地球科学データに対するキーワード推薦手法

石田 陽一[†,a]　清水 敏之[†,b]　吉川 正俊[†,c]

† 京都大学大学院情報学研究科

a) *yishida@db.soc.i.kyoto-u.ac.jp* b) *tshimizu@i.kyoto-u.ac.jp* c) *yoshikawa@i.kyoto-u.ac.jp*

概要　近年，農業や海洋等の様々な分野に関する地球科学データが爆発的に増加しているが，各分野は専門化かつ細分化する方向で発達してきており，分野間でのデータの統合的利用は困難である．そのため，地球環境情報統融合プログラム (DIAS-P) では，分野間でデータを相互利用するためのデータ基盤を構築している．しかし，データの意味や説明の記述を行うメタデータの記述が不十分なデータセットも数多く存在し，互いに関連するデータセットの発見は困難である．そこで本研究では，科学キーワード集の一つである GCMD Science Keyword に着目し，各データセットに対し，適切なキーワードを推薦することで，データの統合的利用の支援を行う．各語の，他分野と地球科学分野における出現傾向を解析することで，地球科学分野特徴語を抽出する．そして，その抽出情報をキーワード定義文やデータセット概要文に適用することで、キーワードを推薦する手法を検討した。

キーワード　地球科学データ，メタデータ，分野特徴語抽出，キーワード推薦，分野連携

1　はじめに

近年，地球観測技術の発達や情報技術の進歩により，多種多様で膨大な量の地球科学データが収集，蓄積，管理されている．それらの地球科学データは，社会に有益な情報へと変換され，環境問題や自然災害への対応，水産資源管理や農業生産管理などへ利用されている．また，農業，海洋，気候など多種多様な地球科学データが爆発的に増加している一方，異なる分野間でデータを統合利用することによる新たな知見の発見が期待されている．

地球科学データは単なる数値や文字の並びでしかなく，非常に専門性が高いものであるので，データ利用者にとって，そのデータの内容理解は非常に困難である．そのため，データ提供者は，そのデータに対してメタデータを付与し，データの取得期間や取得場所，観測内容など，そのデータの意味や説明の記述を行うことで，データ自体の理解支援を行う．そして，それらのメタデータを適切に収集，管理，検索するために，多くのメタデータポータルや地球科学データベースが存在する．

例えば，GCMD(Global Change Master Directory)[1)] と呼ばれる，アメリカ航空宇宙局 NASA が管理している地球科学データのメタデータポータルが存在する．GCMD は，多種多様なメタデータに対する検索機能の提供や，GCMD Science Keywords[1] と呼ばれる科学キーワード集などの様々な統制語彙を管理している．

この GCMD は海外のものであるが，国内でも地球環境情報統融合プログラム [2] (DIAS-P)[2)] と呼ばれる文部科学省主導のプロジェクトが存在する．DIAS-P では，異なる分野間でデータを相互に運用できるデータ基盤の構築を目指している．DIAS-P が管理しているデータベースでは，メタデータを利用して，多様な視点からデータセットを検索できる DIAS データ俯瞰・検索システム [3] が構築されている．

本研究では，この DIAS-P に注目する．DIAS-P では，様々なデータ提供機関からデータやメタデータを収集しており，データセット単位でメタデータを作成する基本方針がとられている．収集したデータに予めメタデータが付与されていない場合は，データ提供者が，専用のウェブツールを用いて，データセット名，問い合わせ先，時空間情報，概要文，キーワードなど，様々なメタデータ項目を手動で入力していく必要がある．

我々は，これらのメタデータ項目の中の “キーワード” の記述量に着目した．メタデータ項目におけるキーワードとは，ある統制語彙の中から，そのデータセットに関連する語彙を選択するものであり，そのキーワード情報により，データセットの分類・関連の取得が可能となる．しかし，実際に，DIAS-P が管理するメタデータにおけるキー ワード付与状況を調査したところ，入力がなされていないものが散見された．原因の一つとして，統制語彙全体の把握が困難であることが挙げられる．統制語彙の中から適切な語彙を選択するには，その統制語彙に関する知識と語彙全体の把握が必要であるが，例えば GCMD Science Keywords のような，2000 語以上もの語彙が階層化されて管理されている統制語彙も存在するので，その全体把握は非常に困難である．そのため，キーワードの入力が不十分なデータセットが多々存在し，データセットの分類・関連の取得が困難な状態である．

