A Keyword Recommendation Method Using CorKeD Words and Its Application to Earth Science Data

Youichi Ishida(✉), Toshiyuki Shimizu, and Masatoshi Yoshikawa

Graduate School of Informatics, Kyoto University,
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
{yishida,tshimizu,yoshikawa}@db.soc.i.kyoto-u.ac.jp

Abstract. In various research domains, data providers themselves annotate their own data with keywords from a controlled vocabulary. However, since selecting keywords requires extensive knowledge of the domain and the controlled vocabulary, even data providers have difficulty in selecting appropriate keywords from the vocabulary. Therefore, we propose a method for recommending relevant keywords in a controlled vocabulary to data providers. We focus on a keyword definition, and calculate the similarity between an abstract text of data and the keyword definition. Moreover, considering that there are unnecessary words in the calculation, we extract CorKeD (Corpus-based Keyword Decisive) words from a target domain corpus so that we can measure the similarity appropriately. We conduct an experiment on earth science data, and verify the effectiveness of extracting the CorKeD words, which are the terms that better characterize the domain.

Keywords: CorKeD words · Domain corpus · Controlled vocabulary · Keyword definition · Abstract text · Earth science

1 Introduction

Due to the rapid advancement in information technologies and the remarkable dissemination of social media in recent years, diverse and vast amount of data has been generated. To classify those data accurately, and to obtain the right information quickly, it is effective to annotate them with metadata. Recently, people have annotated various data, such as user generated content(images, videos, web page bookmarks, and so on), academic research papers, earth science data.

As examples of metadata, there may be mentioned title, creation date, author, abstract text, keyword. We focus on keywords among these metadata. Annotation keywords are used to support search, browse and classification of various data. We consider that there are mainly two ways to add keywords to data. One way is that users themselves annotate various data with keywords [5,8,9], while the other is that data providers themselves add keywords to their own data in a research domain [2,3,6,7]. In the former case, since many general

G. Zuccon et al. (Eds.): AIRS 2015, LNCS 9460, pp. 96–108, 2015.
DOI: 10.1007/978-3-319-28940-3_8

users continuously add keywords to one data, there is an advantage that a set of keywords added to the data finally converges to useful one. Yet, in the latter case, since a data provider is the only person that annotates the data, the utilization value of the added keyword set depends on only the data provider. In addition, in many cases, they restrict keywords to add by using a controlled vocabulary of each domain. By this restriction, they can eliminate noise and omission in retrieval of data which are caused by changes in word form and orthographic variation. However, to select suitable keywords from a controlled vocabulary, it is required to gain extensive knowledge of the research domain and the large-scale controlled vocabulary which typically includes thousands of keywords. Therefore, even a data provider has difficulty in picking out keywords suitably from the vocabulary. In this paper, we focus on this latter case and propose a method for recommending suitable keywords in a controlled vocabulary on various research domains.

Abstract text of the dataset D8NDVI_J managed by DIAS-P
This dataset contains the daily value of the Normalize Difference Vegetation Index (NDVI) from 1982 to 2000 over the terrestrial areas of the Japan Islands that was derived from Pathfinder AVHRR Land (PAL) dataset. The horizontal resolution is 8 x 8 km. To reduce the cloud contamination, the original daily NDVI was temporally smoothed by Temporal Window Operation (TWO) method.

Keyword definition about ACID_RAIN in GCMD Science Keywords
Definition: Rain having a pH lower than 5.6, representing the pH of natural rainwater; the increased acidity is usually due to the presence of sulfuric acid and/or nitric acid, often attributed to anthropogenic sources.

Fig. 1. An example of an abstract text and a keyword definition

In this paper, we make use of an abstract text in metadata. In general, data providers annotate data with an abstract text describing the content of the data. For example, in the case of earth science data, the information of observation items, an observation method, usage of the data and so on are described in the abstract text. Researches on keyword recommendation [2,8] often propose a method for recommending keywords which are added to similar data to a target data in such text information. Yet, metadata quality is actually a pressing problem in the metadata portal called Europeana[1]. When, as in Europeana, the amount and quality of existing metadata set is insufficient in a metadata portal, their methods do not seem to be effective. In this paper, to propose a method which does not depend on the existing metadata set other than a target data, we utilize definition information given to each keyword itself as well as an abstract text given to the target data itself. In most cases, each keyword in

[1] http://pro.europeana.eu/publication/metadata-quality-task-force-report.

a controlled vocabulary has a keyword definition explaining the meaning of it. Fig. 1 shows an example of the keyword definition in the controlled vocabulary called GCMD(Global Change Master Directory) Science Keywords [1] and an abstract text of a dataset managed by the metadata portal called DIAS-P(Data Integration Analysis System Program)[2], which is managed in Japan.

