統制語彙における階層構造を考慮した 地球科学データに対するキーワード推薦

城戸　譲次†　　清水　敏之††　　吉川　正俊††

† 京都大学工学部情報学科　〒 606-8501 京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606-8501 京都市左京区吉田本町
E-mail: †kido.joji.84c@st.kyoto-u.ac.jp, ††{tshimizu,yoshikawa}@i.kyoto-u.ac.jp

あらまし　地球科学データを理解し，適切に利用するにあたっては，データがどのようなものであるかを表す地球科学メタデータが十分な情報を含んでいることが重要である．メタデータ項目の一つとして，キーワードがあり，地球科学データに対してキーワードを付与する際には，階層構造を持つ統制語彙から適切なキーワードを選択して付与することが一般的である．キーワード情報はデータの検索，分類等に重要な役割を果たすが，多くキーワードを含む統制語彙から適切なキーワードを選択するコストは高く，実際に十分な数のキーワードが付与されていない場合も多い．本研究では，統制語彙中のキーワードに与えられた定義文を用いて地球科学データに対してキーワード推薦を行うことを考え，統制語彙における階層構造を考慮した推薦手法を提案する．

キーワード　キーワード推薦, 地球科学データ, メタデータ, 統制語彙, 階層キーワード

1　はじめに

1.1　研究背景

地球科学の発展によって，地球科学データは膨大かつ多様に存在している．また，大量のデータを解析する技術の発達により，地球科学データの分析の需要は非常に高まっている．このような背景から，ある地球科学データがどのような分野のデータであるかを理解すること，自分が分析に利用したいと考えている地球科学データを適切に検索できることは非常に重要となってくる．地球科学データの理解・検索においては，地球科学メタデータが非常に重要な役割を果たす．本研究で我々が想定している地球科学メタデータは，地球科学データに対してデータセット単位で与えられるものであり，データセット名，データセット作成者，データセット作成日，メタデータ作成者，メタデータ作成日，キーワード，データセット概要文などで構成されている．

このような地球科学メタデータは，データに対する知識を持っているデータ提供者が手作業で作成することが一般的である．地球科学メタデータの項目の一つとして，キーワード情報があるが，我々は，地球科学データの検索・理解においては，地球科学メタデータの中でも，特にキーワードが正確に付与されていることが重要だと考えている．ここで，我々が想定している地球科学メタデータにおけるキーワードは任意のキーワードではなく，統制語彙から選択して付与するものであり，地球科学の分野においては，GCMD サイエンスキーワード [3] の利用が一般的になっている．GCMD サイエンスキーワードは，3,000 語以上の地球科学に関する専門的なキーワードを含む統制語彙である．キーワードが正確についていることによって，地球科学データを利用したい人はそのデータがどの分野に関連したものかを大まかに理解することができる．また，キーワードが正確に付与されていることによって，自分が利用したいと考えている分野の地球科学データの検索の質が上がると考えられる．

しかしながら，データ提供者がメタデータを記述する際に多くのキーワードを含む統制語彙から適切なキーワードを選択するコストは高く，実際に，キーワードが正確に付与されていないと思われるものが非常に多い．地球科学データの管理を行っている DIAS (Data Integration Analysis System)[1] [5] では，各データセットに対してデータ作成者が地球科学メタデータの作成を行うが，項目の一つであるキーワードには GCMD サイエンスキーワードや AGU Index Terms，GEOSS 社会利益領域などのキーワード辞書が用いられている．図 1 は DIAS に存在する地球科学データに対する GCMD サイエンスキーワードの付与状況を示したものであるが，DIAS に存在する地球科学データ 519 個のうち，358 個は GCMD サイエンスキーワードが全く付与されていない．このような状況を改善するために，我々は，各データセットに対して適切と思われるキーワードを推薦することでキーワード付与を補助できると考え，地球科学データに対するキーワード推薦の手法を研究している．

1.2　キーワード推薦のアプローチ

GCMD サイエンスキーワードにおけるそれぞれのキーワードには，そのキーワードの説明を表すキーワード定義文が付随している．また，データベースに存在している既にキーワードやデータセット概要文などの地球科学メタデータが付与されている地球科学データのメタデータ集合を既存メタデータ集合と呼ぶ．キーワード推薦には，地球科学データの概要を文章で表

