International Journal on Digital Libraries (2020) 21:307–327
https://doi.org/10.1007/s00799-020-00279-3

An analysis and comparison of keyword recommendation methods for scientific data

Youichi Ishida[1] · Toshiyuki Shimizu[1] · Masatoshi Yoshikawa[1]

Received: 4 September 2018 / Revised: 15 August 2019 / Accepted: 5 January 2020 / Published online: 7 February 2020

Abstract
To classify and search various kinds of scientific data, it is useful to annotate those data with keywords from a controlled vocabulary. Data providers, such as researchers, annotate their own data with keywords from the provided vocabulary. However, for the selection of suitable keywords, extensive knowledge of both the research domain and the controlled vocabulary is required. Therefore, the annotation of scientific data with keywords from a controlled vocabulary is a time-consuming task for data providers. In this paper, we discuss methods for recommending relevant keywords from a controlled vocabulary for the annotation of scientific data through their metadata. Many previous studies have proposed approaches based on keywords in similar existing metadata; we call this the *indirect method*. However, when the quality of the existing metadata set is insufficient, the indirect method tends to be ineffective. Because the controlled vocabularies for scientific data usually provide definition sentences for each keyword, it is also possible to recommend keywords based on the target metadata and the keyword definitions; we call this the *direct method*. The direct method does not utilize the existing metadata set and therefore is independent of its quality. Also, for the evaluation of keyword recommendation methods, we propose evaluation metrics based on a hierarchical vocabulary structure, which is a distinctive feature of most controlled vocabularies. Using our proposed evaluation metrics, we can evaluate keyword recommendation methods with an emphasis on keywords that are more difficult for data providers to select. In experiments using real earth science datasets, we compare the direct and indirect methods to verify their effectiveness, and observe how the indirect method depends on the quality of the existing metadata set. The results show the importance of metadata quality in recommending keywords.

Keywords Keyword recommendation · Metadata quality · Controlled vocabulary · Keyword definition

1 Introduction

1.1 Background

To accurately classify vast amounts of scientific data and to allow the desired information to be quickly obtained, it is useful to annotate these data with metadata. Metadata are not data in their own right but rather describe information related to data. The annotation of each distinct datum or dataset with semantic information such as metadata can support the understanding of the data themselves and allow relevant data to be quickly extracted from a vast amount of available data. Examples of metadata for a scientific dataset include the title, creation date, author, data format, abstract text and keywords of an entry. Of these types of metadata, we particularly focus on keywords because of their importance. Viewing annotated keywords enables us to roughly understand the content of the corresponding dataset, to determine the associations between related datasets, and to support the searching, browsing and classification of various datasets. Therefore, keyword annotation of every dataset is very important for building a convenient and useful database of scientific datasets.

In the folksonomy approach to keyword annotation, which is adopted in many social networking service (SNS) sys-

This work was partially supported by the Data Integration and Analysis System (DIAS) of the Ministry of Education, Culture, Sports, Science and Technology, Japan (MEXT).

✉ Toshiyuki Shimizu
tshimizu@i.kyoto-u.ac.jp

Youichi Ishida
yishida@db.soc.i.kyoto-u.ac.jp

Masatoshi Yoshikawa
yoshikawa@i.kyoto-u.ac.jp

[1] Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan

tems, any user is allowed to freely annotate various data with keywords [1,11,13]. Because many general users are continuously adding keywords to a single dataset or other data object, the advantage arises that the keywords added to that data object will ultimately converge to a useful keyword set based on how the data are used. Unlike in such a folksonomy approach, in the keyword annotation of scientific data, data providers themselves add keywords to their own highly specialized data [4,9,19,22,24]. In this case, because the data provider is the only person to annotate his own data objects with descriptive keywords provided in their metadata, the utilization value of such keyword sets depends solely on the data providers. Therefore, once the metadata have been defined, their value will not be further improved by general users. Furthermore, in many cases, the keywords that can be added are restricted through the use of a specific controlled vocabulary for the relevant domain. This restriction effectively eliminates noise in data retrieval that may be caused by differences in word form and orthographic variations.

We focus on the latter case, in which data providers annotate their own scientific data, and consider methods for recommending suitable keywords from a controlled vocabulary for the annotation of scientific data. The selection of suitable keywords from a controlled vocabulary requires extensive knowledge of both the research domain and the controlled vocabulary, which typically includes thousands of keywords. Therefore, even an expert data provider will experience difficulty in selecting suitable keywords from the provided vocabulary. Controlled vocabularies exist in various research domains; examples for earth science, agriculture and biology, and the life sciences include GCMD Science Keywords [7], the Centre for Agricultural Bioscience (CAB) Thesaurus,[1] and Medical Subject Headings (MeSH),[2] respectively. Because it is hard to understand the whole context of a controlled vocabulary, keyword annotation is regarded as a very time-consuming task for data providers.

We investigated metadata for earth science data that are managed by the metadata portal called the Global Change Master Directory (GCMD)[3] and the project called the Data Integration Analysis System (DIAS)[4] [10] as examples. Both the GCMD and DIAS datasets are annotated with keywords from GCMD Science Keywords [7], and the GCMD Science Keywords vocabulary includes about 3000 keywords. Figures 1 and 2 show the distributions of the number of keywords from the GCMD Science Keywords vocabulary with which each dataset is annotated. An investigation of the metadata in the GCMD and DIAS databases revealed many poorly annotated datasets. In the GCMD metadata portal,[5] the total number of datasets is 32,731, and approximately one-fourth of these datasets have fewer than 5 keywords. In DIAS, among 437 datasets[6] annotated in English, 220 are annotated with no keywords from the GCMD Science Keywords vocabulary, and the average number of GCMD Science Keywords with which each dataset is annotated is only approximately 3. Therefore, there is a need to increase the number of keywords with which each dataset is annotated by recommending keywords for each dataset. Note that the focus of this paper is keyword recommendation methods, and we are assuming the framework that the final selection of right keywords from recommended keywords is done by data providers.

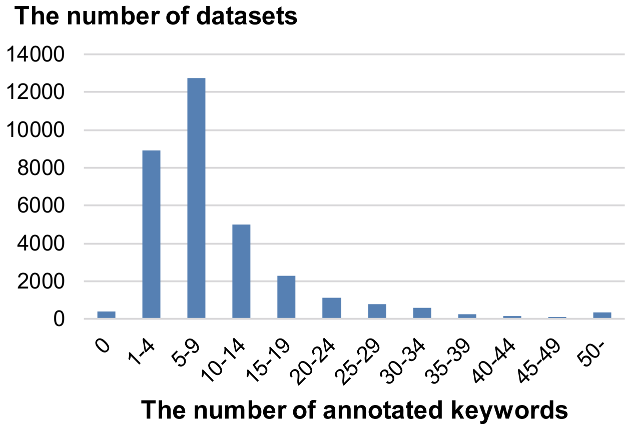


Fig. 1 Distribution of the number of annotated keywords per dataset in the GCMD

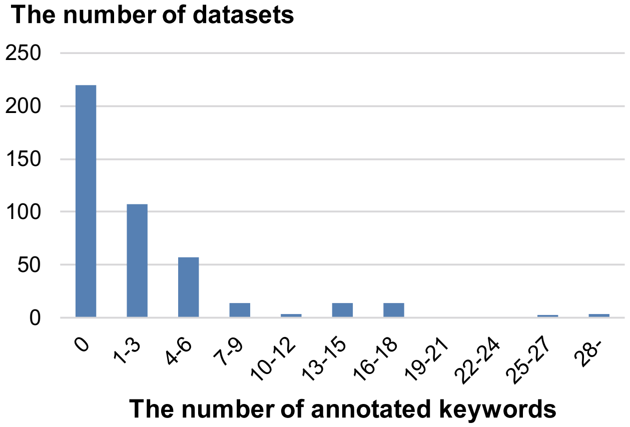


Fig. 2 Distribution of the number of annotated keywords per dataset in DIAS

[1] https://www.cabi.org/cabthesaurus/.

[2] https://www.nlm.nih.gov/mesh/.

[3] https://gcmd.nasa.gov/.

[4] https://www.diasjp.net/.

[5] As of August 2019.

[6] As of August 2019.

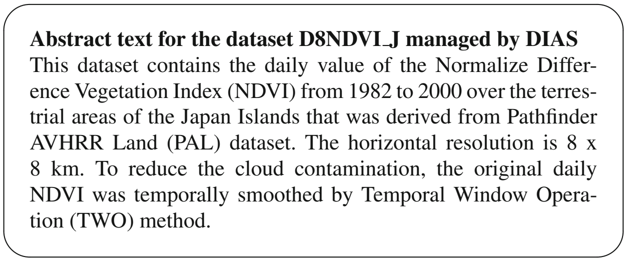
1.2 Outline of our contributions

In this paper, we discuss methods for recommending keywords when data providers create their metadata. We consider the *indirect method*, which recommends keywords based on similar existing metadata, and the *direct method* which recommends keywords based on the keyword definitions. This study makes the following three contributions:

1. We consider the importance of metadata quality in recommending suitable keywords from a controlled vocabulary for annotating scientific data, and we compare the *indirect method* of keyword recommendation with the *direct method*.
2. We propose evaluation metrics that consider a controlled vocabulary with a hierarchical structure. By applying our evaluation metrics, we can evaluate the extent to which the cost of keyword annotation is reduced through keyword recommendation.
3. We present experiments conducted on real datasets managed by the GCMD and verify the effectiveness of the *indirect* and *direct methods*.

Below, we describe each of the contributions listed above in greater detail.

Considering metadata quality in keyword recommendation As the first contribution of this paper, we discuss keyword recommendation methods from the viewpoint of metadata quality. The text information that is provided in metadata is often used for keyword recommendation. In general, this text information typically includes an abstract text, which is a free-text description of the data. For example, in the case of an earth science dataset, information regarding the items observed, the observation methods, the usage of the data and so on is provided in the abstract text, and this information is considered to be useful for recommending suitable keywords. Figure 3 shows an example of the abstract text for a dataset managed by DIAS. In this paper, we assume the abstract text in the metadata of a target dataset for which we wish to recommend keywords is available, and we utilize it for the recommendation. Previous studies on keyword recommendation [13,24] have typically proposed methods for recommending keywords to be added to a target dataset based on those associated with similar existing data by calculating the similarity between the text information in the metadata of the target and in similar existing metadata. We call this method the *indirect method*. Figure 4 illustrates the underlying concept of the indirect method. However, because this method relies on existing metadata, the effectiveness of the indirect method depends on the quality of the existing metadata set, such as the number of existing datasets with available metadata, the keywords with which the datasets are annotated, and the words contained in the abstract texts.



Abstract text for the dataset D8NDVI_J managed by DIAS
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.