As an initial attempt, we calculate the similarity between an abstract text of data and keyword definitions, and recommend keywords which have high degree of similarity. However, considering that not all the words in those documents contribute for deciding which keywords to recommend, we extract "CorKeD (Corpus-based Keyword Decisive) words" so that we can measure the similarity appropriately. We first consider that decisive words for keyword recommendation are domain specific words in a target domain, and extract the domain specific words by analyzing the occurrence tendency of each word between the target domain corpus and the other domain corpora. Secondly, by further refining useful words for recommendation from the domain specific word list, we restrict the words to use in the calculation. We call the restricted words "CorKeD words". Some researches have been conducted on extracting domain specific words [4, 11,12], but we moreover extract from the domain specific word list the CorKeD words, which are the decisive words for keyword recommendation.

Our proposed method can be applied to various research domains, and this paper deals with earth science among such domains. Owing to the recent progress in earth observation technologies, the total amount of earth science data has explosively increased in various domains such as atmosphere, ocean, climate. Therefore, it is required to manage metadata portals so that those metadata can be properly handled. For instance, the metadata portal called GCMD[3] provides a search function for searching various metadata and manages the controlled vocabulary such as GCMD Science Keywords. As mentioned above, there is also a project called DIAS-P in Japan. DIAS-P is aiming to build a database which promotes the interoperability of heterogenous data collected from multiple fields, places, times.

A keyword of metadata in earth science is added to a dataset by selecting keywords relevant to the dataset from a controlled vocabulary. For example, a dataset on rainfall observations is most likely to be annotated with the keyword "PRECIPITATION AMOUNT". In DIAS-P, data providers themselves annotate their provided datasets manually with metadata such as keywords. Therefore, it is hard to select suitable keywords from a large-scale controlled vocabulary. As a result of investigating metadata in DIAS-P, there are actually many poorly annotated datasets. In this paper, we conduct an experiment on datasets managed by DIAS-P, and verify the effectiveness of our method.

Contributions. This study makes three contributions as follows:

1. Unlike the previous methods, we propose the method which does not depend on the quality of the other existing metadata set. We make use of not only an

[2] http://www.diasjp.net/.
[3] http://gcmd.nasa.gov/.

abstract text of metadata but also keyword definitions, which are associated with each keyword in a controlled vocabulary.
2. We restrict the words to use in the calculation by extracting domain specific words and moreover selecting the CorKeD words, which contribute for deciding suitable keywords. Some previous researches [4,11,12] only extract domain specific words.
3. We conduct an experiment on real datasets managed by DIAS-P and verify the effectiveness of extracting the CorKeD words.

2 Related Works

In recent years, researches on keyword recommendation based on the system of folksonomy have attracted attention [5,8,9]. However, most of those researches focus on personalized keyword recommendation utilizing a user's history. In such works, it is common for the users themselves to annotate multiple data arbitrarily with keywords without using a controlled vocabulary. On the other hand, for a highly specialized data such as a research data, it is not users but data providers that add keywords to those data with a controlled vocabulary. Since, in this case, sufficient information of their history is unavailable, content-based methods are considered to be useful. This section presents some related works which propose content-based methods for keyword recommendation.

We describe some researches on supporting social tagging with a content-based method [8,9]. In social bookmarking services such as Delicious[4], Lu et al. [8] propose a method of recommending suitable keywords for a webpage lacking tag information. Their approach calculates an assignment probability of each tag for a webpage, based on how much each tag is appearing in a set of tags added to the webpage and the similarity between the webpages. However, this work presupposes that multiple users annotate one webpage with the same tags as the other users do. Hence, this method cannot be applied to highly specialized research domains because in such domains only the data providers add keywords. They also calculate how trustworthy the webpage is, based on the total number of tags added to the webpage. Yet, in research domains, the number of keywords added to data has nothing to do with the reliability of the data. Belem et al. [9] propose a formula to calculate the relevance of each tag for a resource with learning-to-rank technologies, combining various indicators such as tag co-occurrence, descriptive power, term predictability. However, this work does not use a controlled vocabulary, and extracts recommended keywords from the whole terms of documents.

