73,304. On the other hand, 123,283 datasets (62.7%) do not have tags. Therefore, a tag recommendation system is needed. The number of datasets with only one tag is 4,608. The maximum number of

[6] http://www.j-shis.bosai.go.jp/en/



Internal Revenue Service

Publisher

IRS Research, Analysis, and Statistics (RAS) Statistics of Income(SOI)

Contact

Kevin Pierce

Share on Social Sites

Google+

Twitter

Facebook

ZIP Code Data

Metadata Updated: November 7, 2017

This study provides detailed tabulations of individual income tax return data at the state and ZIP code level.

Access & Use Information

Public: This dataset is intended for public access and use.

License: No license information was provided. If this work was prepared by an officer or employee of the United States government as part of that person's official duties it is considered a U.S. Government Work.

Downloads & Resources

Web Page

Link is ok ☆☆☆☆☆ 0 out of 5 Openness score

Landing Page

(a)

Dates

Metadata Created Date	March 3, 2014
Metadata Updated Date	November 7, 2017
Data Update Frequency	irregular

Metadata Source

Data.json Metadata

Download Metadata

Harvested from Treasury JSON

form-1040 income-tax irs soi state statistics-of-income zip-code

Additional Metadata

Resource Type	Dataset
Metadata Created Date	March 3, 2014
Metadata Updated Date	November 7, 2017
Publisher	IRS Research, Analysis, and Statistics (RAS) Statistics of Income(SOI)
Unique Identifier	015-IRS-SOI-ZIPCODE
Maintainer	Kevin Pierce
Maintainer Email	sis@irs.gov

(b)

Figure 1: Web page[7] of OGD on Data.gov.

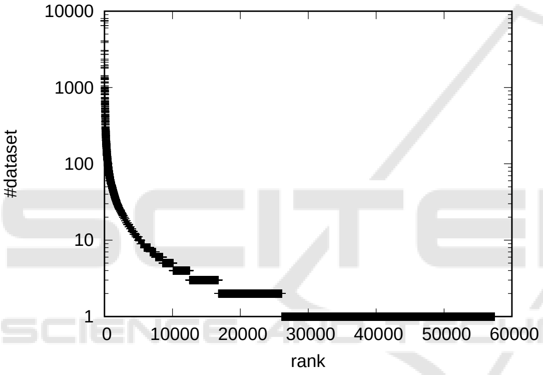


Figure 2: Number of datasets that each tag appears.

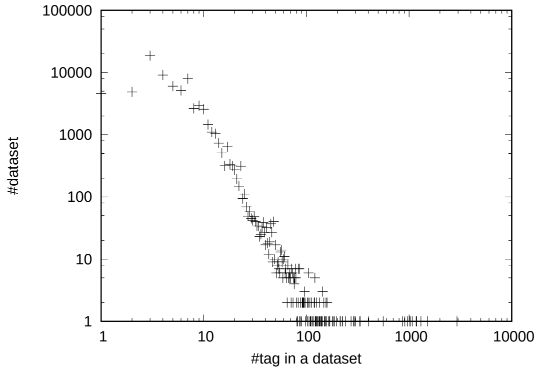


Figure 3: Number of tags in a dataset.

tags that a dataset has is 2,932, and the average number of tags in the 73,304 datasets is 6.59.

Table 2: Top 20 most frequent tags on Data.gov.

tag	#datasets
animal-studies	7,997
project	7,514
coral-reef	7,513
coral	7,477
aquatic-habitats	7,445
transect	7,442
marine-systems	7,432
photo-quadrat	7,432
completed	6,459
general-management-natural-resources-management-wildlife-management	6,433
general-management-inventory	4,070
waterfowl	4,046
earth-science	3,868
annual-narrative	3,063
pocillopora	2,988
annual-narrative-report	2,730
porites	2,698
oceans	2,357
general-management-monitoring	2,262
montipora	2,136

4 TAG RECOMMENDATION USING MULTI-LABEL CLASSIFICATION

The first approach to recommend tags for OGD is multi-label classification. The target tags are

[7] https://catalog.data.gov/dataset/zip-code-data

Table 3: Examples of tags appearing once.