1：http://www.diasjp.net/

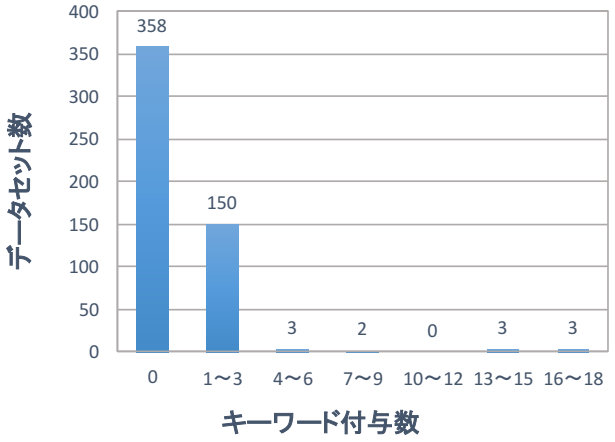

図 1　DIAS データの GCMD サイエンスキーワード付与状況

現したデータセット概要文の情報，キーワード定義文の情報そして既存メタデータ集合の情報を用いることを本研究では想定している．

石田は，地球科学データに対するキーワード推薦のアプローチとして大きく二つの手法を議論した [4]．1 つ目は，推薦対象データセット概要文とキーワード定義文を用いる手法である．この手法は，キーワードごとに付随しているキーワード定義文と推薦対象データセット概要文の類似度を算出し，推薦対象データセット概要文との類似度が高いキーワードを推薦するものである．2 つ目は，推薦対象データセット概要文と既存メタデータ集合を用いる手法である．この手法は，既存メタデータ集合のデータセット概要文と推薦対象データセット概要文との類似度を算出し，既存メタデータのうち，類似度の高いメタデータに存在しているキーワードを推薦するものである．1 つ目の手法は，推薦対象データセットと関連性の高いキーワードを直接推薦する手法であり，2 つ目の手法は，推薦対象データセットと関連性の高い既存メタデータ中に存在するキーワードを推薦対象データセットと関連性の高いキーワードとして間接的に推薦する手法となっているため，前者を直接手法，後者を間接手法と呼ぶこととする．

間接手法は既存メタデータ集合を用いるため，キーワード推薦の精度はデータベースに存在している地球科学メタデータの数や質に依存してしまうが，直接手法は既存メタデータ集合を用いないため，キーワード推薦の精度はデータベースに存在している地球科学メタデータの数や質に依存しない．したがって，既存メタデータ集合の質が低い際には，間接手法よりも直接手法が有効である [4]．我々は，既存メタデータ集合の質が低い状況は，データ管理を始めたばかりのシステムが必ず直面する状況であり，このような状況に対処することは非常に重要であると考え，本研究では直接手法についていくつかの推薦手法を検討した．

GCMD サイエンスキーワードは階層構造となっている．例えば，DROPLET GROWTH というキーワードは，ATMOSPHERE > CLOUDS > CLOUD MICROPHYSICS > DROPLET GROWTH というパスで表現される．我々は，キーワード推薦において，階層を考慮することが重要であると考え，階層を考慮した推薦手法を提案する．

本論文の構成は以下の通りである．まず 2 節において，キーワードに付随するテキスト情報を利用した直接的なキーワード推薦，キーワードの階層構造を考慮したキーワード推薦という二つの観点から関連研究を紹介する．3 節では，3.1 節においてキーワード定義文を用いたナイーブ手法の説明，3.2 節においてその手法の問題点を述べる．4 節では，3.2 節で述べた問題点を緩和するような提案手法について説明を行う．5 節では，3 節において述べたナイーブ手法と 4 節で述べた提案手法の精度を比較する実験の説明を行い，実験の結果及びその結果に対する考察を述べる．最後に 6 節では，本研究のまとめと今後の研究の課題を述べる．

2　関 連 研 究

キーワードやタグは様々な場面で利用されており，現在キーワード推薦に関する研究は，多い [4, 8, 9, 12]．我々はその中でも付与するキーワードが階層構造となっている地球科学データに対するキーワード推薦を対象とした研究を行っており，キーワードの階層構造を考慮したキーワード推薦の手法を検討している．さらに本研究では，キーワード定義文というキーワードに付随する情報を用いた直接手法について，キーワード推薦の手法を検討している．ここではキーワードまたはそれに類するものが階層構造になっているキーワード推薦手法について，そしてキーワード側に付随したテキスト情報を用いたキーワード推薦手法についての二つの観点で関連研究を整理した．