As a research of keyword recommendation for earth science data, Tuarob et al. [2] propose a method for recommending tags for data missing tag information from a controlled vocabulary. They create the feature vector of each dataset from the text information in the metadata, and recommend tags which are added to similar datasets by calculating the similarity between the feature vectors. Each

[4] https://delicious.com/.

document is represented with either a TF-IDF vector [13] or a probability distribution of LDA (Latent Dirichlet Allocation)[14]. However, when the amount and quality of existing metadata set is insufficient in a metadata portal, their method does not seem to be effective. In contrast, we propose a method which does not depend on an existing metadata set, and our proposed method can be applied to a new controlled vocabulary which has not been used much. Shimizu et al. [3] suggests the 14 keywords which represent categories of earth science with Labeled LDA [15]. They define the 14 keywords as labels, learning correspondence between an abstract text of a dataset and added keywords. Then, they recommend suitable keywords by applying the learning results to a target dataset. As in this study, when the number of the labels is small, Labeled LDA is useful for recommendation. Yet, it is very hard to prepare enough training data to define thousands of keywords as labels.

We also introduce some works on supporting an annotation for an academic research paper. Chernyak [6] propose a method for recommending topics from the controlled vocabulary called ACM Computing Classification System. Using self-learning methods such as TF-IDF, BM25, annotated suffix tree, they calculate the similarity between the topics and each paper's abstract. Santos et al. [7] address the problem of multi-label classification for research papers with machine learnings such as SVM, KNN, naive Bayes classification. Although the studies of annotations for research papers are different from earth science in that their studies can guess suitable keywords from reference information, they have much in common with our study in that both studies need a controlled vocabulary and in that keywords are added by a specific person such as an author. Our proposed method can be applicable to the annotation of research papers.

3 Proposed Method

In the following, we explain the case of applying our proposed method to earth science data. In this paper, we made use of an abstract text of a dataset in metadata. By viewing the abstract text, users can roughly comprehend the content of the dataset. We give an example of added keywords with GCMD Science Keywords in Fig. 2. As shown in Fig. 2, keywords are hierarchically managed in GCMD Science Keywords, but in this paper, we propose the method where we do not take the hierarchical structure into account so that our proposed method can be applied to a controlled vocabulary without hierarchical model.

At the beginning, as a method of simple string matching, we extracted keywords from an abstract text of a dataset in DIAS-P by matching each keyword in GCMD Science Keywords. However, as a result of applying the method, we could only recommend the average of about 2.7 keywords. Therefore, we decided to utilize implicit information such as a keyword definition as well as explicit information such as a keyword name. As an initial attempt, we considered that we recommend keywords which have high degree of similarity between an abstract text and the keyword definitions. Moreover, considering that there are unnecessary words in the calculation, we extracted the CorKeD words, which are the

- Atmosphere > Atmospheric Water Vapor > Humidity
- Atmosphere > Atmospheric Water Vapor > Water Vapor
- Atmosphere > Precipitation > Precipitation Amount
- Oceans > Oceans Temperature > Sea Surface Temperature
- Cryosphere > Snow/Ice > Snow Water Equivalent
- Land Surface > Soils > Soil Moisture/Water Content

Fig. 2. Keywords added to Aqua AMSR-E dataset managed by DIAS-P

decisive words for keyword recommendation. We first created the domain specific word list from the domain corpus of earth science, and then by further refining useful words for recommendation from the list, we extracting the CorKeD words to analyze in calculating the similarity. We preprocess the abstract texts and the keyword definitions by removing stopwords, and stemming each word.

3.1 Definition of a Domain Specific Word of Earth Science

As Kubo et al. [4] points out, a domain specific word in certain target domain is considered as a word which has a higher appearance frequency in the target domain than in the other domains. In other words, we can define a domain specific word of earth science as a word which appears at a higher frequency in a corpus of earth science. In this paper, as the other domains other than earth science, we used biology, chemistry and physics, which belong to the same natural science. The reason why we used those three domains is because we considered that we can extract the domain specific words of earth science more properly by comparing with those domains than with non-natural science domains such as the humanities or social science.