Fig. 3 An example of an abstract text

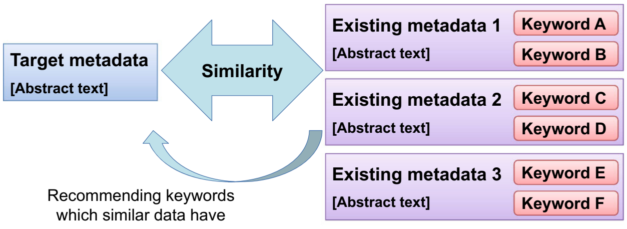


Fig. 4 Indirect method

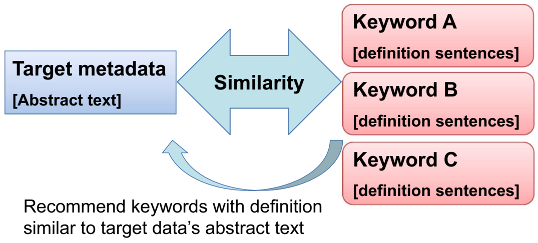
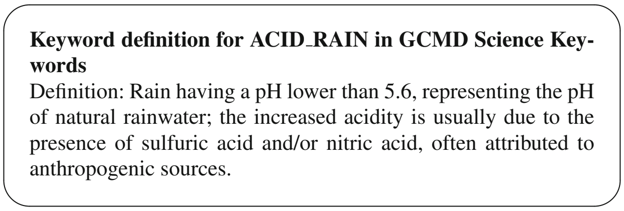


Fig. 5 Direct method

For example, if the datasets that are similar to the target dataset are poorly annotated with keywords, then the indirect method cannot provide useful recommendations even if similar datasets are available in a given portal. In fact, even beyond earth science metadata portals such as DIAS and the GCMD, it has been reported that the quality of the available metadata set is also a pressing problem in Europeana [5], which is the largest-scale digital library portal in Europe. The scope and the quantification of metadata quality as addressed in this paper will be discussed in detail later. The examples given above show that many cases exist in which the existing metadata quality is insufficient.

In this paper, we also consider a keyword recommendation method that does not depend on the quality of the existing metadata set, which we call the *direct method*. In the direct method, instead of existing metadata information, the definition sentences provided for each keyword are used in combination with the abstract text provided for the target data. Figure 5 illustrates the underlying concept of the direct method. In most cases, each keyword in a controlled vocabulary is associated with a keyword definition that explains its meaning; this is also the case for GCMD Science Keywords. Figure 6 shows an example of the definition of a



Keyword definition for 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. 6 An example of a keyword definition

keyword in GCMD Science Keywords. If a data provider provides an abstract text that includes sufficient information to describe his dataset, which is a target dataset of keyword recommendation, then the direct method can be applied to recommend suitable keywords for the dataset even if the quality of the existing metadata is insufficient. Therefore, if each data provider improves the quality of the abstract text associated with his data, then the direct method can be used to improve the quality of an entire metadata portal by increasing the number of annotated keywords. By contrast, if the existing metadata quality is insufficient, the indirect method will be unable to recommend suitable keywords for a dataset even if the data provider improves the quality of the associated abstract text. Therefore, the indirect method is less likely to increase the number of annotated keywords or to improve the quality of an entire portal. Note that if the existing metadata quality is sufficient, then the indirect method can be effectively applied for keyword recommendation. Hence, we can say the direct method, which is independent of the metadata quality, is preferable while the existing metadata quality is insufficient, and it can contribute to creating an environment in which the indirect method can function effectively.

Evaluation metrics considering a hierarchical vocabulary structure As the second contribution of this paper, we propose evaluation metrics based on a hierarchical vocabulary structure, which is a distinctive feature of most controlled vocabularies. The purpose of this study is to reduce the cost incurred by data providers for keyword annotation. Many controlled vocabularies, including GCMD Science Keywords, are hierarchically structured. We consider that the cost of keyword annotation differs depending on the position of the keyword in the controlled vocabulary that is being used. For example, because keywords in upper layers, which represent category names, such as "OCEANS," tend to be easier to select, the cost of selecting such keywords is smaller. By contrast, because keywords in lower layers, such as "SEA SURFACE TEMPERATURE," tend to be more difficult to select because they are buried under many higher keywords, the cost of their selection is larger. Considering that the value of recommending a keyword in a lower layer should not be considered equal to that of recommending a keyword in a higher layer, we propose evaluation metrics that consider a hierarchical vocabulary structure, with greater emphasis placed on keywords that are more difficult for data providers to select. The difficulty of selecting a keyword is considered to depend on either the hierarchical depth to which the keyword belongs or the number of descendants and siblings of the keyword in the controlled vocabulary. Moreover, when a method mistakenly recommends an incorrect keyword that is not desired by the data provider, we apply a penalty for that mistake.

Experiments on real datasets As the third contribution of this paper, we present experiments conducted on real datasets. We assume that the indirect method, which relies on an existing metadata set, cannot provide useful recommendations when the quality of the existing metadata is insufficient. In this paper, to verify this assumption, we consider several existing metadata sets that differ in their degrees of metadata quality. We consider several quantitative measures related to the quality of the existing metadata sets. These measures include the average number of words in each abstract text, the average number of annotated keywords per dataset and the number of existing datasets. Moreover, using different sets of existing metadata generated by varying these quantitative measures, we compare the direct and indirect methods of keyword recommendation to verify their effectiveness.

1.3 Application to earth science data

In this paper, we consider the field of earth science as one possible target research domain. Earth science data include satellite observation data, vegetation index data, data on earthquakes and so on. In earth science, the desire is to integrate these diverse data and apply them in heterogeneous domains. Therefore, to promote the use of such data, we need to support their classification and the ability to extract relevant data through keyword annotation.

Earth science metadata portal By virtue of recent advances in observation technologies and progress in information technologies, the total amounts of earth science data that are available in various fields, such as atmospheric studies and oceanic studies, have increased at an explosive pace. In addition, each of these fields is developing in the direction of increasing subdivisions or greater specialization. Therefore, because it is difficult to share the information from each field and to utilize data between fields in an integrated manner, metadata portals must be properly managed to allow the associated metadata to be effectively collected, handled and searched. For instance, the GCMD collects various kinds of earth science data and manages controlled vocabularies related to project names and platform names in addition to GCMD Science Keywords. In addition, the GCMD portal provides a search function for searching and classifying various metadata using those keywords. Moreover, in Japan, DIAS provides various search functions for earth science data based on keywords contained in controlled vocabularies and the spatiotemporal information of the data of interest.

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

Fig. 7 Keywords added to the Aqua AMSR-E dataset managed by DIAS

The aim is to build a database that promotes the application of earth science data in heterogeneous domains and the integrated use of diverse data collected in multiple fields from different places and at different times. For example, by applying precipitation data from the meteorological domain in the geological domain, it is possible to use those data for disaster prevention. Moreover, considering that global warming is likely to be a factor in fostering viral infections, global warming data can be applied in the field of medical science. In addition, a simulator has been developed that enables the prediction of the cultivation prospects for individual species of rice in a given area by integrating data on rice growth with meteorological observation data. To promote the cross-domain application and integrated use of data, each dataset should be annotated with sufficient keywords. Therefore, it is necessary to recommend sufficient keywords to allow users to easily classify various kinds of data and identify data of high relevance for a given purpose.

Keyword annotation for earth science data Metadata keywords for earth science data are added to a dataset by selecting keywords that are relevant to that dataset from a controlled vocabulary. For example, using the GCMD Science Keywords vocabulary, a dataset containing rainfall observations can be annotated with the keyword “PRECIPITATION AMOUNT.” In addition, as mentioned in Sect. 1.2, each keyword is managed hierarchically in GCMD Science Keywords. For example, consider the top-layer keyword “OCEANS”; below this keyword, “OCEAN TEMPERATURE” is one of the keywords in the second layer, and “SEA SURFACE TEMPERATURE” is a keyword in the next deeper layer. In this paper, we represent these keywords in their hierarchical structure as follows: “OCEANS > OCEAN TEMPERATURE > SEA SURFACE TEMPERATURE.” We provide examples of the hierarchical structures of several GCMD Science Keywords in Fig. 7.[7]

1.4 Outline of the paper

The remainder of this paper is organized as follows. Section 2 introduces previous studies on keyword recommendation for web pages, earth science data and academic research papers. This section also introduces research on metadata quality. Section 3 explains the two methods of keyword recommendation for scientific datasets. One is the *indirect method,* which relies on an existing metadata set, and the other is the *direct method,* which relies on the metadata of the target itself and the keyword definitions. In addition, we consider the role of metadata quality in relation to these two methods. Section 4 proposes evaluation metrics that consider a controlled vocabulary with a hierarchical structure. In Sect. 5, we present experiments conducted on real datasets managed by the GCMD and compare the results obtained using the direct and indirect methods by applying our evaluation metrics. Section 6 concludes the paper and discusses possibilities for future work.

2 Related works

In recent years, keyword recommendation based on the folksonomy approach has attracted attention from researchers [1,11,13]. However, most of these studies have focused on personalized keyword recommendations based on a user’s history. In such works, it is common for users themselves to annotate multiple data using arbitrary keywords rather than a controlled vocabulary. By contrast, our focus is the cases that only the data providers annotate their own data with keywords from a controlled vocabulary. In this case, because sufficient information on the histories of the data providers is typically unavailable, content-based methods are considered to be more useful. This section presents several related works that have proposed content-based methods of keyword recommendation. In addition, we also introduce studies on metadata quality.

2.1 Social tagging

Several studies have addressed the use of content-based methods to support social tagging [1,11,13]. For social bookmarking services such as Delicious, Lu et al. proposed a method of recommending suitable keywords for a web page lacking tag information [13]. In their method, an assignment probability is calculated for each potential tag for a web page based on the similarities between web pages and how often that tag appears in sets of tags added to similar web pages. However, this approach presupposes that multiple users will annotate a particular web page with the same tags as other users will. Hence, this method cannot be applied to the cases that only the data providers add keywords to their own data. The cited authors also calculated the trustworthiness of a web page based on the total number of tags added to that web page. However, in research domains, the numbers of keywords added to data have little relation to the reliability of those data. Belem et al. proposed a formula for calculating the relevance

[7] AMSR-E is a microwave sensor that estimates various physical quantities related to water.

of each tag for a resource using learning-to-rank technologies, combining various indicators such as tag co-occurrence, descriptive power and term predictability [1]. However, a controlled vocabulary is not used in this approach; instead, recommended keywords are extracted from among all terms present in a document. Krestel et al. suggested tags for a web page by applying latent Dirichlet allocation (LDA) [2] to a set of preassigned tags [11]. This method assumes that less frequent or unique tags that have been added by only a few users are inappropriate. However, in the case of scientific data, such as earth science data, because controlled vocabularies are used, important keywords are often assigned that appear with low frequencies. Therefore, if LDA were to be used in the earth science domain, appropriate but less frequently used keywords would likely not be recommended.

2.2 Keyword recommendation for earth science data

In the context of keyword recommendation for earth science data, Tuarob et al. [24] proposed a method of recommending tags drawn from a controlled vocabulary for data lacking tag information. In this method, the feature vector for each dataset is created from the text information available in the metadata, and tags are recommended to be added to similar datasets by calculating the similarities between feature vectors. Each document is represented by either a term frequency-inverse document frequency (TF-IDF) vector [18] or an LDA probability distribution. However, when the quality of the existing metadata set in a metadata portal is insufficient, this method can be expected to be ineffective. Therefore, we also consider the direct method, which does not depend on an existing metadata set and can be applied to a new controlled vocabulary that has not seen extensive use. Shimizu et al. proposed a method of recommending keywords that represent different earth science categories using Labeled LDA [16,22]. They defined 14 keywords as labels for learning the correspondence between the abstract text associated with a dataset and the keywords added to that dataset. Then, they recommended suitable keywords by applying the learned results to a target dataset. As demonstrated in that study, when the number of labels is small, labeled LDA is useful for recommendation. However, it is very difficult to prepare sufficient training data to define thousands of keywords as labels.