ecological-history
lmi-energy-data
nest-tree-condition
water-quality-data-standards
yellow-billed-magpie
washington-suburban-sanitary-commission
depository-institution
recreation-information-database
human-conflict

ones which have already been used in the datasets. Let $L = \{l_1, l_2, \ldots, l_m\}$ be a set of labels, and $D = \{(\boldsymbol{x}_1, Y_1), (\boldsymbol{x}_2, Y_2), \ldots (\boldsymbol{x}_n, Y_n)\}$ be a set of training examples where $\boldsymbol{x}_i$ is the feature vector and $Y_i \subseteq L$. The multi-label learning task is to make a classifier for L from D. Then, given an unlabeled example $\boldsymbol{x}$, the classifier predicts labels for $\boldsymbol{x}$.

When we apply the multi-label classification to the datasets on the site Data.gov, we make the feature vector $\boldsymbol{x}_i$ of a dataset as a vector for the weighting of nouns appearing in the title of the dataset. The weighting is the term frequency of a noun in the title. The set of labels Y_i corresponds to the tags in a dataset.

We employ the one-vs-rest strategy, which generates classifiers for each label to distinguish a label from other labels. We use support vector machine, random forests (Breiman, 2001) and multinomial naive Bayes (Manning et al., 2008) methods for making a classifier that distinguishes a label from the rest. The methods are implemented by scikit-learn[8]. We compare the accuracy of the three typical methods in experiments.

5 PARTICULAR NOUN PHRASE EXTRACTION

The multi-label classification in the previous section uses tags already given to OGD. This classification, therefore, can select only from the tags, and it cannot predict words which are not in the tags. The second approach considers extracting words as appropriate tags from the OGD rather than using the tags.

We counted the words in each tag of the datasets on Data.gov. We found that 36,340 (63.3%) of all 57,430 tags consist of multiple words[9]. When limited to tags whose frequency is less than 6 in the datasets, 32,139 (65.8%) of 48,856 tags consist of multiple words. Therefore, many infrequent tags are multiple words, such as noun phrases.

First, we extract noun phrases from the title and the description of a dataset and its resources of OGD. We see them as a text. We use patterns for the noun phrases reported in (Kang et al., 2015). The patterns are as follows:

< NP >::=< Pre >< NN > | < NN >
| < NP > “in” < NP >
< Mod >::= “jj”|“nn”|“nn$”|“np”
< Pre >::=< Mod > | < Pre >< Mod >
< NN >::= “nn”|“np”|“nns”

where “jj” means adjective, “nn” means noun, “np” means proper noun, “nn$” means possessive noun, “nns” means plural noun, and “in” means preposition. We use TreeTagger[10] for morphological analysis of these English texts.

Next, we examine the frequency of noun phrases in the datasets. Frequent noun phrases are commonly used in many of the datasets. Therefore, the phrases are important. Such phrases would have already extracted manually as tags. We look at the discriminative power of noun phrases in the datasets. The simplest metric for the discriminability is document frequency of a noun phrase. In the case of OGD, the document frequency of a noun phrase is defined as the number of datasets that the phrase appears. However, there are a lot of infrequent phrases with the same document frequency based on Zipf’s law. We can not distinguish the infrequent phrases from the view point of the discriminability.

We extract particular noun phrases for each dataset by modifying the term weighting method for noun phrases proposed in (Yamada et al., 2018)[11]. It is natural that the frequency of words within a noun phrase np is higher than the frequency of np itself in a set of datasets because np includes the words. However, if a noun phrase does not satisfy this natural assumption, then the words mostly appear only within the noun phrase. That is, the words are related to only the noun phrase. Therefore, the noun phrase is considered to be particular in the datasets.

Let $np = w_1\_w_2\_\cdots\_w_m$ be a noun phrase which matches the above pattern in a dataset, where “_” is a space, and $w_1, w_2, \cdots, w_m$ are words. If the words $w_1, w_2, \cdots, w_m$ appear within only the noun phrase np, then the words are strongly associated with only the phrase. In addition, the frequency of $w_1, w_2, \cdots, w_m$ and np is the same.

The average of the difference between the frequency of noun phrase $np = w_1\_w_2\_\cdots\_w_m$ and the

[8] http://scikit-learn.org/stable/

[9] Words in a tag are joined by the character “-” on Data.gov.