そこで，データ提供者が各メタデータを記述時する際に，キーワードをランキング形式で推薦することで，データセットの分類や統合的利用の支援ができると考え

1) http://gcmd.nasa.gov/

2) http://www.editoria.u-tokyo.ac.jp/dias/

た．我々は，各キーワードの意味を説明した定義文を利用し，各キーワードの持つ内包的な情報まで考慮に入れた手法を提案する．まず，地球科学分野と，その他の分野における各語の出現傾向を解析することで，地球科学分野に特徴的な単語を抽出する．そして，その抽出情報を，各キーワードの定義文やデータセットの内容を表す概要文に適用することで、キーワードを推薦する手法を検討した。なお，本研究では，英語で記述されたメタデータを対象としている．

本論文の構成を以下に示す．第 2 節では関連研究について述べる．第 3 節では，地球科学分野特徴語の抽出法やキーワード定義文の利用法を含めたキーワード推薦手法の具体的な内容を示す．第 4 節では評価実験について述べ，第 5 節では，まとめと今後の課題について述べる．

2 関連研究

地球科学データを対象としたタグ推薦の研究として，十分な数のタグが付与されていないデータに対し，そのデータと類似しているデータに付与されたタグを推薦する Tuarob らの研究 [4] がある．各メタデータのテキスト情報から，各データセットの特徴ベクトルを導出し，そのベクトル間のコサイン類似度を測定する．特徴ベクトルの各要素には，TF-IDF 値や LDA の確率分布を用いている．しかし，この研究の手法では，テキスト情報が少ない場合には有効に働かない．それに対し我々は，概要文が短いデータセットに対しても，その概要文に含まれる地球科学に特徴的な単語の情報を介して，キーワードを推薦できるような手法を提案している．

また，我々はこれまでにも地球科学データのメタデータ中のキーワードの付与支援を行う研究を行ってきた [5, 8]．これらの研究では，推薦する語彙を，分野を表す 14 個のキーワードに限定し，機械学習の一つである Labeled LDA を利用する．14 個のキーワードをラベルとして，データセット概要文とキーワードとの対応関係を学習させ，その学習結果を対象データセットに適用することで，キーワード推薦を行う．この手法はラベルの数が少ないので有効に働いているが，今回我々が対象とする 2000 語以上もの語彙集合をラベルとして Labeled LDA を適用すれば，推薦の精度は極端に低下するだろう．

また，Web ページなどのリソースに対するソーシャルタギングを支援する Krestel らの研究 [9] があり，タグ集合に LDA を適用する手法が提案されている．ソーシャルタギングの場合，数少ないユーザにより付与された特異なタグは重要ではないことが多いが，地球科学データの場合，少ない頻度でしか出現しない重要な単語は少なからず存在する．同様に LDA を用いて 2000 語以上もの語彙に対する豊富な学習データを用意することは，非常に困難である．

- 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

図 1 “Aqua/AMSR-E Satellite dataset”に付与されているキー ワード

3 キーワードの推薦手法

本研究では，各データセットのメタデータ中の「概要文」の情報を利用，解析することで，キーワードを推薦する手法を提案する．各データセットのメタデータでは，そのデータセットの意味を説明する概要文を記述することが一般的であり，各概要文を閲覧することで，そのデータセットの意味を大まかに把握する事ができる．また，推薦されるキーワードは，地球科学に関する統制語彙の一つである GCMD Science Keywords に収録されている語彙とした．DIAS-P が管 理しているデータセットを例にとり，GCMD Science Keywords の利用によるキーワードの付与例を図 1 に示す．