2.3 Annotation for academic research papers

This section introduces several works that have addressed methods of supporting annotation for academic research papers. Chernyak proposed a method for recommending topics based on a controlled vocabulary called the ACM Computing Classification System[8] [4]. Using techniques such as TF-IDF, BM25 and annotated suffix trees, this method calculates the similarities between the topics and an abstract text of a paper of interest. Santos and Rodrigues addressed the problem of multi-label classification for research papers using machine learning techniques such as support vector machines (SVMs), the k-nearest neighbor (k-NN) approach and naïve Bayes classification [19]. Certainly, these methods are also applicable to our topic of study. Because they utilize other existing research papers available in a portal, these methods depend on the quality of the existing metadata. Notably, although studies of annotation methods for research papers are different from those for sets of observational data, such as earth science data, in that annotation methods for research papers can utilize reference information, they have much in common with our study in that both types of approaches require a controlled vocabulary and the keywords are added by a specific person, such as the author, in both cases. Our discussion can thus be considered applicable to the annotation of research papers.

2.4 Dataset profiling

Dataset profiles consist of topics for a dataset and their relevance. The researches on dataset profiling proposed approaches based on graphical models [6], machine learning techniques [21] and calculation of cosine similarity [17]. Fetahu et al. [6] proposed an approach for creating dataset profiles through the combination of the topic extraction from reference datasets and the ranking based on graphical models. Schaible et al. addressed the problem of reusing appropriate vocabulary terms to represent linked data using a machine learning technique and a data mining approach. Those methods are based on how other data providers on the LOD cloud have used RDF classes and properties [20]. Ramnandan et al. proposed TF-IDF-based cosine similarity approach. They ranked semantic labels in decreasing order of the cosine of the angle between a target document vector and the other existing document vectors [17]. The target of these studies is general datasets such as the datasets from the LOD cloud, while the target of our study is scientific datasets and keywords from a controlled vocabulary with a hierarchical structure, which are associated with definition sentences. These studies can be considered as indirect methods and thus depend on metadata quality.

2.5 Studies on metadata quality

In this section, we introduce several studies on metadata quality. To evaluate metadata quality, Bruce and Hillmann

[8] https://www.acm.org/about-acm/class.

introduced the seven metrics of "completeness," "accuracy," "conformance to expectations," "logical consistency and coherence," "accessibility," "timeliness" and "provenance" [3]. These metrics indicate whether the metadata include sufficient information to understand the data contents, how accurate the information provided by the metadata is, whether the metadata satisfy user needs, how closely the metadata follow a standard definition, how accessible the data are, whether there are differences between the real data and the metadata because of the passage of time, and how trustworthy the manner in which the metadata were processed is, respectively. Ochoa and Duval [14] proposed formulas for quantifying these seven quality metrics for metadata in a digital library. For example, these authors quantified "conformance to expectations" by calculating the sum of the TF-IDF score of each word in the abstract text or the entropy of each keyword. However, there is no general consensus regarding how metadata quality should be quantified. Tani et al. [23] published a survey paper on the issue of metadata quality. This work summarizes various frameworks for evaluating metadata quality, including that of Bruce and Hillmann [3]. However, the authors state that there is no consensus with respect to the question "What is metadata quality?". Because metadata quality refers to the complex concept of "fitness for use," its definition changes depending on the real usage of the data in question. Moreover, these authors explain that it is difficult to define metadata quality in a way that is applicable in all contexts because the understanding and evaluation of metadata are different for each community unit. In this paper, we focus on the number of existing datasets with available metadata, the number of words in the abstract texts and the number of annotated keywords per dataset, which we regard as important aspects of the concept of metadata quality, and we define these quality measures as factors that can influence the performance of methods of keyword recommendation.

3 Metadata quality and recommendation methods

We consider that the effectiveness of the indirect method of keyword recommendation is likely to depend on the quality of the existing metadata set that is being used for reference, whereas the direct method, which uses keyword definitions, is independent of the quality of the existing metadata. First, in Sect. 3.1, we discuss the scope of metadata quality as considered in this paper, and we consider the quantification of the quality of an existing metadata set. Section 3.2 introduces the indirect method, and Sect. 3.3 explains the direct method in detail. In this paper, GCMD Science Keywords is used as the controlled vocabulary that provides the keywords, and our keyword recommendations include the entire keyword path, such as "ATMOSPHERE > ATMOSPHERIC RADIATION > HEAT FLUX." In other words, if the relevance of "HEAT FLUX" to the metadata for a target dataset is judged to be high, then we recommend "HEAT FLUX" and provide all keywords above it in its path.

3.1 Quantitative measures of metadata quality

We consider that the quality of metadata reflects the degree to which the metadata include sufficient information to characterize the associated data. With regard to an abstract text that is included in the metadata for a dataset, the quality of the abstract text is considered to represent the degree to which it describes the information required to understand the contents of the dataset, such as the process by which the data were obtained and their utility value. In addition, with regard to annotated keywords, the quality of a set of keywords is considered to represent the degree of comprehensiveness with which the keywords express various aspects of the dataset and how useful each keyword is in finding the data.

To verify whether the performance of the indirect method depends on the quality of the available metadata, we consider several existing metadata sets that differ in quality and observe the changes in the values of various evaluation metrics when the indirect method is applied to these different existing metadata sets. To perform such a verification study, it is necessary to establish a standard on which we can judge whether the existing metadata quality is sufficient. However, as mentioned in Sect. 2.5, because the concept of metadata quality encompasses several abstract aspects that cannot be quantified, it is very difficult to calculate some measure of the quality in which all of these aspects are considered. Therefore, we approximate the metadata quality in terms of quantitative measures such as the number of words in the abstract text for each dataset and the number of annotated keywords per dataset, and we then consider several existing metadata sets created to exhibit different levels of quality based on these measures.

In addition, the number of existing datasets for which associated metadata are available can itself be regarded as a measure of the metadata quality. When only a small amount of existing metadata is available in a portal, a particular dataset that suits the specific needs of any given user is less likely to be found in that portal. In this case, the quality of the existing metadata set is considered to be insufficient. Thus, in this paper, we identify the following three measures, Q_t, Q_k and Q_n, as factors related to the quality of an existing metadata set.

1. Q_t: The average number of words in the abstract text for each dataset.
2. Q_k: The average number of keywords with which each dataset is annotated.

3. Q_n: The number of existing datasets with associated metadata available in a given portal.

This study addresses the case in which, based on the three measures listed above, the quality of the existing metadata set is insufficient; we assume that in such a case, the indirect method of keyword recommendation is ineffective.

3.2 Indirect method

In this paper, as the indirect method of keyword recommendation based on an existing metadata set, we adopt Tuarob et al.'s method, which was introduced in Sect. 2.2 [24]. This method is chosen because it was developed for keyword recommendation for earth science data based on textual metadata information. In this method, using the abstract texts contained in the metadata, feature vectors are created for the metadata of a target dataset and for the metadata of existing datasets in a given portal, and these feature vectors are used to calculate the similarity between the target metadata and the existing metadata for each dataset. Then, keywords are recommended from among those used to annotate the similar existing datasets. This process is described in detail below.

$$P(k \mid q, K, D, M) = \frac{TagScore_M(k, q, D)}{\sum_{k \in K} TagScore_M(k, q, D)} \tag{1}$$

where $k (\in K)$ represents a keyword, K is a controlled vocabulary, q is the metadata for a target dataset, D represents the set of existing metadata in a portal, and M represents the method used for feature vector creation. Given K, D and M, $P(k \mid q, K, D, M)$ represents the probability that q should be annotated with the keyword k. $TagScore_M(k, q, D)$, as calculated using Eq. (2), measures the relevance between the metadata q of the target dataset and the keyword $k (\in K)$.

$$TagScore_M(k, q, D) = \sum_{d \in D} DocSim_M(q, d, D) \cdot isTag(k, d) \tag{2}$$

In Eq. (2), $DocSim_M(q, d, D)$ calculates the cosine similarity between the metadata $d \in D$ of an existing dataset and the metadata q of the target dataset. $isTag(k, d)$ is a binary function that returns a value of 1 if the keyword k is present in the existing metadata $d \in D$ and a value of 0 otherwise. Thus, $TagScore_M(k, q, D)$ represents the sum of the similarities of the target dataset with all existing datasets that are annotated with a set of keywords that includes k. $P(k \mid q, K, D, M)$ represents the normalization of $TagScore_M(k, q, D)$ such that the output will lie on the interval [0, 1], and keywords are recommended in descending order of $P(k \mid q, K, D, M)$.

In Sect. 2.2, we noted two methods that can be used to create the feature vectors, namely the calculation of TF-IDF vectors and the determination of the topic distribution for the metadata of each dataset. Tuarob et al. reported that the latter method yields more suitable keyword recommendations than the former. Therefore, in the indirect method considered here, we adopt the latter approach, in which the topic distribution of the metadata of each dataset is used as the feature vector for that dataset. In the method of Tuarob et al., the metadata for each dataset are considered to consist of a mixture of topics, and LDA is used to calculate the metadata topic distribution for each dataset. LDA is a technique that is widely used for topic modeling and assumes that a data provider has a particular set of topics in mind when annotating his data, leading the data provider to define the metadata for a dataset by selecting words related to those topics. LDA is used to infer the topic distributions intended by the data provider by analyzing the text information present in a metadata set. In this paper, for consistency with the experimental conditions considered by Tuarob et al., we use the LDA algorithm with 300 topics and 1000 iterations. In addition, when using LDA, we must determine the parameters α and β that describe the per-document topic distributions and the per-topic word distributions, respectively. No description of the detailed values is provided in Tuarob et al.'s paper, but $(\alpha, \beta) = (50/L, 0.1)$ is often adopted, where L is the number of labels [8]. Therefore, we adopt $(\alpha, \beta) = (50/300 \simeq 0.167, 0.1)$. We use the Stanford Topic Modeling Toolbox[9] to apply LDA.

$$TopicDistribution(q, T) = [t_1, t_2, \ldots, t_{300}] \tag{3}$$

T represents the set of topics; thus, $|T| = 300$. $t_j \in T$ is the probability of assigning topic j to the metadata q of the target dataset. In this way, the indirect method treats the topic distribution of the metadata associated with each dataset as the feature vector for that dataset and recommends keywords by calculating the similarity between the target dataset and each existing dataset as indicated by their associated metadata.

3.3 Direct method

In this paper, we use the abstract texts contained in the metadata associated with datasets of interest as a basis for the direct method of keyword recommendation. By viewing the abstract text, a user can gain a rough understanding of the content of a dataset. Initially, using a simple string matching method, we extracted keywords from the abstract text associated with a dataset in DIAS by matching the words present in the text with each keyword in the GCMD Science Keywords vocabulary. Note that we preprocessed the abstract texts and the keywords by removing stopwords and stemming each

[9] https://nlp.stanford.edu/software/tmt/tmt-0.3/.

word. However, although the average number of annotated keywords per dataset in the GCMD is approximately 10, the string matching method was only able to recommend an average of approximately 2.7 keywords per dataset. Therefore, we decided to utilize the implicit information provided by the keyword definitions in addition to the explicit information of the keyword names. The GCMD Science Keywords vocabulary includes a definition for each keyword. We calculate the similarities between the abstract text provided in the metadata of a target dataset and the definition of each keyword, and we then recommend keywords with a high degree of similarity.