[10] http://www.cis.uni-muenchen.de/ schmid/tools/ TreeTagger/

[11] In (Yamada et al., 2018), a noun phrase is defined as nouns appearing successively in a text.

frequency of words $w_1, w_2, \ldots, w_m$ is defined as follows:

$$\mathrm{diff}(np) = \frac{1}{m}\sum_{i=1}^{m}(\mathrm{freq}(w_i) - \mathrm{freq}(np)) \qquad (1)$$

where freq($*$) is the total frequency of $*$ in all datasets. Clearly, $\mathrm{freq}(w_i) \geq \mathrm{freq}(np)$. The smaller the value of diff(np) is, the more particular the noun phrase is.

We assume that $np = w_1\_w_2\_w_3$. If $\mathrm{freq}(np) = \mathrm{freq}(w_1) = \mathrm{freq}(w_2) = \mathrm{freq}(w_3) = 10$, then $\mathrm{diff}(np) = \frac{1}{3}\{(10-10)+(10-10)+(10-10)\} = 0$. If $\mathrm{freq}(w_1) = 20$, $\mathrm{freq}(w_2) = 50$, and $\mathrm{freq}(w_3) = 110$, then $\mathrm{diff}(np) = \frac{1}{3}\{(20-10)+(50-10)+(110-10)\} = 50$.

The proposed particular noun phrase extraction procedure for OGD follows the steps below. Given a set D of datasets with the title and the description of a dataset and its resources, the procedure counts the total frequency of words and noun phrases in D. Then, it calculates the formula 1 of all noun phrases in each dataset. Finally, it sorts the noun phrases by diff(np) for each dataset and outputs the sorted noun phrases.

Noun phrases which are output by the procedure are not always infrequent in the datasets. However, we can consider that noun phrases with a small value of diff(np) are particular even if the phrases appear in some datasets. As described in Section 1, it is desirable that a tag recommendation system outputs candidate tags from various viewpoints and publishers of OGD select appropriate tags from the candidates. This section proposed a new viewpoint about the discriminability of tags.

6 EXPERIMENT

This section shows experiments of multi-label classification for OGD on Data.gov by using the support vector machine, the random forest and multinomial naive Bayes methods. This section also shows noun phrases extracted by the method of the previous section.

6.1 Multi-label Classification

6.1.1 Dataset

We collected 196,587 datasets of Data.gov on September 12, 2017. The total number of tags in the datasets is 57,430. From datasets with tags, the training data are 90% of the datasets selected randomly, and the test data are the rest of the datasets. In advance, we eliminated tags which appear less than twenty times in the training data because it is difficult to predict infrequent tags in the training data and the learning time is too long. After the elimination, the number of tags in the training data is 2,917. We also eliminated tags that appear in the test data but do not appear in the training data because it is impossible to predict the tags. Table 4 shows the statistics of the training and test data. Each dataset in the training and the test data is vectorized by using the term frequency of nouns in the title of each dataset.

Table 4: Training data and test data in the experiment of multi-label classification.

# of all datasets	68,832
# of datasets in training data	62,203
# of datasets in test data	6,629
# of tags in training data	2,917
# of words in training data	26,187

6.1.2 Evaluation Measures

We use the micro f measure, the macro f measure and the average precision to evaluate the tags predicted by a classifier (Tsoumakas et al., 2010). Let $T = \{t_1, t_2, \ldots, t_n\}$ be a set of tags in the training data. First, we define the recall and the precision of tag t_i in the datasets as follows:

$$recall_{t_i} = \frac{TP_{t_i}}{TP_{t_i} + FN_{t_i}},$$

$$precision_{t_i} = \frac{TP_{t_i}}{TP_{t_i} + FP_{t_i}},$$

where TP_{t_i} denotes the number of examples in the test data with correctly predicted tag t_i, FN_{t_i} is the number of examples that have t_i but are not predicted t_i by a classifier, and FP_{t_i} is the number of examples that do not have t_i but are predicted t_i by a classifier.

Then, the f measure of t_i is defined as follows:

$$f\_measure_{t_i} = \frac{2 \times recall_{t_i} \times precision_{t_i}}{recall_{t_i} + precision_{t_i}}.$$

We describe an exception, which is that a tag t_i appears in the training data but does not appear in the test data, to calculate the f measure. In this case, if a classifier does not predict t_i in all examples of the test data, then both $recall_{t_i}$ and $precision_{t_i}$ are 1. Therefore, $f\_measure_{t_i}$ is 1. If it predicts t_i in any of the examples, then $recall_{t_i}$ is 1 but $precision_{t_i}$ is 0. Therefore, $f\_measure_{t_i}$ is 0.