2.1　キーワードの階層構造を考慮したキーワード推薦手法

まずは，キーワードの階層構造を考慮したキーワード推薦手法に関する関連研究を紹介する．文献 [12] は統制語彙からキーワードを選択する状況におけるキーワード推薦手法でなく，任意の語をデータ作成者が付与する状況におけるキーワード推薦手法に関する研究である．文献 [12] では，Q & A コミュニティに存在する深さ 1 ～ 3 の階層となっている 389 個のカテゴリは，データの分類に不十分であることを指摘した上で，類似度の高い Q & A コミュニティのカテゴリをタグとして推薦するだけでなく，タイトルや概要文に含まれる語から主題タグ，キーワードタグを推薦する手法を述べている．この手法はカテゴリや概要文，タイトルが既に付与されているデータセットからナイーブベイズ式を用いて評価スコアを算出し，スコアの高いカテゴリタグ，主題タグを推薦し，更に入力文書に含まれる語の中で TF-IDF 値の高い語をキーワードタグとして推薦する手法となっている．

文献 [1] および文献 [8] は階層構造となっている統制語彙の中からキーワードを選択して付与する手法である．文献 [8] では，ACM Classification Scheme に存在する深さ 1～4 の階層クラスを用いたクラス分類の手法について述べられている．ACM digital library に存在する論文のタイトル，概要文，キーワード，クラスの情報を用いて，分類器を作成する．分類器作成は

階層ごとに行い，ある階層においてクラス分類を行った後，分類されたクラスの子のクラスに関する分類器を作成することを繰り返し，最終的にどのクラスに属しているかを選択する手法となっている． 文献 [1] では，特徴ベクトルとそれに対するラベルベクトルの組について学習を行い，作成した予測器を用いてテストデータの特徴ベクトルを入力した際のラベルベクトルを求め，求めたラベルベクトルのそれぞれの値を木構造となっているラベルのノードの重みとし，更に木構造となっているノードに関する制約を定義し，最適化問題を解くことで階層を考慮したラベル推薦を行っている．文献 [1]，文献 [8] の研究は，推薦対象データセットだけでなく，キーワードが既に付与されているデータセットを利用してキーワード推薦を行う手法である．

2.2 キーワードに付随するテキスト情報を利用した直接的なキーワード推薦手法

キーワードもしくはキーワードに類するものに付随するテキスト情報を用いるキーワード推薦手法に関する関連研究を紹介する．この手法は，推薦対象データセットとキーワードの類似度から関連性の高いキーワードを直接推薦するような手法である．文献 [11] では，名古屋工業大学の学生の目標記述に沿った科目推薦機能について述べられている．この研究で対象となっている名古屋工業大学では学生は入学初年度に C プランと呼ばれる目標記述を作成する．この C プランと科目のシラバスについて TF–IDF や sentence2vec を用いて類似度算出を行い，類似度の高い科目を推薦する実験を行っている．文献 [6] では，Khan Academy というオンラインのコース資料は，前提となる知識が必要なこともあるためにコース資料単体では不十分な場合が存在することを指摘し，オープンソースの教科書のチャプターを Khan Academy のコース資料に割り当てる手法を述べている．この研究では，主に三つの手法が述べられている．一つ目は Khan Academy のメインページに存在する短い概要文やタイトル情報，教科書のチャプターのテキスト情報を TF–IDF でベクトル化し，最も類似度の高いチャプターを割り当てる手法である．二つ目は，Khan Acaedmy のメインページに存在する短い概要文やタイトル情報からルールベースで学習目標となる主要語句を抜き出し，それぞれの語句について最も TF 値の高いチャプターを割り当てる手法である．三つ目，二つ目の手法を改良したもので，教科書のチャプターについて，TF 値を順番に求め，TF 値が閾値を超えた際にそのチャプターを主要語句に割り当てる手法である．文献 [10] では，大学学部計算機科学プログラムで扱うべき知識をまとめた指導要領である CS2013 を専門分野知識体系として用い，実際の大学の科目や MOOC (Massive Open Online Courses，大規模公開オンライン講座) に存在する科目に CS2013 のタグを推薦する手法が提案されている．CS2013 は，18 個の Knowledge Area (以下 KA) で構成されており，更に各 KA が複数の Knowledge Unit (以下 KU) を含む．更に各 KU は複数の Topic を含んでいる．この研究で提案されている手法の一つである Wikipedia 構造手法は CS2013 の KU に存在する複数の Topic をその KU の説明文として利用し，説明文及び科目シラバスの出現語と同じ Wikipedia 記事を結びつけグラフを作成し，そのグラフから各 KU について科目シラバスとの関連度を計算し，科目シラバスと関連度の高い KU を推薦する手法となっている．