まず，我々は，各データセット概要文と GCMD Science Keywords に収録されている語彙を文字列マッチングさせ，一致したものをデータセットに関連するキーワードとして推薦する手法を検討した．しかし，先述した GCMD では，1 データセットあたり平均約 12 個のキーワードが付与されているのに対し，この文字列マッチングの手法では，平均約 2.7 個のキーワードを推薦できるのみである．データセット概要文が短いデータセットも多く存在し，十分な数のキーワードを推薦することは困難であったので，我々はさらに別の手法を検討した．

3.1 キーワード定義文の利用法

そこで，本研究では，データセット概要文やキーワード名という表面的な情報を使うだけでなく，各キーワードの定義文という内包的な情報まで考慮に入れた手法を提案する．GCMD Science Keywords に収録されているキーワードには，そのキーワードの意味を説明した定義文が存在する．我々は，データセット概要文中の単語が多く含まれるような定義文を持つキーワード，または，それらの単語をキーワード名中の一部に含むようなキーワードを推薦すべきと考えた．しかし，地球科学分野に関連のない単語を介したことによる誤った推薦を行う可能性があるので，我々は，地球科学に特徴的な単語 (以下，地球科学分野特徴語と呼ぶ) を定義し、予め地球科学分野特徴語リストを作成する必要があると考えた．提

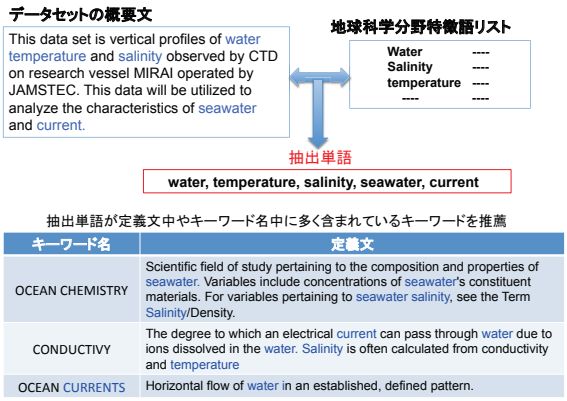

図 2 キーワード定義文の利用法

案手法の概要 (図 2) としては，まず，データセット概要文と予め作成した地球科学分野特徴語リストを照合した結果，一致した単語 (以下，抽出単語と呼ぶ) を抽出する．そして，それらの抽出単語を定義文中やキーワード名中に多く含むキーワードを推薦すべきという考えを基に，定義文中やキーワード名中に含まれる各抽出単語に付与された重みの値を利用し，各データセットに対する各キーワードの適合度を算出する．地球科学分野特徴語の定義，各単語への重みの付与，キーワードの適合度算出法など，提案手法の具体的内容は 3.2 節以降で述べる．なお，全単語にステミング処理を施した上で，提案手法を適用する．

3.2 地球科学分野特徴語の定義

久保ら [6] が指摘するように，ある対象分野における特徴語とは，対象分野と他分野それぞれにおける単語の出現頻度を比較し，対象分野における出現頻度が相対的に高い単語であると考えられる．つまり，他分野と比較して，地球科学分野に相対的に高頻度で出現する単語を，地球科学分野特徴語と定義する．本研究では，他分野として，地球科学と同じ自然科学分野に属する，生物学，化学，物理学の 3 つの分野を利用する．自然科学分野以下に属する分野間で比較する方が，社会科学分野や人文科学分野と比較するより，厳密に地球科学分野特徴語を抽出できると考えたためである．

3.3 各分野のコーパス

分野間で比較するためには，各分野のコーパスが必要となる．地球科学分野のコーパスとしては，American Geophysical Union(以下 AGU)[3] と呼ばれる地球科学に関する学会の，2013 Fall Meeting における発表要旨をまとめた概要文集合を用いる．20028 件の概要文から約 600 万語の単語を取得した．

他分野のコーパスとしては，各分野に対応するジャーナルに含まれる論文の概要文集合を用いた．利用したジャーナルの詳細は付録に記載している．また，他分野のコーパスサイズを等しくする必要があるので，今回は約 20 万語で統一した．