The macro f measure of all tags is defined as follow:

$$macro\_f\_measure = \frac{1}{n}\sum_{i=1}^{n} f\_measure_{t_i}$$

where n is the cardinal number of T.

We define the micro recall and precision of all tags as follows:

$$micro\_recall = \frac{\sum_{i=1}^{n} TP_{t_i}}{\sum_{i=1}^{n} TP_{t_i} + \sum_{i=1}^{n} FN_{t_i}},$$

$$micro\_precision = \frac{\sum_{i=1}^{n} TP_{t_i}}{\sum_{i=1}^{n} TP_{t_i} + \sum_{i=1}^{n} FP_{t_i}}.$$

Then, the micro f measure for all tags is defined as follows:

$$micro\_f\_measure = \frac{2 \times micro\_recall \times micro\_precision}{micro\_recall + micro\_precision}.$$

The average precision evaluates the rank of tags predicted by the classifiers. First, we define precision at rank k as follows:

$$\text{precision}(k) = \frac{1}{k}\sum_{i=1}^{k} r_i$$

where $r_i = 1$ if a tag at rank i is one of the tags of an example in the test data; otherwise, $r_i = 0$. Then, the average precision is defined as follows:

$$average\_precision = \frac{1}{|E|}\sum_{i=1}^{|E|}\frac{1}{|Y_i|}\sum_{k=1}^{n} r_k \times \text{precision}(k)$$

where $|E|$ is the number of examples in the test data, $|Y_i|$ is the number of tags of the i-th example in the test data, and n is the cardinal number of T.

6.1.3 Result

We implemented the function predict_prob() in scikit-learn to order tags for an example in the test data when using the random forest and multinomial naive Bayes methods. The function predict_prob() returns probability estimates. In the experiments on the micro and the macro f measure, we see tags for which the probability estimate is greater than 0.5 as predicted tags. The support vector machine predicts whether each tag should be assigned to the example using the function pred() in scikit-learn. Therefore, the predicted tags are not ordered. This is the reason that the cell of the average precision of the support vector machine is blank in Table 5.

Table 5 shows the results for each method. The support vector machine provided the best results among the methods. The results for the random forest are approximately the same as those for the support vector machine and are better than the results for the multinomial naive Bayes method. Roughly speaking, the random forest and the support vector machine can correctly predict three out of four tags that are assigned to an example. Moreover, when they predict four tags for an example, three out of the four tags are correct.

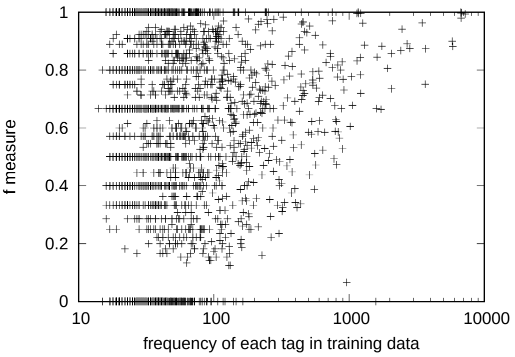


Figure 4: F measure for each frequency of tags in training data using the random forest.

Figure 4 shows the f measure for each frequency of tags in the training data using the random forest. The horizontal axis is plotted on a log scale. Many of the frequent tags in the training data have a high f measure. On the other hand, some tags for which the f measure is 0 appear fewer than 100 times in the training data.

As shown in Table 5, the macro f measure of all methods is lower than the micro f measure. The macro f measure is the average of the f measures of all tags. Therefore, if a tag is infrequent in the test data and the f measure of the tag is 0, the macro f measure is decreased. On the other hand, the micro f measure is not significantly affected.

Figure 4 and Table 5 show that the f measure of some infrequent tags in the training data is low, even though we eliminated tags that appear less than twenty times in the data in advance. This shows that predicting infrequent tags using the multi-label classification is difficult.