To calculate the similarities, we create feature vectors for the metadata of the target dataset and each keyword definition. Note that we preprocess the abstract texts and the keyword definitions by removing stopwords and stemming each word. We represent the abstract text associated with a target dataset by A_i, and we represent the set of words contained in A_i by $T(A_i)$. In addition, we represent the definition of a keyword by D_j and represent the set of words contained in D_j by $T(D_j)$. Then, we represent a feature vector of A_i by $\mathbf{DA}(A_i)$, and a feature vector of D_j by $\mathbf{KD}(D_j)$. Each element of $\mathbf{DA}(A_i)$ and $\mathbf{KD}(D_j)$ is the weight for the word $t_l \in T(A_i) \cup T(D_j)$, and thus, the sizes of both $\mathbf{DA}(A_i)$ and $\mathbf{KD}(D_j)$ are $|T(A_i) \cup T(D_j)|$. For $\mathbf{DA}(A_i)$, the weight of t_l represented by $DA(t_l)$ is whether t_l appears in A_i or not.

$$\mathbf{DA}(A_i) = [DA(t_1), DA(t_2), \ldots, DA(t_l), \ldots, DA(t_n)] \tag{4}$$

$$DA(t_l) = \begin{cases} 1 & (t_l \in T(A_i)) \\ 0 & (\text{otherwise}) \end{cases} \tag{5}$$

$$t_l \in T(A_i) \cup T(D_j) \quad (1 \leq l \leq |T(A_i) \cup T(D_j)|{=}n) \tag{6}$$

For $\mathbf{KD}(D_j)$, the weight of t_l represented by $KD(t_l)$ is the TF-IDF value of t_l. For the term frequency (TF), we use the Length Regularized TF (LRTF) introduced in [15]. These quantities are described in detail below.

$$\mathbf{KD}(D_j) = [KD(t_1), KD(t_2), \ldots, KD(t_l), \ldots, KD(t_n)] \tag{7}$$

$$KD(t_l) = \begin{cases} \mathrm{LRTF}(t_l, D_j) \cdot \mathrm{IDF}(t_l, C) & (t_l \in T(D_j)) \\ 0 & (\text{otherwise}) \end{cases} \tag{8}$$

$$\mathrm{LRTF}(t_l, D_j) = \mathrm{TF}(t_l, D_j) \cdot \log_2\left(1 + \frac{ADL(C)}{\mathrm{len}(D_j)}\right) \tag{9}$$

$$\mathrm{IDF}(t_l, C) = \log_2\left(\frac{|C|}{DF(t_l, C)}\right) + 1 \tag{10}$$

Let C be the set of keyword definitions, where $D_j \in C$, and let $|C|$ be the number of keywords. In addition, $\mathrm{len}(D_j)$ is the length of keyword definition D_j, $ADL(C)$ is the average of all $\mathrm{len}(D_j)$, and $\mathrm{TF}(t_l, D_j)$ is the appearance frequency of word t_l in D_j. The LRTF is calculated using a formula that normalizes the TF value by considering the ratio of $\mathrm{len}(D_j)$ to $ADL(C)$. We consider the LRTF to be appropriate in this situation, in which an abstract text is treated as the object of a query, because it is stated in [15] that the LRTF is useful for long queries consisting of more than 5 words. The inverse document frequency (IDF) is calculated using the standard formula, in which $|C|$ is divided by $DF(t_l, C)$. In this paper, we utilize the cosine similarity, which is generally used when comparing documents in vector-space models. The cosine similarity takes values in the interval [0, 1], and it represents the angle formed by two vectors: a value close to 1 indicates that the vectors are similar to each other, whereas a value close to 0 indicates that they are independent of each other. We recommend keywords in descending order of cosine similarity. The calculation is presented in detail below.

$$\mathit{CosineSim}(\mathbf{DA}(A_i), \mathbf{KD}(D_j)) = \frac{\mathbf{DA}(A_i) \cdot \mathbf{KD}(D_j)}{|\mathbf{DA}(A_i)| \cdot |\mathbf{KD}(D_j)|} \tag{11}$$

4 Evaluation metrics considering a hierarchical vocabulary structure

This section describes our proposed evaluation metrics that consider a hierarchical vocabulary structure, which is a distinct characteristic of most controlled vocabularies. The purpose of our study is to reduce the cost incurred by data providers for keyword annotation. We consider that for each keyword, this cost depends on the position of that keyword in the controlled vocabulary. Recommendations of keywords in lower layers, which are more difficult to select, lead to greater reductions in cost than recommendations of keywords in upper layers, which tend to represent category names, such as “OCEANS” or “ATMOSPHERE” in the case of GCMD Science Keywords. Thus, considering a hierarchical vocabulary structure, we propose evaluation metrics that place greater emphasis on keywords that are more difficult for data providers to select. First, Sect. 4.1 explains the need for our evaluation metrics by comparing our metrics with precision and recall, which are generally used to evaluate recommendations. Next, as an example of a recommendation evaluation metric that considers a hierarchical structure of the items to be recommended, Sect. 4.2 introduces the normalized eXtended Cumulated Gain (*nxCG*) [12], which is an evaluation metric for the retrieval of XML documents. Finally, Sect. 4.3 explains our proposed evaluation metrics for a controlled vocabulary such as GCMD Science Keywords. As mentioned in Sect. 3, we note that each recommended keyword is given as its entire path, such as “ATMOSPHERE > ATMOSPHERIC RADIATION > HEAT FLUX.”

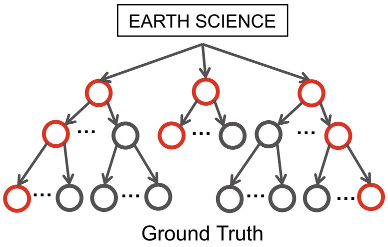


Fig. 8 An example of a ground truth

4.1 The need for our metrics

In this section, we explain the need for our evaluation metrics by considering an example in which we use GCMD Science Keywords as the controlled vocabulary. As mentioned above, the keywords in a controlled vocabulary are often managed hierarchically. Figures 8 and 9 show the hierarchical structure of a controlled vocabulary, in which each node represents one keyword. In Fig. 8, we represent the correct keywords in the controlled vocabulary for a certain dataset as red bordered nodes. In Fig. 9, we present two patterns of recommended results for this dataset, in which each recommended keyword is indicated by a yellow node. In Result (1), three correct keywords are recommended, all in the top layer, whereas Result (2) also contains a total of three correct keywords, but they include keywords in the lower layers. In terms of precision or recall, both of these results earn the same evaluation score of 50%. However, we cannot consider both sets of recommendations to be of equivalent value. Although Result (1) provides correct keywords, which represent the earth science categories “ATMOSPHERE,” “BIOSPHERE” and “AGRICULTURE,” these keywords are easy for a data provider to select and also have many descendants. Therefore, the recommendation provided by Result (1) cannot significantly reduce the cost of keyword annotation. By contrast, Result (2) recommends correct keywords in lower layers, such as “HEAT FLUX.” The selection of such keywords requires extensive knowledge of both the domain and the controlled vocabulary. Therefore, because it recommends keywords that are more difficult to select, Result (2) is considered to reduce the cost incurred by the data provider to a greater extent. Considering that the cost differs depending on the positions of the keywords in the controlled vocabulary, these two sets of recommendations should not be regarded as equivalently valuable, even though their evaluation scores in terms of commonly used metrics such as precision and recall are identical. To differentiate between such recommendations, we must use evaluation metrics that consider a hierarchical vocabulary structure and place greater emphasis on keywords that are more difficult to select.

4.2 Normalized eXtended Cumulated Gain

As an example of a recommendation evaluation metric that considers a hierarchical structure of the items to be recommended, we can consider *nxCG* [12]. *nxCG* (Normalized eXtended Cumulated Gain) is an evaluation metric for the retrieval of XML documents. In this context, the items recommended in search results can include the *Section*, *Paragraph* or *Sentence* of interest or even divisions with greater granularity. The *nxCG* value represents the relative cumulated gain achieved by an actual system compared with that achieved by an ideal system. The features of *nxCG* include the ability to consider partially correct items and the overlaps between recommended items in a search result.

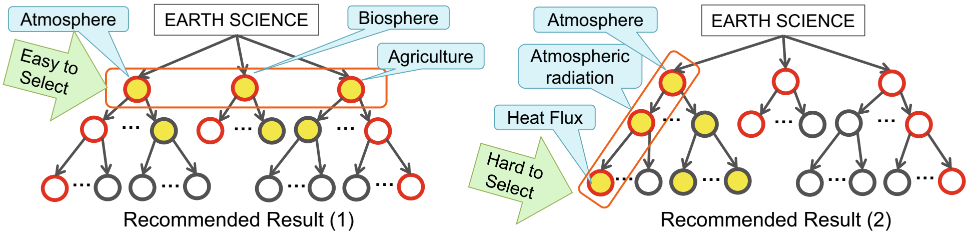
4.2.1 Partially correct items

When a system presents $Section(A)$, which contains $Sentence(B)$, which is a perfectly correct item, $Section(A)$ can be regarded as a partially correct item. Therefore, we should discount the gain achieved through the identification of $Section(A)$. Similarly, when the system presents $Paragraph(D)$, which contains $Sentence(E)$, which is a perfectly correct item, we must discount the gain achieved through the identification of $Paragraph(D)$. In this case, considering that the user will be obliged to read some part of $Section(A)$ or $Paragraph(D)$ that is irrelevant to the original query, we can regard this discounting of the gain as the generation of a penalty.

4.2.2 Overlaps between recommended items

We consider the case in which a system presents $Paragraph$ (D) as the highest ranked item and $Sentence(E)$ as the second highest ranked item. Because $Paragraph(D)$ contains $Sentence(E)$, the gain achieved by presenting $Sentence(E)$

Fig. 9 Two examples of sets of recommended results



as the second item is offset by the overlap. In other words, because there is an overlap between $Paragraph(D)$ and $Sentence(E)$, the gain achieved through the identification of $Sentence(E)$ is reduced by the gain represented by $Paragraph(D)$, which is the highest ranked item. Therefore, *nxCG* calculates the cumulated gain, reducing the total gain in the case of an overlap between recommended items.

4.2.3 How to calculate *nxCG*

Considering partially correct items and the overlaps between recommended items, *nxCG* calculates a relative value with respect to the cumulated gain that would be achieved by an ideal system. $nxCG@n$ can be expressed as follows, where n represents the number of recommended items presented in a search result.

$$nxCG@n = \frac{\sum_{j=1}^{n} xG(j)}{\sum_{j=1}^{n} xI(j)} \tag{12}$$

$xG(j)$ and $xI(j)$ represent the gains for the jth-ranked items presented by the actual system and the ideal system, respectively.

4.3 Proposed evaluation metrics

4.3.1 Differences from evaluation metrics for XML documents

The two features of *nxCG* discussed above are also applicable to a controlled vocabulary such as GCMD Science Keywords. However, the evaluation concept differs from that for XML documents in the following two respects.

The first is the hierarchical depth at which an item may be partially correct. One of the features of *nxCG* is that a penalty is generated if a recommended item includes information that is not relevant to the user's original request. However, controlled vocabularies and XML documents differ in the opportunities they provide for penalty generation. The top layer and third layer in a controlled vocabulary are analogous to the *Section* and *Sentence* levels, respectively, in XML documents. In the case of XML documents, a *Section* may be only partially correct when it is a recommended item in a search result, whereas when a *Sentence* is recommended, it is either perfectly correct or perfectly incorrect. By contrast, in the case of a controlled vocabulary, there is the possibility that a recommended keyword in the top layer may be perfectly correct or incorrect, and when a keyword in the third layer is recommended, it may be only partially correct. For example, suppose that "ATMOSPHERE > ATMOSPHERIC RADIATION > HEAT FLUX" is a correct keyword. If we were to recommend another keyword in the third layer, such as "ATMOSPHERE > ATMOSPHERIC RADIATION > EMISSIVITY," then we could judge the "ATMOSPHERE > ATMOSPHERIC RADIATION" portion of the recommended keyword to be partially correct. Thus, we would consider the mistaken recommendation of "EMISSIVITY" to generate a penalty.