3.4 地球科学分野特徴語リストの作成法

地球科学分野特徴語リストを作成するためには，各単語に対し，分野間での相対的な出現頻度を比較する必要がある．本研究では，分野間での相対的な出現頻度を求めるために，久保ら [6] が提案している DP(*the Difference between Population Proportions*) と呼ばれる計算式を利用する．この計算式は，統計学における母比率の差の検定に基づくものである．以下に詳細を示す．なお，今回はストップワードを除いた単語一語毎に対して相対的出現頻度の計算を行う．

$$DP_c(t) = \frac{\frac{f_0(t)}{W_0} - \frac{f_i(t)}{W_i}}{\sqrt{\pi_i(t)(1-\pi_i(t))\left(\frac{1}{W_0}+\frac{1}{W_i}\right)}} \quad (1)$$

$$\pi_i(t) = \frac{f_0(t)+f_i(t)}{W_0+W_i} \quad (2)$$

$f_0(t)$ は単語 t の地球科学分野コーパス内での出現回数，$f_i(t)$ は単語 t の他分野コーパス内での出現回数，W_0 は地球科学分野コーパスの総単語数，W_i は各他分野コーパスの総単語数である．集合 C の要素は { 生物学，化学，物理学 } であり，$DP_c(t)$ は分野 $c \in C$ との比較時における単語 t の相対的出現頻度を表す．この $DP_c(t)$ は正規分布に従う．そして，地球科学分野と他分野を比較した際の，各単語 T の相対的出現頻度 $w(t)$ を以下のように求める．

$$w(t) = \frac{\sum_{c \in C} DP_c(t)}{|C|} \quad (3)$$

$|C|$ は集合 C の大きさであり，今回は $|C| = 3$ である．(3) 式により，各分野比較時に算出される相対的出現頻度の平均を計算し，$w(t) > 0$ となる単語 t を地球科学分野特徴語リストに含めることとする．

実際に先述のコーパスを利用し，$w(t)$ を計算した．表 1 に $w(t)$ の値の上位 10 件を提示する．

上位に出現している単語は全て，地球科学に特徴的な単語と考えられる．しかし，“data”, “model” , “region” といった推薦すべきキーワードを決定付ける力を持っていないような単語も上位に出現したので，我々はさらに，これらのキーワードの決定力を持たない単語を除去する手法について検討した．

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

表 1 各単語 t に対する $w(t)$ 上位 10 件

単語 t	$w(t)$	単語 t	$w(t)$
data	27.89	soil	19.88
model	24.46	atmospher	19.00
climat	24.26	fault	18.18
water	20.18	ic	18.15
region	20.03	event	18.07

3.4.1 キーワードの決定力の有無

地球科学分野の中には，大気，農業，海洋など，さらに細分化された分野 (以下，細分野と呼ぶ) が存在する．我々は，“climat”, “soil”, “atmospher” といったキーワードの決定力を備えた単語は，それらの細分野間で出現頻度に偏りがあり，一方で，“data”, “model”, “region” といったキーワードの決定力を持たない単語は，その細分野に依存せずに平均的に出現すると仮定した．例えば，“climat(気候)” という単語は，地球科学分野内の “大気” という細分野に偏った形で出現するだろうが，“data” という単語は，どの細分野にも偏りなく出現するだろう．このように，各単語に対し細分野間での出現頻度分布の偏りを数量化することで，キーワードの決定力の有無を判別できると考えた．

本研究では，細分野として，AGU が提供する科学キーワード集である AGU index terms[4] での分類軸に用いられている 49 個の分野を利用した．そして我々は，分布の偏りを数量化する方法として一般的である χ 二乗値を用いた．χ 二乗値とは，観測度数と期待度数の適合度を数量化した値である．また，本研究では，その単語が出現する AGU での発表の概要文の件数を度数とみなし，細分野毎に実際に測定した度数をその単語の観測度数，各対象単語が細分野に依存せずに平均的に出現すると仮定した場合の度数をその単語の期待度数と考えた．以下に，その詳細を示す．

$$\chi^2(t) = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{4}$$

$$E_i = A_i \times \frac{A_t}{D} \tag{5}$$

n は細分野数を表し，今回は $n = 49$ である．O_i は i 番目の細分野における観測度数，E_i は i 番目の細分野における期待度数，A_t は単語 t が含まれる概要文件数，A_i は i 番目の細分野に属する概要文件数を表す．また，D は概要文総件数を表し，今回用いた AGU 2013 Fall Meeting における概要文総件数は 20028 件である．