Corpus of Each Domain. To compare among the domains, we must construct a corpus of each domain. We created a corpus of earth science from the presentation summaries in 2013 Fall Meeting held by AGU(American Geophysical Union)[5], which is the organization of earth science. We obtained approximately 6 million words from 20028 summaries. As corpora of the other domains, we used summaries of papers published in journals of each domain[6]. In addition, we equalized a corpus size of each other domain at about 200 thousand words.

The Method of Creating the Domain Specific Word List. To construct the domain specific word list of earth science, we need to compare the relative

[5] http://sites.agu.org/.

[6] Chemistry : Journal of the American Chemical Society
Physics : The European physical journal
Biology : International journal of biological sciences, Journal of evolutionary biology.

frequency for each one word in the corpus of earth science between earth science and each other domain. This study utilized a formula called DP(the Difference between Population Proportions) that Kubo et al. [4] propose. This formula is based on 2-sample test for equality of proportions in statistics. It is described in detail below.

$$DP_d(t) = \frac{\frac{f_0(t)}{W_0} - \frac{f_d(t)}{W_d}}{\sqrt{\pi_d(t)(1-\pi_d(t))\left(\frac{1}{W_0}+\frac{1}{W_d}\right)}}, \quad \pi_d(t) = \frac{f_0(t)+f_d(t)}{W_0+W_d} \tag{1}$$

Let $f_0(t)$ and $f_d(t)$ be the appearance frequency of word t in the corpus of earth science and the other domain d, respectively. W_0 and W_d is the total number of words in the corpus of earth science and the other domain d, respectively. $\pi_d(t)$ is the ratio of the appearance frequency of word t in the both corpora, and the set D consists of {biology, chemistry, physics}. $DP_d(t)$ represents the relative frequency of word t in comparing between earth science and the other domain $d \in D$. This $DP_d(t)$ follows a normal distribution. Then, by Eq. 2, we calculated the average of the relative frequency obtained by comparing with each other domain. $|D|$ is the size of D, that is, $|D| = 3$.

$$w(t) = \frac{\sum_{d \in D} DP_d(t)}{|D|} \tag{2}$$

When $w(t)$ was positive as calculation results, we defined the word t as a domain specific word. Table 1 represents the top 10 scores of $w(t)$. Certainly, all of the highly ranked words can be considered as domain specific words of earth science.

Table 1. The top 10 of $w(t)$'s score

	word t	$w(t)$
1	data	27.89
2	model	24.46
3	climat	24.26
4	water	20.18
5	region	20.03
6	soil	19.88
7	atmospher	19.00
8	fault	18.18
9	ic	18.15
10	event	18.07

However, it seems that ranking highly words such as "data","model", "region" have little information for deciding which keywords to recommend. Therefore, we furthermore discussed a method for refining decisive words for recommendation from the domain specific word list.

3.2 Whether a Word Contributes for Deciding Keywords

In earth science, there are further subdivided domains, such as atmosphere, agriculture, oceans. In the case of a word which contributes for deciding keywords, we considered that there is a bias in the appearance frequency for such words among the subdivided domains. On the other hand, in the case of a word which has little information for the decision, we considered that such words appear without depending on the subdivided domains. For instance, the word "climat" is likely to appear disproportionately in the subdivided domain "atmosphere", while the word "data" tends to appear at about the same frequency among the subdivided domains. Thus, by quantifying the bias of the frequency distribution for each word among the subdivided domains, we can judge whether the word contributes for deciding keywords or not.