There are two different approaches to deal with this problem. The first approach is to improve an algorithm of multi-label classification for infrequent labels. Jain et al. proposed PfastreXML, which is a multi-label classification algorithm for predicting infrequent labels (Jain et al., 2016). Another approach is to re-sample examples in the training data(Haixiang et al., 2017). For example, over-sampling increases examples of infrequent labels. SMOTE (Chawla et al., 2002) selects k nearest neighbors of an example of an infrequent label and then makes a new example between the neighbors and the example. Using these approaches to recommend tags of OGD is the subject of a future study.

The average precision of the random forest method is 0.766. For example, if an example in the test data has two tags, which are predicted to have ranks 1 and 4, then the average precision is 0.750. If an example has seven tags, which are predicted to have ranks ranging from 2 to 8, then the average pre-

Table 5: Results for multi-label classification methods.

	micro f measure	macro f measure	average precision
support vector machine	**0.775**	**0.597**	—
random forest	0.763	0.538	**0.766**
multinomial naive Bayes	0.597	0.244	0.619

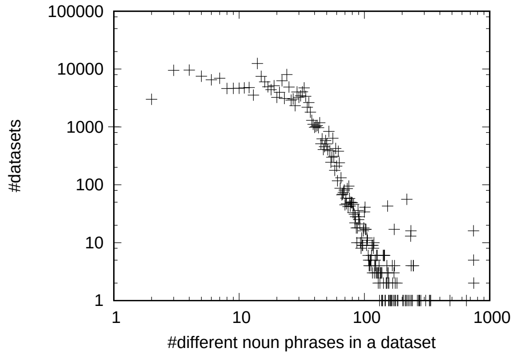


Figure 5: Number of different noun phrases extracted by our method from a dataset and the number of datasets with each number of the different phrases.

cision is 0.754.

We consider developing a tag recommendation system for publishers of OGD. After inputting the title of the OGD into the system, the system displays approximately twenty predicted tags, each of which has a degree of recommendation. The publisher then selects appropriate tags from the predicted tags. Based on the experiments, we can reasonably conclude that the random forest provides good results because correct tags are ranked at the top of prediction.

6.2 Particular Noun Phrase Extraction

We extracted a total of 3,912,648 noun phrases from the title and description of 196,587 datasets and their resources. The number of different noun phrases extracted was 645,183. Figure 5 shows the number of different noun phrases extracted by the proposed method from a dataset and the number of datasets with each number of different phrases. The numbers of datasets that were not extracted noun phrases and that which had only one noun phrase are 277 and 1,003, respectively. The maximum and average numbers of different noun phrases extracted from a dataset are 745 and 19.9, respectively.

A total of 3,448 different noun phrases out of the top noun phrases extracted by the proposed method from datasets are included in the 57,430 tags of Data.gov. Figure 6 shows the frequency of tags of Data.gov that are the same as noun phrases extracted by the proposed method. The tags are sorted in ascending order of frequency. The frequency as tags of 2,827 of the 3,448 noun phrases is less than 10. Therefore, the noun phrase extraction method proposed in Section 5 extracted noun phrases corresponding to infrequent tags on Data.gov.

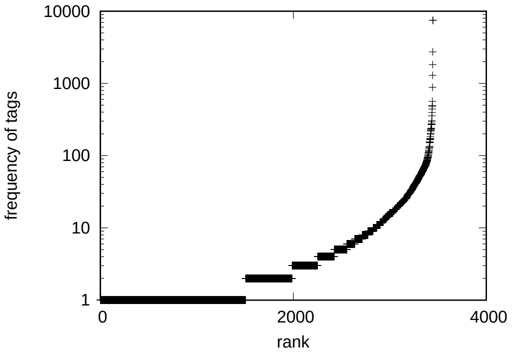


Figure 6: Frequency of tags of Data.gov which are the same as noun phrases extracted by our method.

Since the frequency of prepositions is high, noun phrases with prepositions tend to increase the value in formula 1. We should have excepted the frequency of them from the calculation of the formula 1.

We proposed a method by which to calculate the discriminative power of noun phrases in a new light. It is desirable and natural that there are various viewpoints with respect to appropriate tags for OGD, such as the popularity and the coverage of tags, as described in Section 2. Again, suppose that we develop a tag recommendation system that displays candidate tags. The system should display candidate tags extracted based on various viewpoints, and publishers of OGD can then select correct tags from among the candidates.