また我々は，キーワードの決定力がない単語は，AGU での発表の，いずれの概要文にも出現する可能性が高いと仮定した．AGU での概要文集合に対し，各単語の DF 値 (document frequency) を求め，その DF 値上位 0.5%に含まれる単語に対し，χ 二乗検定値を計算した．実際の計算結果の一部を表 2,3 に示す．

表 2 決定力の無い単語

単語 t	$w(t)$
data	660.8
model	695.1
region	801.8
time	282.3
base	352.5

表 3 決定力の有る単語

単語 t	$w(t)$
climate	5735.9
water	3678.5
soil	3439.1
atmospher	4375.0
temperatur	1729.7

表 2,3 より，キーワードの決定力がある単語に対する χ 二乗値は，比較的大きな値を示している．これは，これらの単語が一部の細分野に偏った形で出現していることを指している．一方で，キーワードの決定力がない単語は χ 二乗値の値が小さい．これは，これらの単語が細分野に依存せずに出現していることを指している．我々は，計算結果から閾値を 1700 と設定し，χ 二乗値が閾値以下の単語を 3.4 節の地球分野特徴語リストから省くことで，キーワード推薦の精度向上に努めた．

3.5 各データセットに対するキーワードの適合度算出法

各データセットに対する推薦キーワードの適合度を算出するために，3.4 節で述べた各単語の相対的出現頻度を利用する．相対的出現頻度の値は，いかに地球科学に特徴的な単語であるかの度合を表す．そこで，相対的出現頻度の値を，各単語に付与される重みと考えた．適合度の算出法としては，まず，データセット概要文中と地球科学分野特徴語リストを照合し，抽出単語を求める．そして，定義文中やキーワード名中の一部に含まれる抽出単語の個数や重みの値から、以下のように，各データセットに対する各キーワードの適合度を算出する．

$$ndl = (1 - s) + (s \times \frac{dl}{avgdl}) \tag{6}$$

$$k\_score = \frac{\sum_{t \in E} w(t)}{ndl} \tag{7}$$

定義文に含まれる抽出単語の個数は，その定義文の文書長に依存するので，Liu らの研究 [7] でも利用された ndl により，定義文の平均文書長で正規化を行った．dl は定義文の文書長，$avgdl$ は定義文の平均文書長を表す，s はパラメータであり，通常は $s = 0.2$ と設定する．E は抽出単語を要素とする集合であり，集合 E の要素である各単語の重みの総和を ndl で正規化することで，各データセットに対するキーワードの適合度 k_socre を算出する．つまり，k_score の値を昇順に並べることで，キーワードをランキング形式で推薦する．

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

4 評価実験

4.1 実験概要

提案手法の有用性を確かめるため，提案手法を適用した結果得られた推薦キーワード一覧を，実際の各データ提供者へ提出し，各キーワードが有用か否かを判定した．実験として，DIAS-P が管理しているデータセット 20 個を対象とした．今回の実験では，地球科学分野特徴語リスト作成と，χ 二乗値を利用した各単語への重み付与の有用性を確かめる．評価実験の結果を，表 4 に示す．表 4 では，各手法により推薦されたキーワード上位 10 件中における適合率を示している．