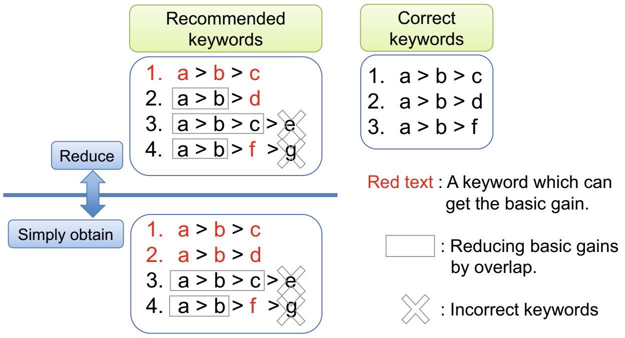


Fig. 10 Procedure for determining the gain in the case of overlap

The second difference is the way in which overlaps between recommended items are addressed. For example, suppose that "ATMOSPHERE > ATMOSPHERIC RADIATION > HEAT FLUX" is a correct keyword, the highest ranked keyword is "ATMOSPHERE > ATMOSPHERIC RADIATION > EMISSIVITY," and the second highest ranked keyword is "ATMOSPHERE > ATMOSPHERIC RADIATION > HEAT FLUX." If the obtained gain were to be reduced by the overlap in the same manner as in the case of *nxCG*, then the gain from the second recommended keyword would be reduced by the overlap of "ATMOSPHERE > ATMOSPHERIC RADIATION," which is the common portion of the first and second recommended keywords. If this approach were to be used, larger gains would tend to be achieved by recommending keywords that avoid overlaps with higher ranked keywords. This would lead to an emphasis on the recommendation of a wide variety of keywords that belong to different categories. However, from the perspective of data providers, it is not necessarily more useful to recommend a wide variety of keywords. Moreover, some data providers may find it valuable to receive correct recommendations for multiple keywords that belong to the same category. Therefore, we consider both the case in which we reduce the obtained gain by the overlaps between recommended items and the case in which we simply calculate all gains provided by correctly recommended keywords without reducing them by their overlaps. However, even in the latter case, if a keyword in a lower layer in a recommended keyword path is incorrect, then we reduce the gain by the overlap in the same manner as in the former case. The details of the gain calculation procedure are illustrated in Fig. 10.

4.3.2 Overview of the proposed evaluation metrics

Considering these two differences with respect to the evaluation of retrieved XML documents, this section presents the overview of the proposed evaluation metrics. We first explain the assignment of gains and penalties in the case of a controlled vocabulary with a hierarchical structure and then introduce our evaluation metrics considering the two approaches of handling the gains in the case of overlap.

We consider that the cost of annotation for each keyword depends on either the hierarchical depth to which that keyword belongs or how many descendants and siblings that keyword has, and we approximate the cost reduction associated with the recommendation of each keyword by a gain that is assigned to each keyword in advance. In this study, this gain is referred to as the "basic gain." As methods for calculating the basic gain for each keyword, we propose the following two approaches, which consider the cost-affecting factors mentioned above. One approach is to use the hierarchical depth, and the other is to use the numbers of descendants and siblings. In addition, we define appropriate penalties to be applied to the final gain in each of these two approaches. The overview of the procedure for calculating the final gain is presented in Fig. 11, and the details of the calculations are discussed later.

We call our evaluation metrics normalized hierarchical Cumulated Gain (*nhCG*) metrics. Based on the two methods of calculating the basic gains and the two methods of handling the gains in the case of overlap, we propose the following four evaluation metrics: $nhCG_{HDnoreduce}$, $nhCG_{HDreduce}$, $nhCG_{DSnoreduce}$ and $nhCG_{DSreduce}$. Table 1 summarizes these four metrics.

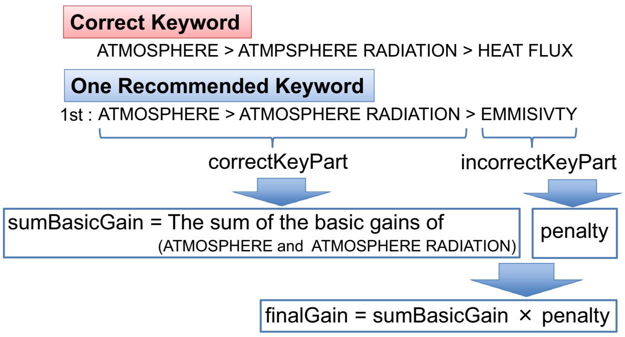

Correct Keyword
ATMOSPHERE > ATMPSPHERE RADIATION > HEAT FLUX
One Recommended Keyword
1st : ATMOSPHERE > ATMOSPHERE RADIATION > EMMISIVTY
correctKeyPart
incorrectKeyPart
sumBasicGain = The sum of the basic gains of
(ATMOSPHERE and ATMOSPHERE RADIATION)
penalty
finalGain = sumBasicGain × penalty

Fig. 11 Procedure for calculating the final gain for each recommended keyword

Table 1 The four proposed metrics

Proposed metric	Basic gain	Case of overlap
$nhCG_{HDnoreduce}@n$	Hierarchical depth	Simply obtain
$nhCG_{HDreduce}@n$	Hierarchical depth	Reduce
$nhCG_{DSnoreduce}@n$	Descendants and siblings	Simply obtain
$nhCG_{DSreduce}@n$	Descendants and siblings	Reduce

As in the case of *nxCG*, these metrics represent relative values compared with the cumulated gain achieved by an ideal system. The formula for calculating $nhCG_{HDnoreduce}$ is given in detail below. Although the methods of calculating the final gain differ among the different metrics, the calculation of the relative value compared with the ideal gain is the same for all of them. *finalGain* represents the final gain obtained for each recommended keyword.

$$nhCG_{HDnoreduce}@n = \frac{\sum_{j=1}^{n} finalGain(j)}{\sum_{j=1}^{n} xI(j)} \tag{13}$$

4.3.3 Basic gain considering hierarchical depth

In this approach, we assign a basic gain to each keyword based on its hierarchical depth. As mentioned above, we wish to assign larger basic gains to keywords in lower layers, which are more difficult to select. We consider that in general, as the hierarchical depth increases, the number of keywords belonging to that depth also increases. Therefore, we calculate the basic gain for each keyword as a value related to the number of keywords that belong to the same depth compared with the total number of keywords in the controlled vocabulary. However, if the number of keywords that belong to a deeper depth is much greater than the number that belong to a shallower depth, a direct proportionality may lead to an excessive difference between the basic gains assigned to keywords at these two depths. Therefore, we use the following nonlinear formula to calculate the basic gains:

$$basicGain_{HD}(k) = \frac{\frac{keynum(h)}{|K|} \cdot 10}{\log_2 h} \tag{14}$$

h represents the hierarchical depth to which keyword k belongs, and *keynum* is a function that returns the number of keywords that belong to depth h. In addition, $|K|$ represents the total number of keywords in the controlled vocabulary

Table 2 Basic gain considering hierarchical depth for each keyword in GCMD Science Keywords

Depth	$\#of keywords$	Basic gain
First layer	14	0.089
Second layer	121	0.429
Third layer	921	2.596
Fourth layer	652	1.474
Fifth layer	183	0.253
Sixth layer	17	0.018

K. Table 2 shows the results of applying this formula to GCMD Science Keywords used in our experiments. In the case of GCMD Science Keywords, because the numbers of keywords in the fourth and deeper layers are monotonically decreasing, relatively small scores are assigned to their basic gains. This indicates, for example, that there is little difference between the costs for selecting keywords in the fifth and sixth layers because the numbers of keywords in those layers are small.

4.3.4 Basic gain considering descendants and siblings

In this approach, we calculate the basic gain for each keyword based on the numbers of descendants and siblings it possesses. As in Sect. 4.3.3, we wish to estimate larger basic gains for keywords that are more difficult to select from the controlled vocabulary. Therefore, we assign larger basic gains to keywords with more siblings because it is more difficult to select a suitable keyword from among a larger number of keywords that belong to the same upper keyword. As shown in Fig. 12, a larger basic gain can be achieved by recommending a keyword that has 15 siblings than by recommending a keyword that has only 2 siblings.

In addition, we wish to assign larger gains to more specific keywords that allow better discrimination among datasets. Because more abstract keywords tend to be located at shallower depths, such keywords are easier to select from a controlled vocabulary. By contrast, because more specific keywords tend to be located at deeper depths, such keywords are more difficult to select. Let us consider an example based on GCMD Science Keywords. We propose that we should estimate smaller basic gains for more abstract keywords, which represent general categories of earth science, such as "ATMOSPHERE," whereas we should estimate larger basic gains for more specific keywords, which represent specific types of observations or phenomena, such as "ACID RAIN." Based on these considerations, we calculate the basic gain for each keyword by considering the number of descendants of that keyword. That is, we regard a keyword with more descendants as a more abstract keyword, which encompasses various meanings, and a keyword with fewer descendants as a more specific keyword, which embodies only a particular meaning. As illustrated in Fig. 13, a larger basic gain can be achieved by recommending a keyword with only 2 descendants than by recommending a keyword with 20 descendants.

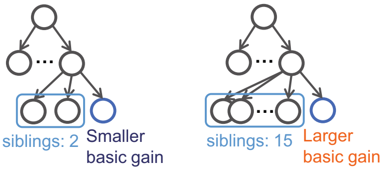


Fig. 12 Considering the number of siblings

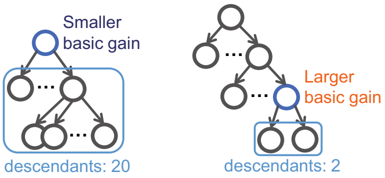


Fig. 13 Considering the number of descendants

When the numbers of descendants and siblings are considered as described above, the formula for calculating the basic gain for each keyword is as given below.

$$basicGain_{DS}(k) = \log_2 \left\{ 1 + \left(\frac{sibling(k)}{descendant(k) + 1} \right) \right\} \quad (15)$$

$sibling(k)$ is a function that returns the number of siblings of a keyword k, and $descendant(k)$ is a function that returns the number of descendants of the keyword k. The reason why we add 1 to $descendant(k)$ is to account for the case of a keyword that has no descendants. Similarly, we add 1 to the argument of the logarithm to prevent its value from being less than 1. As before, to avoid excessive differences among the basic gains for different keywords, we define the formula to vary in a nonlinear form.