In this paper, as the subdivided domains, we used the 49 categories taken as a classification axis of AGU index terms[7], which is a controlled vocabulary managed by AGU introduced in Sect. 3.1. Furthermore, we utilized χ square value, which is generally used as a method for quantifying a bias of a distribution. χ square value shows difference between an observed and an expected value. As the observed value, we calculated the document frequency(DF) of each word in the summaries of AGU. Besides, as the expected value, we calculated the DF of the word by assuming that the word appears at about the same proportion among the subdivided domains. It is described in detail below.

$$\chi^2(t) = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \quad E_i = S_i \times \frac{S_t}{S} \tag{3}$$

S represents the total number of the summaries in AGU 2013 Fall Meeting, which is 20028. n represents the number of the subdivided domains, that is, $n = 49$. Let O_i be the observed value and let E_i be the expected value in i^{th} subdivided domain. S_i is the number of the summaries in ith subdivided domain and S_t is the total number of the summaries containing word t. In addition, we considered that a word which has little information is highly likely to appear in any summaries, and calculated χ square value for each word contained in the top 0.5 % of DF values. Tables 2 and 3 show the part of the calculation results.

In Table 2, χ square value for each word which is likely to contribute for the decision shows a relatively large value. This indicates that these words appear disproportionately in some subdivided domains. Conversely, in Table 3, χ square value for each word which has little information shows a relatively small

[7] http://abstractsearch.agu.org/keywords.

Table 2. CorKeD words

word t	χ square value
climat	5735.9
water	3678.5
soil	3439.1
atmospher	4375.0
temperatur	1729.7

Table 3. Not CorKeD words

word t	χ square value
data	660.8
model	695.1
region	801.8
time	282.3
base	352.5

value. This shows that these words appear without depending on the subdivided domains. From the result of a preliminary investigation, we set a threshold 1700, and by eliminating words less than the threshold from the domain specific word list, we finally created a set of CorKeD words, which is used in the calculation.

3.3 How to Calculate the Similarity

Using only the CorKeD words as previously described, we calculated the similarity between an abstract text and the keyword definitions. We represented a query abstract A_i by a feature vector $\boldsymbol{DA}(A_i)$. Let Cs be the set of the CorKeD words. When a word is included in both the query abstract A_i and Cs, the word's element of $\boldsymbol{DA}(A_i)$ is 1. Conversely, when a word is not included in A_i or Cs, the word's element of $\boldsymbol{DA}(A_i)$ is 0.

$$t_{ij} = \begin{cases} 1 & (t_{ij} \in A_i \wedge t_{ij} \in Cs) \\ 0 & (otherwise) \end{cases} \tag{4}$$

$$\boldsymbol{DA}(A_i) = \{t_{i1}, t_{i2}, \cdots, t_{im}\} \tag{5}$$

We represented a keyword definition D_j by a feature vector $\boldsymbol{KD}(D_j, Cl)$, and each element of the feature vector is TF-IDF value for each word. On this occasion, we used LRTF(Length Regularized TF) introduced in [10] as TF(Term Frequency). These are described in detail below.

$$LRTF(t, D_j) = TF(t, D_j) \times \log_2 \left(1 + \frac{ADL(Cl)}{len(D_j)}\right) \tag{6}$$

$$IDF(t, Cl) = \log_2 \left(\frac{|Cl|}{DF(t, Cl)}\right) + 1 \tag{7}$$

$$KD(t, D_j, Cl) = LRTF(t, D_j) \times IDF(t, Cl) \tag{8}$$

$$\boldsymbol{KD}(D_j, Cl) = \{KD(t_1, D_j, Cl), \cdots, KD(t_n, D_j, Cl)\} \tag{9}$$

Let Cl be the keyword definitions set, and let $|Cl|$ be the number of the keywords. In addition, $len(D_j)$ is the length of the keyword definition D_j, $ADL(Cl)$ is the average of $len(D_j)$, and $TF(t, D_j)$ is the appearance frequency of word t in D_j. LRTF is a formula which normalizes TF value, considering the proportion between $len(D_j)$ and $ADL(Cl)$. We considered that LRTF is appropriate to this

situation, where an abstract text is regarded as a query, because [10] says that LRTF is useful to a long query composed of more than 5 words. IDF (Inverse Document Frequency) value was calculated by the most standard formula, in which $|Cl|$ is divided by $DF(t, Cl)$. In this paper, we calculated the cosine similarity between the above two feature vectors using only the CorKeD words, and recommended keywords in descending order of cosine similarity values.

$$CosineSim(\boldsymbol{DA}(A_i), \boldsymbol{KD}(D_j, Cl)) = \frac{\boldsymbol{DA}(A_i) \cdot \boldsymbol{KD}(D_j, Cl)}{\|\boldsymbol{DA}(A_i)\| \times \|\boldsymbol{KD}(D_j, Cl)\|} \quad (10)$$

4 Evaluation

To verify the effectiveness of our proposed method, we conducted an experiment on 20 datasets managed by DIAS-P. We submitted the recommended keywords to each data provider to judge whether each recommended keyword is correct or not. We used GCMD Science Keywords as a controlled vocabulary, which includes 2017 keywords. To demonstrate effectiveness of creating the set of CorKeD words, we compare our approach with a method for calculating the similarity in using all words included in the keyword definitions and an abstract text.