4.2 地球科学分野特徴語リスト作成に関する実験

地球科学分野特徴語リストの作成による効果のみを確かめるため，地球科学分野特徴語リストと照合せずに，データセット概要文中に含まれる全単語 (ストップワードは除く) を抽出単語とみなす手法 (表 4 における手法 1) を考える．そして，その比較対象として，全単語の重みを 1 とした地球科学分野得特徴語リストを適用する手法 (表 4 における手法 2) を考える．この二つの手法を比較した結果，表 4 より，一部のデータセットにおいて，適合率が向上した．適合率が向上したデータセットでは，“degree” や “study” など，地球科学分野には関連がないと考えられる単語が多く出現していた．予め地球科学分野特徴語リストを作成することで，これらの単語を抽出単語から排除したことが，適合率の向上につながったと考えられる．しかし，多くのデータセットでは，この二つの手法における適合率に変化は見られなかった．原因として，キーワードの定義文に含まれる多くの語彙が，既に地球科学分野特徴語である可能性が高いことが挙げられる．また，適合率が低下したデータセットも存在した．本来推薦に必要となる単語がリストに含まれなかったことに原因があると考え，今後は，リスト作成の精度を向上させる必要がある．

4.3 χ 二乗値利用に関する実験

χ 二乗値の利用効果を確かめるため，比較対象として，χ 二乗値を計算せずに作成した地球科学分野特徴語リストを適用する手法 (表 4 における手法 3) を，提案手法と比較する．その結果，ほぼ全てのデータセットにおいて，適合率が向上した．適合率が向上したデータセット概要文には，“data”,“field”,“system” など，キーワードの決定力を持たない単語が多く出現していた．この結果より，χ 二乗値を計算し，これらの単語を地球科学分野特徴語リストから省いたことが，適合率の向上につながったと考えられる．また，一部のデータセットで，適合率が低下した．原因として，閾値の設定の問題が挙げられる．例えば，“flux(流速)” という単語は，地球科学分野で頻出であるが，今回は χ 二乗値計算時に，閾値以下の単語として除去した．このように，今後は適切な閾値を定める必要がある．

4.4 各単語への重み付与の有用性に関する実験

各単語への重み付与の有効性を確かめるため，全単語の重みを 1 とした地球科学分野特徴語リストを適用する手法 2 を，提案手法と比較する．提案手法と比較した結果，一部のデータセットで，適合率が向上した．“sea” ,“ocean”, “climat”, “land” など，地球科学分野に顕著に出現し，かつ，キーワードの決定力を備えた単語に対し，比較的大きな重みを加えたことで，より適切な推薦を行えたと考えられる．特に，データセット ID : SSM_I では，各単語に重みを加えた効果が顕著に現れ，手法 2 では全く正解キーワードを推薦できなかったが，提案手法により複数の正解キーワードを推薦することができた．しかし，適合率が低下したデータセットも存在した．原因として，各データセットによって重要視する単語が変化する場合があるにも関わらず，データセットに依存せずに，各単語へ重みを付与したことが挙げられる．例えば，植生分布に関するデータセットの場合や，観測衛星によるデータセットの場合，他のデータセットよりもそれぞれ，“biosphere(生物圏)” という単語や “microwave(マイクロ波)” という単語を重視すべきだと考えられる．今後は，各データセットによって，可変的に各単語の重みを変化させることも視野に入れるべきである．

5 おわりに

5.1 まとめ

我々は，地球科学データに対してデータ提供者にキーワードを推薦することで，データセットの分類やデータの統合的利用の支援ができると考えた．提案手法の評価実験においては，一部のデータセットに対して，地球科学分野特徴語リストを作成した効果が見受けられ，また，細分野間の各単語の偏り度合を観測することで，キーワード推薦の精度を向上させることができた．しかし，提案手法が有意に働かないデータセットも存在したので，さらなる手法の改善を行う必要があると思われる．

5.2 今後の課題

今回は他分野として，生物学，化学，物理学という 3 つの分野を利用したが，今後は，地球科学分野特徴語リストの精度向上のため，社会科学分野や人文科学分野など，様々な分野も含めて比較する必要があると考える．

また，4.4 節で記したように，各データセットによって，可変的に各単語の重みを変化させることも考えるべきである．例えば，各データセットのタイトル名に含まれる単語に対して重みを加える方法や，各データセット概要文中の各単語の TF 値 (term frequency) を，重みの