4.3.5 Assigning penalties considering hierarchical depth

This section explains the method used to assign a penalty for a mistakenly recommended keyword when the gain is calculated based on hierarchical depth. As stated above, we assign a penalty to each recommended keyword path that contains an incorrect keyword. For example, suppose that "ATMOSPHERE > ATMOSPHERIC RADIATION > HEAT FLUX" is a correct keyword and that "ATMOSPHERE > ATMOSPHERIC RADIATION > EMISSIVITY" is recommended as the highest ranked keyword. We consider the mistaken recommendation of "EMISSIVITY" to generate a penalty. This penalty reduces the final gain with respect to the sum of the basic gains obtained for "ATMOSPHERE > ATMOSPHERIC RADIATION." In this case, because two of the three keywords in the recommended keyword path are correct, we consider that the final gain is two-thirds the sum of the basic gains. Algorithms 1 and 2 present pseudocodes demonstrating the procedures for calculating the final gain for a recommended keyword subject to a penalty. *recomKeyword* represents the full path of a recommended keyword, such as "ATMOSPHERE > ATMOSPHERIC RADIATION > EMISSIVITY." *correctKeyPart* is an array that stores each correct keyword in the recommended keyword path, and *sumBasicGain* is the sum of the

Algorithm 1 Final gain for a recommended keyword using $nhCG_{HDnoreduce}@n$

```
Input: $recomKeyword$, $correctKeyPart$
Output: $finalGain$ for $recomKeyword$
$sumBasicGain \leftarrow 0$
if $correctKeyPart.length \neq recomKeyword.length$ then
    for $keyword$ in $correctKeyPart$ do
        if $!overlap(keyword)$ then
            $sumBasicGain \leftarrow sumBasicGain + basicGain_{HD}(keyword)$
        end if
    end for
else
    for $keyword$ in $correctKeyPart$ do
        $sumBasicGain \leftarrow sumBasicGain + basicGain_{HD}(keyword)$
    end for
end if
$finalGain \leftarrow sumBasicGain \cdot (correctKeyPart.length/recomKeyword.length)$
```

Algorithm 2 Final gain for a recommended keyword using $nhCG_{HDreduce}@n$

```
Input: $recomKeyword$, $correctKeyPart$
Output: $finalGain$ for $recomKeyword$
$sumBasicGain \leftarrow 0$
for $keyword$ in $correctKeyPart$ do
    if $!overlap(keyword)$ then
        $sumBasicGain \leftarrow sumBasicGain + basicGain_{HD}(keyword)$
    end if
end for
$finalGain \leftarrow sumBasicGain \cdot (correctKeyPart.length/recomKeyword.length)$
```

basic gains obtained for the keywords in the recommended keyword path. *overlap*(*keyword*) is a function that identifies whether a keyword has previously appeared at a higher ranking.

4.3.6 Assigning penalties considering descendants and siblings

This section explains the method used to assign a penalty for a mistakenly recommended keyword when the gain is calculated based on the numbers of descendants and siblings. We determine the magnitude of the penalty to be assigned depending on the numbers of siblings of the incorrect keywords that are included in the path of the recommended keyword. For the example presented in Sect. 4.3.5, we assign a penalty to the sum of the basic gains obtained for "ATMOSPHERE > ATMOSPHERIC RADIATION" by considering the number of siblings of the keyword "EMISSIVITY." As shown in Fig. 14, if an incorrect keyword has dozens of siblings, then the data provider must investigate whether those dozens of siblings are correct, whereas if an incorrect keyword has only two siblings, then all the data provider must do is to investigate whether those two siblings are correct. Therefore, when an incorrect keyword has more siblings, the cost of keyword annotation is considered to be higher, and we assign a larger penalty. Algorithms 3 and 4 present pseudocodes demonstrating the procedures for calculating the final gain for a recommended keyword subject to a penalty. *incorrectKeyPart* is an array that stores the incorrect keywords in the recommended keyword path. *sumSiblingNum* is the total sum of the numbers of siblings of the incorrect keywords. When calculating the final gain, we suppress any excessive penalty effect by introducing a logarithm into the formula. Moreover, to ensure that a penalty is also assigned to a keyword with one sibling, we add 2 to *sumSiblingNum*.

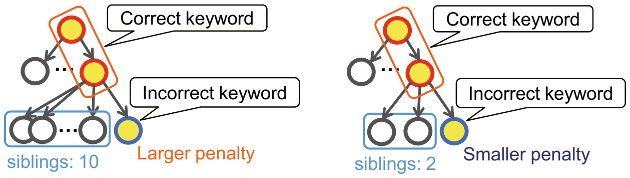


Fig. 14 Penalty assignment considering descendants and siblings

Algorithm 3 Final gain for a recommended keyword using $nhCG_{DSnoreduce}@n$

```
Input: $recomKeyword$, $correctKeyPart$, $incorrectKeyPart$
Output: $finalGain$ for $recomKeyword$
$sumBasicGain \leftarrow 0$
$sumSiblingNum \leftarrow 0$
if $correctKeyPart.length \neq recomKeyword.length$ then
    for $keyword$ in $correctKeyPart$ do
        if $!overlap(keyword)$ then
            $sumBasicGain \leftarrow sumBasicGain + basicGain_{DS}(keyword)$
        end if
    end for
    for $keyword$ in $incorrectKeyPart$ do
        $sumSiblingNum \leftarrow sumSiblingNum + sibling(keyword)$
    end for
else
    for $keyword$ in $correctKeyPart$ do
        $sumBasicGain \leftarrow sumBasicGain + basicGain_{DS}(keyword)$
    end for
end if
$finalGain \leftarrow sumBasicGain \cdot (1/\log_2(sumSiblingNum + 2))$
```

5 Experimental evaluation

In this paper, we consider the indirect method and the direct method of keyword recommendation. The direct method is independent of the quality of the existing metadata set because it exploits the information available in keyword definitions. By contrast, if the quality of the existing metadata set is insufficient, the indirect method, which relies on existing

Algorithm 4 Final gain for a recommended keyword using $nhCG_{DSreduce}@n$

```
Input: $recomKeyword$, $correctKeyPart$, $incorrectKeyPart$
Output: $finalGain$ for $recomKeyword$
$sumBasicGain \leftarrow 0$
$sumSiblingNum \leftarrow 0$
for $keyword$ in $correctKeyPart$ do
    if $!overlap(keyword)$ then
        $sumBasicGain \leftarrow sumBasicGain + basicGain_{DS}(keyword)$
    end if
end for
if $incorrectKeyPart \neq []$ then
    for $keyword$ in $incorrectKeyPart$ do
        $sumSiblingNum \leftarrow sumSiblingNum + sibling(keyword)$
    end for
end if
$finalGain \leftarrow sumBasicGain \cdot (1/\log_2(sumSiblingNum + 2))$
```

metadata, is less likely to provide useful recommendations. In this section, to verify our hypothesis, we consider several existing metadata sets that differ in quality and observe the changes in the values of the evaluation metrics that occur when the indirect method is applied to these sets of existing metadata. We use the evaluation metrics proposed in Sect. 4.3 to compare the direct and indirect methods.

In our experiments, we use real metadata on datasets managed by the GCMD. Because the metadata in the GCMD include GCMD Science Keywords, we can regard the annotated GCMD Science Keywords as the set of correct keywords for that dataset. We create the ideal result for each of our evaluation metrics by ranking these correct keywords such that the sum of the gains obtained with the considered evaluation metric is as large as possible.

5.1 Creation of existing metadata sets that differ in quality

In the case of GCMD metadata used in our experiments, the average number of words in the abstract text is approximately 112 words; there are approximately 10,000 datasets whose abstract texts contain fewer than 50 words, including 6000 datasets whose abstract texts contain fewer than 25 words. Figure 15 shows the distribution of the number of words in the abstract text in the GCMD metadata set.[10] Thus, if an abstract text contains fewer than 50 words or 25 words, or approximately one half or one quarter of the average number of words in the abstract text in the GCMD metadata set, respectively, then we consider the metadata quality for the corresponding dataset to be insufficient from the perspective of the abstract text. By contrast, if an abstract text contains more than 200 words, or approximately twice the average number of words in the abstract text in the GCMD metadata set, then we consider the corresponding metadata to include sufficient information to understand the contents of the associated dataset. We also consider the number of annotated keywords per dataset. For example, if fewer than 5 keywords are used to annotate a dataset (for which at least 10 keywords should be required), then the metadata quality for that dataset is considered to be insufficient from the keyword perspective. By contrast, considering the fact that most datasets in the GCMD are annotated with fewer than 10 keywords as we can see in Fig. 1, if a dataset is annotated with approximately 10 keywords, then those keywords can be considered to cover, to some extent, multiple aspects of the dataset.

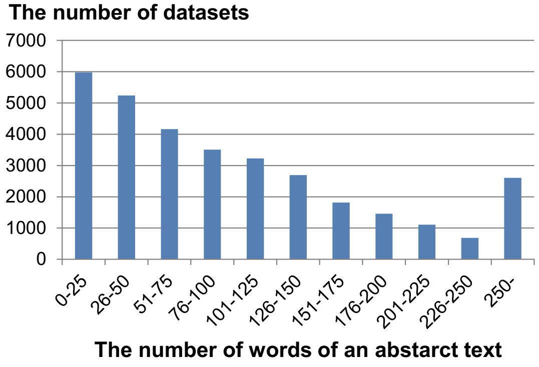


Fig. 15 Distribution of the number of words in the abstract text provided for each dataset in the GCMD

Here, we explain the process used to create existing metadata sets that differ in quality for our experimental evaluation. To consider all cases in which the quality is either sufficient or not in terms of each of the three quality measures introduced in Sect. 3.1, namely Q_t (average number of words in the abstract text), Q_k (average number of annotated keywords) and Q_n (number of existing datasets), we constructed one existing metadata set with each of the 2^3 corresponding patterns. Table 3 lists the created sets MS_1 to MS_8 and their characteristics as well as the number of datasets (metadata) assigned to each set. First, because we can observe most datasets in the GCMD are annotated with fewer than 10 keywords, we consider datasets that are annotated with more than 10 keywords to have sufficient quality in terms of keywords. Thus, to create a metadata set with an insufficient number of keywords, we randomly removed keywords from the metadata of these datasets until each had fewer than 5 keywords, which is less than half of the average number. Second, to obtain metadata sets with sufficient and insufficient quality in terms of the abstract text, we divided the datasets into those whose metadata contained abstract texts of fewer than 50 words and more than 200 words, respectively. Finally, by randomly selecting the metadata associated with 100 datasets from each metadata set created as described above, we created metadata sets with insufficient quality in terms of the

[10] We removed stopwords.

Table 3 Artificially created sets of existing metadata

Existing metadata set	Abstract text	Annotated keywords	Datasets
MS_1 (small Q_t, large Q_k, small Q_n)	$\leq$ 50 words	$\geq$ 10 keywords	100 cases
MS_2 (small Q_t, large Q_k, large Q_n)	$\leq$ 50 words	$\geq$ 10 keywords	2487 cases
MS_3 (large Q_t, large Q_k, small Q_n)	$\geq$ 200 words	$\geq$ 10 keywords	100 cases
MS_4 (large Q_t, large Q_k, large Q_n)	$\geq$ 200 words	$\geq$ 10 keywords	1981 cases
MS_5 (small Q_t, small Q_k, small Q_n)	$\leq$ 50 words	$\leq$ 5 keywords	100 cases
MS_6 (small Q_t, small Q_k, large Q_n)	$\leq$ 50 words	$\leq$ 5 keywords	2487 cases
MS_7 (large Q_t, small Q_k, small Q_n)	$\geq$ 200 words	$\leq$ 5 keywords	100 cases
MS_8 (large Q_t, small Q_k, large Q_n)	$\geq$ 200 words	$\leq$ 5 keywords	1981 cases

number of existing datasets for which metadata are available. Thus, we created 8 metadata sets to serve as references for recommending keywords for a target dataset.

5.2 Influence of metadata quality on the performance of the indirect method

In this section, we discuss the variations in the evaluation values observed among the results for the different existing metadata sets and describes the influence of each of the three considered quality measures on the performance of the indirect method.

To consider the possibility that the values of the evaluation metrics may vary depending on the metadata quality of the target dataset, we prepared five target metadata sets, each consisting of metadata for 100 target datasets and one each with abstract texts consisting of fewer than 50 words, 51–100 words, 101–150 words, 151–200 words and more than 200 words. To ensure reliability, each target dataset was annotated with more than 10 keywords, and there was no overlap between the target metadata sets and the existing metadata sets.