4.1 Evaluation Metric

This experiment evaluated precision of top 10 keywords recommended by the two methods. In most cases, when precision is evaluated, recall and F-value are calculated at the same time. However, since it is hard to understand the whole keywords in the large-scale controlled vocabulary, even data providers have difficulty in obtaining perfectly the correct keywords set. Thus, we considered that accurate recall and F-value are difficult to calculate.

4.2 Results

Table 4 shows the average of precisions and each precision evaluated by the two methods. Table 4 indicates that our proposed method outperforms the comparative method. The reason is because we can calculate the similarity more properly by using only the CorKeD words. Table 5 describes an example of recommended keywords for the dataset called "GCOM_W1", whose precision is particularly improved. The correct keywords are shown in bold text. In Table 5, our proposed method can recommend many correct keywords which cannot be recommended by the comparative method. We give an example of the similarity between the dataset and the keyword "DEGREE DAYS". In this case, when we used the comparative method, the words to use in the calculation were "atmospher", "one" , "temperatur", "day", "measur", "degre", whereas by applying our method, we could use only the CorKeD words such as "atmospher", "temperatur", which are useful words for deciding suitable keywords. The reason why the accuracy is

Table 4. The evaluation results of keyword recommendation for each dataset

Dataset ID	The Comparative Method	The Proposed Method
ALOS_AVNIR2	30 %	10 %
ALOS_PALSAR	10 %	10 %
ALOS_PRISM	10 %	10 %
AMY_HARIMAU_WPR_dataset	40 %	30 %
Aqua_AMSR_E	0 %	0 %
AVISO_SLA	20 %	10 %
CEOP_CAMP_Eastern_Siberian_Taiga	10 %	30 %
D8NDVI_J	40 %	40 %
D8NDVI_J	50 %	50 %
DIAS_ODAPv2.1	40 %	50 %
DIAS_ODAPv2.1	40 %	60 %
Fuji_Hokuroku_Flux	30 %	50 %
GCOM_W1	0 %	30 %
Global_map	20 %	10 %
Global_map	40 %	20 %
GPV	0 %	10 %
MAHAPGP	30 %	20 %
MIRAI_CTD	30 %	50 %
MOM_rNP	30 %	40 %
MSST	0 %	20 %
ODA_rNPhigh	30 %	40 %
ODA_rNPhigh	40 %	60 %
SSM_I	0 %	20 %
TRMM_PR	10 %	10 %
Average of precisions	**22.92 %**	**28.33 %**

(Note : When more than two data providers evaluates the same dataset, the precision evaluated by each data provider is described)

improved can be because our method could eliminate the words such as “one”, “day”, “measur”, which have useless information for recommendation.

On the other hand, there are some datasets whose precision decrease. We give as an example the similarity between the dataset “ALOS_AVNIR2” and the keyword “LAND USE”. In this case, when we used the comparative method, the words to use in the calculation were “earth”, “land”, “observ”, “area”, **“use”**. However, although the keyword is included in the abstract text, our method with the CorKeD words eliminated the word “use”, which is in the part of the keyword name. In consequence, it was difficult for our method to recommend words related to “LAND USE”, resulting in low recommendation accuracy. In addition, we can find some examples where, for the same reason, we cannot recommend keywords which are included in the abstract texts. These keywords can be extracted by processing in phrase units. Therefore, in the future task, we