表 4 各データセットに対するキーワード推薦の評価結果

データセット ID	手法 1	手法 2	手法 3	提案手法
D8NDVI_J	30%	40%	30%	50%
D8NDVI_J	40%	60%	60%	60%
Global_map	30%	20%	20%	10%
Global_map	30%	30%	20%	10%
MIRAI_CTD	10%	10%	10%	10%
AMY_HARIMAU_WPR_dataset	20%	20%	0%	20%
AVISO_SLA	30%	40%	0%	30%
DIAS_ODAPv2.1	50%	50%	40%	60%
DIAS_ODAPv2.1	50%	50%	40%	50%
MOM_rNP	30%	30%	30%	20%
MSST	0%	10%	10%	30%
ODA_rNPhigh	60%	60%	40%	60%
ODA_rNPhigh	80%	60%	50%	50%
SSM_I	0%	0%	10%	30%
MAHAPGP	20%	20%	10%	20%
ALOS_ANVIR2	20%	30%	20%	20%
ALOS_PALSAR	10%	10%	0%	10%
ALOS_PRISM	0%	0%	10%	0%
Aqua_AMSR_E	10%	10%	0%	20%
GCOM_W1	20%	30%	0%	50%
TRMM_PR	20%	20%	0%	0%
GPV	0%	0%	0%	20%
Fuji_Hokuroku_Flux	0%	20%	10%	10%
CEOP_CAMP_Eastern_Siberian_Taiga	20%	10%	10%	20%
適合率平均	24%	26%	18%	28%

(注: 1 データセットに複数の評価者が存在する場合は，同データセットに対する評価が複数行に記載されている)

値に反映する方法などが考えられる．

さらに，データセット概要文やタイトル名中に存在する，そのデータを収集した観測機器名の情報から，さらなるキーワードを推薦することが考えられる．

最後に，今回は，GCMD Science Keywords に収録されているキーワードのみ推薦対象としたが，今後は，AGU が管理している AGU index terms など，他の科学キーワード集も推薦対象として検討していきたい．

謝辞

研究全般に関して多くの助言を頂きました東京大学地球観測データ統融合連携研究機構の絹谷弘子氏，小野雅史氏，海洋研究開発機構の石川洋一氏に心より感謝致します．さらに，キーワード推薦の有効性検討に御協力下さいましたすべての方々に感謝致します．また，利用したデータセットは，地球環境情報統融合プログラム DIAS-P の枠組みの下で収集・提供されたものであります．ここに記して謝意を表します．

参考文献

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

[2] 絹谷弘子，清水敏之，吉川正俊，喜連川優：DIAS における多分野研究者連携による地球科学データ公開に向けた協働，電子情報通信学会技術研究報告，Vol. 110, No.328，pp. 45-50，2010.

[3] 清水敏之，絹谷弘子，吉川正俊：多様な地球科学データに対する俯瞰・検索システムの開発，電子情報通信学会技術研究報告，Vol. 110，No. 328，pp. 39-44，2010.

[4] 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(JCDL 2013), pp. 239-248, 2013.

[5] Shimizu, T., Sueki, T., Yoshikawa, M. : Supporting keyword selection in generating earth science metadata. In 37th Annual IEEE Computer Software and Applications Conference (COMPSAC 2013), pp. 603-604, 2013.

[6] 久保順子，辻 慶太，杉本重雄：異なる学問分野のコーパスを利用した専門用語抽出手法の提案，情報知識学会誌，Vol. 20，No. 1，pp. 15-31，2010.

[7] Liu, F., Yu, C., Meng, W., Chowdhury, A. : Effective keyword search in relational databases, Proceedings of the 2006 ACM SIGMOD international conference on Management of data, pp. 563-574, 2006.

[8] 石田陽一，清水敏之，吉川正俊：地球科学データに対するタグと検索語の推薦手法，第 5 回データ工学と情報マネジメントに関するフォーラム，2014.

[9] Krestel, R., Fankhauser, P., Nejdl, W. : Latent dirichlet allocation for tag recommendation. In Proceedings of the third ACM conference on Recommender systems, pp. 61-68, 2009.

付録： 本研究で使用したジャーナル一覧

- Journal of the American Chemical Society
- International journal of biological sciences
- Journal of evolutionary biology
- The European physical journal