We regarded the keywords that have been added to each target dataset in the GCMD as the set of correct keywords for that dataset. We evaluated the top 10 keywords recommended by the indirect method and by the direct method using the proposed evaluation metrics. Figures 16, 17, 18, 19, 20, 21, 22 and 23 show the results of applying both methods to the target metadata sets. Figures 16, 18, 20 and 22 present the results obtained using MS_1 to MS_4, in which Q_k is large. Figures 17, 19, 21 and 23 present the results obtained using MS_5 to MS_8, in which Q_k is small. The horizontal axis represents the target metadata set, and the vertical axis represents the average evaluation value.

The number of words in the abstract text For each of Figures 16, 17, 18, 19, 20, 21, 22 and 23, we discuss the influence of Q_t on the performance of the indirect method. Because we should compare existing metadata sets for which the number of datasets with available metadata is comparable, we consider the changes in the evaluation values between

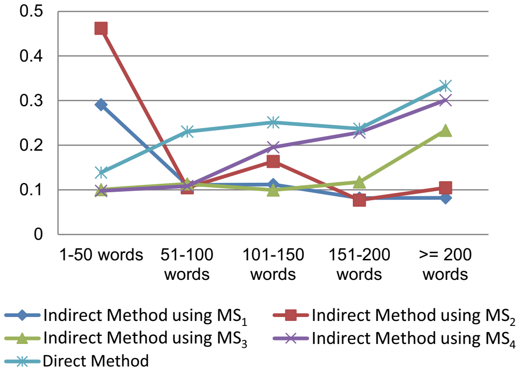


Fig. 16 $nhCG_{HDnoreduce}$@10 when Q_k is large

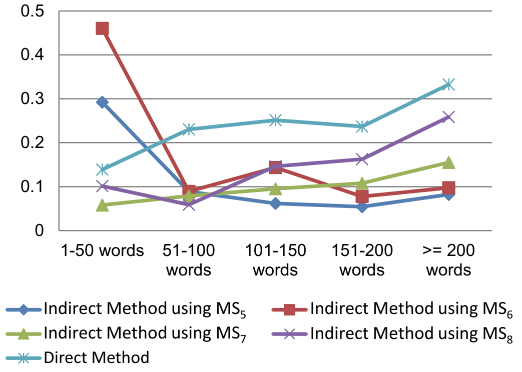


Fig. 17 $nhCG_{HDnoreduce}$@10 when Q_k is small

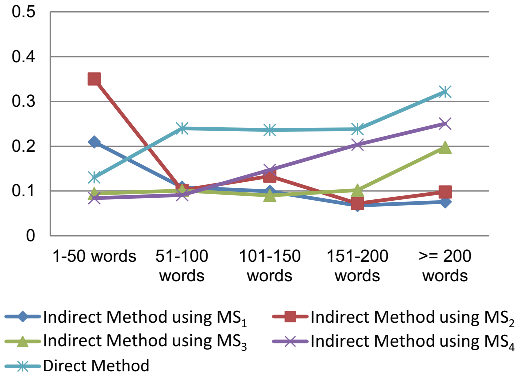


Fig. 18 $nhCG_{HDreduce}$@10 when Q_k is large

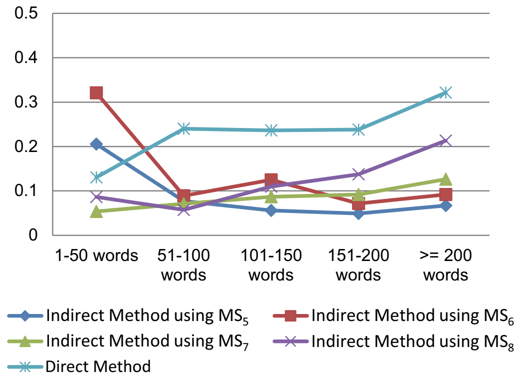


Fig. 19 $nhCG_{HDreduce}@10$ when Q_{k} is small

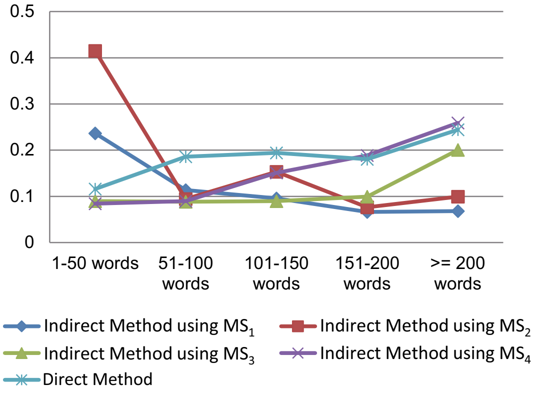


Fig. 20 $nhCG_{DSnoreduce}@10$ when Q_{k} is large

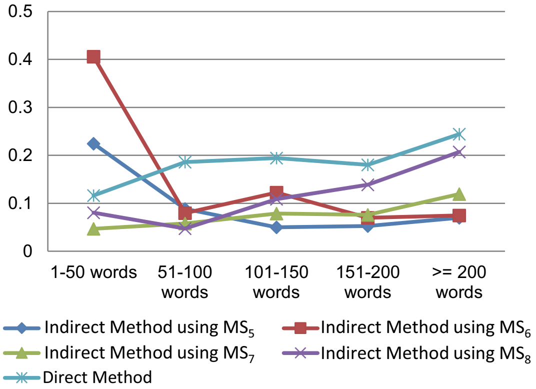


Fig. 21 $nhCG_{DSnoreduce}@10$ when Q_{k} is small

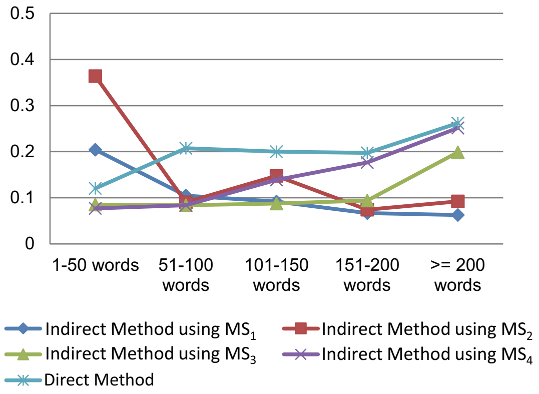


Fig. 22 $nhCG_{DSreduce}@10$ when Q_{k} is large

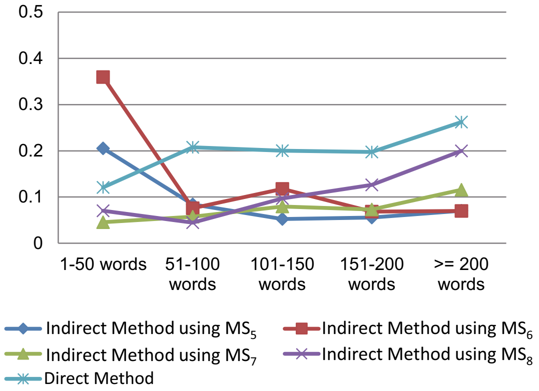


Fig. 23 $nhCG_{DSreduce}@10$ when Q_{k} is small

the following pairs of existing metadata sets: (MS$_1$, MS$_3$), (MS$_2$, MS$_4$), (MS$_5$, MS$_7$) and (MS$_6$, MS$_8$). Except in the case of the target datasets with abstract texts of fewer than 50 words, the evaluation values obtained using MS$_3$, MS$_4$, MS$_7$ and MS$_8$ tend to be higher than those obtained using MS$_1$, MS$_2$, MS$_5$ and MS$_6$. This finding demonstrates that Q_{t} influences the performance of the indirect method. Moreover, the results indicate that as the number of words in the abstract text of the target dataset increases, the differences in the evaluation values due to Q_{t} become larger. This behavior can be explained as follows: if the metadata of the target dataset include sufficient information for the recommendation of suitable keywords, then the recommendation accuracy improves in proportion to an increasing amount of information available in the existing metadata set. From this result, it can be recommended that a data provider should provide an abstract text with a sufficient description of the corresponding dataset.

In the case that the abstract texts in both the existing and target metadata sets consist of fewer than 50 words, however, the evaluation values are significantly higher. To determine the cause of this phenomenon, we investigated the corresponding datasets. We found that for many of the target datasets, their abstract texts contained descriptions that were nearly identical to those in the existing metadata. For example, some data providers produce multiple abstract texts from the same template by simply replacing the names of the observation satellite, the research vessel and the observation location. The metadata of these datasets are likely to have been mass produced through copying and pasting, and the annotated keywords for these datasets are also nearly the same. Thus, the evaluation values of the indirect method become high as it can utilize the information of nearly identical metadata in the existing metadata sets. Such mass production of metadata provides an additional indication that the cost incurred by data providers for metadata annotation is very high.

The number of annotated keywords To investigate the influence of Q_k, we consider the changes in the evaluation values between the following pairs of existing metadata sets: (MS_1, MS_5), (MS_2, MS_6), (MS_3, MS_7), and (MS_4, MS_8). We can observe the changes by comparing lines of the same color between Figs. 16 and 17, 18 and 19, 20 and 21, and 22 and 23. The results show that when Q_k is small, the evaluation values tend to be lower; however, Q_k has less influence on the results of the indirect method in the cases of existing metadata sets with small Q_t (MS_1, MS_2, MS_5, and MS_6). This finding indicates that when Q_t is small, the recommendation accuracy cannot be improved by increasing Q_k. From these results, we conclude that Q_k affects the results, but Q_t should receive greater emphasis.

The number of existing datasets with available metadata In each of Figures 16, 17, 18, 19, 20, 21, 22 and 23, we consider the influence of Q_n on the performance of the indirect method. Because we should compare existing metadata sets of comparable Q_t, we consider the changes in the evaluation values between the following pairs of existing metadata sets: (MS_1, MS_2), (MS_3, MS_4), (MS_5, MS_6) and (MS_7, MS_8). The results show that the evaluation values obtained using MS_2, MS_4, MS_6 and MS_8 tend to be higher than those obtained using MS_1, MS_3, MS_5 and MS_7. This indicates that Q_n affects the results of the indirect method. If Q_n is small, then the number of comparisons considered when calculating the similarities is insufficient, resulting in a low recommendation accuracy. In addition, one of the reasons for this result is that Q_n directly affects the accuracy of the topic distribution obtained via LDA, which is a kind of machine learning technique.

Discussion The results presented in this section show that all three quality measures considered in this study influence the results of the indirect method. Therefore, the performance of the indirect method tends to depend on the quality of the existing metadata set, and there is a high likelihood that the indirect method cannot offer reliable accuracy and comprehensiveness, especially when the existing metadata quality is insufficient. When MS_4, whose quality is sufficient in terms of all three measures, is used as the reference for the indirect method, the evaluation values tend to be highest. By contrast, when MS_5, whose quality is insufficient in terms of all three measures, is used as the reference, the evaluation values tend to be lowest. These observations provide an additional indication that the metadata quality influences the performance of the indirect method. As mentioned in Sect. 1, the quality of the metadata registered in DIAS can generally be regarded as insufficient. Therefore, it can be expected that if we were to use this metadata set as the existing metadata set for keyword recommendation, the indirect method would be ineffective.