7 CONCLUSION

This paper examined tag recommendations for open government data. The two different approaches are multi-label classification and particular noun phrase extraction. We applied three multi-label classification methods, the support vector machine, the random forest and the multinomial naive Bayes. Although the random forest received a good result for a tag recommendation system, further improvement of the accuracy of prediction is important. Our particular noun phrase extraction method extracted some noun phrases which are the same as infrequent tags on Data.gov.

Our future work is to recommend tags which are infrequent in training data. In our current experiments, we eliminated tags which appear fewer than twenty times in the datasets in advance. Nevertheless, the accuracy of infrequent tags in training data was low. Infrequent tags tend to express the concrete content of OGD. Therefore, infrequent tags are important to understand OGD without actually reading the data.

Future work includes the development of a Web system which recommends tags when users input the OGD information. The system displays candidate tags output by multi-label classification and ones extracted by various viewpoints including our particular noun phrase extraction.

ACKNOWLEDGEMENTS

This work was partially supported by JSPS KAKENHI Grant Numbers 15K00426.

REFERENCES

Babbar, R. and Schölkopf, B. (2017). Dismec: Distributed sparse machines for extreme multi-label classification. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, pages 721–729. ACM.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority oversampling technique. *J. Artif. Int. Res.*, 16(1):321–357.

Corrêa, A. S. and Zander, P.-O. (2017). Unleashing tabular content to open data: A survey on pdf table extraction methods and tools. In *Proceedings of the 18th Annual International Conference on Digital Government Research*, pages 54–63. ACM.

Figueiredo, F., Pinto, H., Belém, F., Aleida, J., Gonçalves, M., Fernandes, D., and Moura, E. (2013). Assessing the quality of textual features in social media. *Information Processing and Management*, 49(1):222–247.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems With Applications*, 73:220–239.

Herrera, F., Charte, F., Rivera, A. J., and del Jesus, M. J. (2016). *Multilabel Classification: Problem Analysis, Metrics and Techniques*. Springer Publishing Company, Incorporated, 1st edition.

Jain, H., Prabhu, Y., and Varma, M. (2016). Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 935–944. ACM.

Kang, N., Doornenbal, M. A., and Schijvenaars, R. J. A. (2015). Elsevier journal finder: Recommending journals for your paper. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 261–264. ACM.

Köster, V. and Suárez, G. (2016). Open data for development: Experience of uruguay. In *Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance*, pages 207–210. ACM.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Martins, E. F., Belém, F. M., Almeida, J. M., and Gonçalves, M. A. (2016). On cold start for associative tag recommendation. *J. Assoc. Inf. Sci. Technol.*, 67(1):83–105.

Oliveira, M. I. S., de Oliveira, H. R., Oliveira, L. A., and Lóscio, B. F. (2016). Open government data portals analysis: The brazilian case. In *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research*, pages 415–424. ACM.

Prabhu, Y. and Varma, M. (2014). Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 263–272. ACM.

Ribeiro, I. S., Santos, R. L., Gonçalves, M. A., and Laender, A. H. (2015). On tag recommendation for expertise profiling: A case study in the scientific domain. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pages 189–198. ACM.

Tambouris, E., Kalampokis, E., and Tarabanis, K. (2017). Visualizing linked open statistical data to support public administration. In *Proceedings of the 18th Annual International Conference on Digital Government Research*, pages 149–154. ACM.

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2010). Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685.

Vasa, M. and Tamilselvam, S. (2014). Building apps with open data in india: An experience. In *Proceedings of the 1st International Workshop on Inclusive Web Programming - Programming on the Web with Open Data for Societal Applications*, pages 1–7. ACM.

Venetis, P., Koutrika, G., and Garcia-Molina, H. (2011). On the selection of tags for tag clouds. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pages 835–844. ACM.

Xu, C., Tao, D., and Xu, C. (2016). Robust extreme multi-label learning. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1275–1284. ACM.

Yamada, Y., Himeno, Y., and Nakatoh, T. (2018). Weighting of noun phrases based on local frequency of nouns. In *Recent Advances on Soft Computing and Data Mining - Proceedings of the 3rd International Conference on Soft Computing and Data Mining*, pages 436–445. Springer.