Table 5. The result of recommended keywords for GCOM_W1 dataset

	The Comparative Method	The Proposed Method
1	PLANETARY BOUNDARY LAYER HEIGHT	PLANETARY BOUNDARY LAYER HEIGHT
2	SEA SURFACE HEIGHT	MOISTURE FLUX
3	DEGREE DAYS	**SEA SURFACE TEMPERATURE**
4	STRATOPAUSE	**SEA SURFACE TEMPERATURE INDICES**
5	TROPOSPHERIC/HIGH-LEVEL CLOUDS	SEA SURFACE HEIGHT
6	ALTITUDE	CLOUDS
7	ICE TEMPERATURE	INVERSION HEIGHT
8	DEW POINT TEMPERATURE	**ATMOSPHERIC WATER VAPOR**
9	SENSOR COUNTS	STRATOPAUSE
10	INVERSION HEIGHT	GEOPOTENTIAL HEIGHT

would like to consider the combination of our proposed method and processing in phrase units. For the dataset "Aqua_AMSR_E", the correct keywords are not recommended at all by either of the two methods. This is because this abstract text describes some advantages or features of an observational instrument, not explanation about the contents of the dataset. We consider that the information of the observational instrument is likely to help the keyword recommendation.

5 Conclusions and Future Works

To support keyword annotation for various data of research domains, we proposed the method for recommending keywords in a controlled vocabulary. We utilized each keyword definition itself as well as an abstract text of a target data, and proposed the method which does not depend on the existing metadata set other than a target data. Also, to calculate the similarity more properly, we refined the words by extracting domain specific words and moreover selecting the CorKeD words. In this paper, we conducted the experiment on real datasets managed by DIAS-P, and showed the effectiveness of extracting the CorKeD words.

In the future work, we need to compare our approach with the previous ones such as [2,8], and other recent approaches. In addition, we would like to compare DP[4] with the other measures for calculating the relative frequency of one word, such as self mutual information and log-likelihood ratio. Also, we are interested to use the other controlled vocabularies of earth science, and want to apply our approach to the other domains such as chemistry, biology.

References

1. Olsen, L.M., Major, G., Shein, K., Scialdone, J., Ritz, S., Stevens, T., Morahan, M., Aleman, A., Vogel, R., Leicester, S., Weir, H., Meaux, M., Grebas, S., Solomon, C., Holland, M., Northcutt, T., Restrepo, R.A., Bilodeau, R.: NASA/Global Change Master Directory (GCMD) Earth Science Keywords. Version 8.0.0.0.0 (2013)

2. Tuarob, S., Pouchard, L.C., Giles, C.L.: Automatic tag recommendation for metadata annotation using probabilistic topic modeling. In: JCDL, pp. 239–248 (2013)
3. Shimizu, T., Sueki, T., Yoshikawa, M.: Supporting keyword selection in generating earth science metadata. In: COMPSAC, pp. 603–604 (2013)
4. Kubo, J., Tsuji, K., Sugimoto, S.: Automatic term recognition using the corpora of the different academic areas (in Japanese). J. Jpn Soc. Inf. Knowl. **20**(1), 15–31 (2010)
5. Krestel, R., Fankhauser, P., Nejdl, W.: Latent dirichlet allocation for tag recommendation. In: RecSys, pp. 61–68 (2009)
6. Chernyak, E.: An approach to the problem of annotation of research publications. In: WSDM, pp. 429–434 (2015)
7. Santos, A.P., Rodrigues, F.: Multi-label hierarchical text classification using the acm taxonomy. In: EPIA, pp. 553–564 (2009)
8. Lu, Y.T., Yu, S.I., Chang, T.C., Hsu, J.Y.J.: A content-based method to enhance tag recommendation. In: IJCAI, pp. 2064–2069 (2009)
9. Belem, F., Martins, E., Pontes, T., Almeida, J., Goncalves, M.: Associative tag recommendation exploiting multiple textual features. In: SIGIR, pp. 1033–1042 (2011)
10. Paik, J.H.: A novel TF-IDF weighting scheme for effective ranking. In: SIGIR, pp. 343–352 (2013)
11. Utiyama, M., Chujo, K., Yamamoto, E., Isahara, H.: A comparison of measures for extracting domain-specific lexicons for english education (in Japanese). J. Nat. Lang. Process. **11**(3), 165–197 (2004)
12. Uchimoto, K., Sekine, S., Murata, M., Ozaku, H., Isahara, H.: Term recognition using corpora from different fields. Terminology **6**(2), 233–256 (2001)
13. Salton, G.: Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley, Boston (1989)
14. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
15. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: EMNLP, pp. 248–256 (2009)