5.3 Comparison between the direct and indirect methods

In this section, we compare the direct and indirect methods. First, in the case that the abstract texts in both the existing and target metadata sets consist of fewer than 50 words, the evaluation values obtained using the direct method are lower than those obtained using the indirect method. The cause of this phenomenon is described in Sect. 5.2. With the exception stated above, the results in Figs. 16, 17, 18, 19, 20, 21, 22 and 23 show that the evaluation values achieved using the direct method are higher in most evaluation metrics than those achieved using the indirect method. To determine the reasons for this behavior, we investigated the results obtained by applying these two methods to several datasets. We found that regardless of the characteristics of the existing metadata set used, the indirect method tends to recommend keywords from the top layer of GCMD Science Keywords. This is because Eq. (2), which describes the keyword relevance calculation used in the indirect method, depends on the frequencies of the keywords with which the existing datasets are annotated, and keywords in the top layer appear more frequently. By contrast, our evaluation metrics assign a larger basic gain to keywords that are more difficult for data providers to select. Therefore, when the proportion of the correctly recommended keywords that are located in the top layer is higher, the evaluation values will inevitably be lower. For the recommendation of top-layer keywords in a controlled vocabulary with a hierarchical structure, the indirect method is likely to be more appropriate than the direct method, but it cannot support the selection of keywords in lower layers, which are more difficult to select. In other words, the indirect method cannot effectively reduce the cost incurred by data providers for keyword annotation. By contrast, the direct method is independent of the frequencies of the annotated keywords in

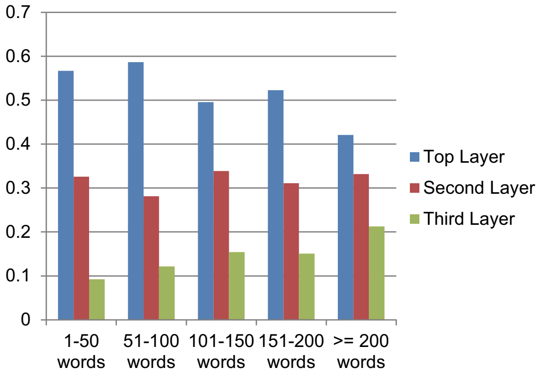


Fig. 24 The proportions of keywords in each layer recommended by the indirect method

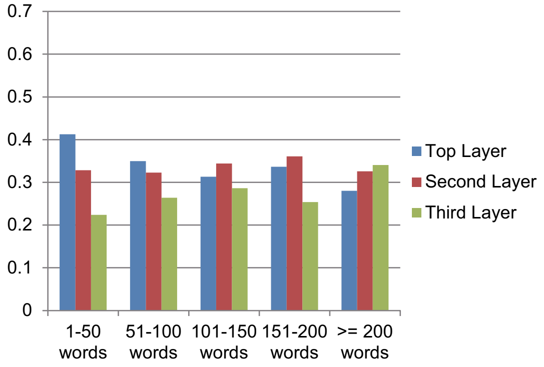


Fig. 25 The proportions of keywords in each layer recommended by the direct method

the existing metadata because those metadata are not used. Therefore, the direct method tends to recommend more keywords in lower layers, which are more difficult to select, and can reduce the cost of keyword annotation to a greater extent. To illustrate this behavior, as shown in Figs. 24 and 25, we calculated the proportions of the correctly recommended keywords belonging to each hierarchical depth for each method. Because of space limitations, only the ratios among the top three hierarchical depths are shown in Fig. 24, which presents the ratios found when using MS_7.

Notably, the evaluation values achieved using the direct method are often higher than even those achieved using the indirect method with MS_4, whose quality is sufficient in all three considered measures. This finding indicates that the indirect method, which depends on the keyword frequencies, has a tendency to recommend keywords in the top layer even when the existing metadata quality is sufficient.

Finally, the results of the direct method in Figs. 16, 17, 18, 19, 20, 21, 22 and 23 show that when the abstract texts in the target metadata contain fewer than 50 words, the evaluation values are the lowest among all target metadata sets. This observation demonstrates that the keyword recommendation accuracy that is achieved by the direct method depends on the number of words in the target abstract text. Therefore, we recommend that data providers should be sure to provide abstract texts that contain sufficient information to adequately describe their datasets.

6 Conclusion and future works

6.1 Summary

To support keyword annotation for data in various research domains, we have discussed methods of recommending keywords drawn from a controlled vocabulary. We believe that the *indirect method* of keyword recommendation is likely to be ineffective when the quality of the existing metadata set is insufficient. By contrast, the *direct method* utilizes the definition of each keyword in addition to the abstract text provided in the metadata associated with the recommendation target, and therefore, it does not depend on the quality of an existing metadata set. To verify the effectiveness of the direct and indirect methods, we created several metadata sets of differing quality in terms of the number of existing datasets with available metadata, the length of the abstract texts and the number of annotated keywords per dataset. We observed the changes in various evaluation values that occurred when these different datasets were used as the existing metadata set for keyword recommendation, and we showed that the performance of the indirect method depends on the quality of the existing metadata set. This observation shows that it is important to consider metadata quality when we adopt keyword recommendation methods. Additionally, we proposed evaluation metrics that consider a hierarchical vocabulary structure, which is a distinctive feature of most controlled vocabularies. Using our evaluation metrics, we evaluated the extent to which the cost incurred by data providers for keyword annotation is reduced through keyword recommendation. In our experimental evaluations, we compared the indirect method with the direct method using the proposed metrics. The results show that the indirect method tends to recommend keywords in the highest layer, which are easy to select, whereas the direct method tends to recommend keywords in lower layers, which are more difficult to select.

6.2 Future works

In this study, we considered keyword recommendation using only the keywords that are included in GCMD Science Keywords. In the future, we wish to investigate other controlled vocabularies in the earth science domain. We are also interested in applying our analysis to various other research domains, such as agriculture, chemistry, biology and computer science, in addition to earth science. Several such controlled vocabularies are available, including the CAB Thesaurus, the ACM Computing Classification System, and MeSH, which are relevant to agriculture, computer science and the life sciences, respectively. Each controlled vocabulary has a hierarchical structure, in which each domain is further subdivided into more specific domains. Notably, the direct method requires keyword definitions, but not all controlled vocabularies include keyword definitions. Therefore, by creating a definition for each keyword from words related to that keyword, we wish to improve the direct method to make it suitable for application to a controlled vocabulary that does not include keyword definitions.

Moreover, we would like to improve the direct and indirect methods with more consideration on a hierarchical vocabulary structure. In this paper, we used simple implementations for the direct and indirect methods in order to focus on the discussion of metadata quality. It is possible to handle each keyword considering the position of the keyword in the controlled vocabulary. Also, we are interested in integrating the direct and indirect methods. Because both of these methods are based on similarity calculations, the assignment probability of each keyword for a recommendation target can be acquired as a linear combination of the similarities calculated using the two methods. For this purpose, we need to define a parameter α to indicate how much emphasis should be placed on each method. The value of this parameter α should depend on the quality of the existing metadata set that is available for reference. In addition, we can consider utilizing the concept of keyword co-occurrence. In the earth science domain, sets of two or more keywords can be identified that tend to co-occur. For example, when researchers observe wind intensity, they also simultaneously observe wind velocity. However, an approach that relies on keyword co-occurrence will necessarily depend on the quality of the existing metadata set.

Finally, in our future work, we would like to incorporate our approach into a real metadata management system and obtain feedback on its performance from researchers and experts in earth science. Because we are assuming the framework that the final keyword selection from the recommended keywords is done by data providers, we would like to consider the significance of keyword recommendation with the keyword selection process.

Acknowledgements We have discussed how to proceed with this research with Dr. Hiroko Kinutani and Dr. Masafumi Ono at the University of Tokyo. We wish to express our deep appreciation for their cooperation. In addition, the datasets used in this paper were collected and provided within the frameworks of DIAS and the GCMD. We would like to express our deep gratitude to all those who work on and support these projects.

References

1. Belém, F., Martins, E.F., Pontes, T., Almeida, J.M., Gonçalves, M.A.: Associative tag recommendation exploiting multiple textual features. In: Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1033–1042 (2011)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Bruce, T.R., Hillmann, D.I.: The continuum of metadata quality: defining, expressing, exploiting. In: Metadata in Practice, ALA Editions (2004)
4. Chernyak, E.: An approach to the problem of annotation of research publications. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 429–434 (2015)
5. Dangerfield, M.C., Kalshoven, L.: Report and recommendations from the task force on metadata quality. Technical Report, Europeana (2015). https://pro.europeana.eu/post/metadata-quality-task-force-report
6. Fetahu, B., Dietze, S., Nunes, B.P., Casanova, M.A., Taibi, D., Nejdl, W.: A scalable approach for efficiently generating structured dataset topic profiles. In: Proceedings of the 11th European Semantic Web Conference, pp. 519–534 (2014)
7. Global Change Master Directory (GCMD): GCMD Keywords, Version 8.6. Greenbelt, MD: Earth Science Data and Information System, Earth Science Projects Division, Goddard Space Flight Center (GSFC) National Aeronautics and Space Administration (NASA). URL (GCMD Keyword Forum Page) (2019). https://wiki.earthdata.nasa.gov/display/gcmdkey
8. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. **101**(suppl 1), 5228–5235 (2004)
9. Ishida, Y., Shimizu, T., Yoshikawa, M.: A keyword recommendation method using CorKeD words and its application to earth science data. In: Proceedings of the 11th Asia Information Retrieval Societies Conference, pp. 96–108 (2015)
10. Kawasaki, A., Yamamoto, A., Koudelova, P., Acierto, R., Nemoto, T., Kitsuregawa, M., Koike, T.: Data integration and analysis system (DIAS) contributing to climate change analysis and disaster risk reduction. Data Sci. J. **16**(41), 1–17 (2017)
11. Krestel, R., Fankhauser, P., Nejdl, W.: Latent dirichlet allocation for tag recommendation. In: Proceedings of the 2009 ACM Conference on Recommender Systems, pp. 61–68 (2009)
12. Lalmas, M., Tombros, A.: Evaluating XML retrieval effectiveness at INEX. SIGIR Forum **41**(1), 40–57 (2007)
13. Lu, Y., Yu, S., Chang, T., Hsu, J.Y.: A content-based method to enhance tag recommendation. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence, pp. 2064–2069 (2009)
14. Ochoa, X., Duval, E.: Automatic evaluation of metadata quality in digital repositories. Int. J. Digit. Libr. **10**(2–3), 67–91 (2009)
15. Paik, J.H.: A novel TF-IDF weighting scheme for effective ranking. In: Proceeding of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 343–352 (2013)
16. Ramage, D., Hall, D.L.W., Nallapati, R., Manning, C.D.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 248–256 (2009)
17. Ramnandan, S.K., Mittal, A., Knoblock, C.A., Szekely, P.A.: Assigning semantic labels to data sources. In: Proceedings of the 12th European Semantic Web Conference, pp. 403–417 (2015)
18. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Boston (1989)
19. Santos, A.P., Rodrigues, F.: Multi-label hierarchical text classification using the ACM taxonomy. In: Proceedings of the 14th Portuguese Conference on Artificial Intelligence, pp. 553–564 (2009)
20. Schaible, J., Gottron, T., Scherp, A.: Termpicker: Enabling the reuse of vocabulary terms by exploiting data from the linked open

data cloud. In: Proceedings of the 13th International Semantic Web Conference, pp. 101–117 (2016a)

21. Schaible, J., Szekely, P.A., Scherp, A.: Comparing vocabulary term recommendations using association rules and learning to rank: a user study. In: Proceedings of the 13th International Semantic Web Conference, pp. 214–230 (2016b)
22. Shimizu, T., Sueki, T., Yoshikawa, M.: Supporting keyword selection in generating earth science metadata. In: Proceedings of the 37th Annual IEEE Computer Software and Applications Conference, pp. 603–604 (2013)
23. Tani, A., Candela, L., Castelli, D.: Dealing with metadata quality: the legacy of digital library efforts. Inf. Process. Manag. **49**(6), 1194–1205 (2013)
24. Tuarob, S., Pouchard, L.C., Giles, C.L.: Automatic tag recommendation for metadata annotation using probabilistic topic modeling. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 239–248 (2013)