文献 [2] では，Khan Academy というオンラインコース資料に CCGPS (Common Core Georgia Performance Standard) という K–12 に基づく教育のカリキュラムを自動ラベル化する手法について述べている．CCGPS は，Grade → Subject → Course → Topic → Instruction というように階層構造となっており，Instruction には短い概要文が付与されている．この概要文から POS tagger や Stanford tagger を用いて特徴語句を抽出し，それらの特徴語句をもとに TF–IDF で特徴ベクトルを作成する．同様にオンラインコース資料に存在する短い概要文やタイトル情報，教科書のチャプターのテキスト情報も TF–IDF で特徴ベクトルを作成する．また，CCGPS において，親ノードは子ノードの特徴語句を含むと考え，Grade，Subject，Course，Topic についても TF–IDF を用いて特徴ベクトルを作成する．あるカリキュラムについて，パス上の，あるノードにおける特徴ベクトルとオンラインコース資料の特徴ベクトルの類似度が子ノードの特徴ベクトルとオンラインコース資料の特徴ベクトルの類似度の 1/10 以下である場合，そのカリキュラムを推薦候補から外し，残ったカリキュラムについて，Instruction の特徴ベクトルとオンラインコース資料の特徴ベクトルを計算し，類似度の高いカリキュラムを推薦する．この研究は，階層も考慮しており，なおかつキーワード側のテキスト情報を用いて直接的にキーワードを推薦する手法であるため，我々の研究と類似しているが，CCGPS ではテキスト情報はリーフノードである Instraction にしか付随しておらず，そのテキスト情報を親ノードにも利用することでリーフノード以外のノードのテキスト情報を補っている．本研究で用いる GCMD サイエンスキーワードでは，全てのノードにキーワード定義文が付随しているため，CCGPS のカリキュラムと GCMD サイエンスキーワードは階層キーワードの構造が異なっている．

このように，キーワードの階層構造を考慮したキーワード推薦手法，キーワードもしくはキーワードに類するものに付随するテキスト情報を用いるキーワード推薦手法に関する研究は様々に存在するが，この両方を考慮したキーワード推薦の手法，すなわちキーワードもしくはキーワードに類するものに付随するテキスト情報を用いる階層キーワード推薦手法に関する研究はほとんどなされていない．我々が本研究で用いる GCMD サイエンスキーワードはキーワード定義文というキーワードに関するテキスト情報が付随しておりかつキーワードが階層構造となっている状況であるため，これらの情報を用いた階層キーワードの推薦手法を新たな手法として模索していく．

3 キーワード定義文を用いたナイーブな推薦手法

3.1 ナイーブな推薦手法

1.2 節で述べたように，直接手法はキーワードの定義文と推薦対象データセット概要文の類似度を計算し，類似度の高い

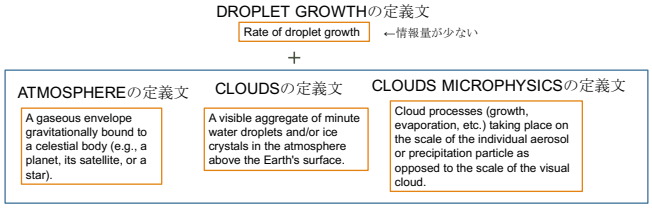

図 2　DROPLET GROWTH の定義文

キーワードを推薦する手法である．したがって，キーワードの定義文や推薦対象データセット概要文は，ある程度の情報量が必要であると考えられる．GCMD サイエンスキーワードでは，キーワード推薦を行う上で，キーワード定義文に含まれる情報量は十分でない場合があるために，精度の低い結果となってしまうことが予備実験により分かっている．予備実験での結果を踏まえ，あるキーワードの定義文は，キーワード自体の定義文だけでなく，パスの上位に存在するキーワードの定義文も考慮したものとすることが適切であると考えた．具体的に述べると，図 2 に示すように，DROPLET GROWTH の定義文は DROPLET GROWTH の定義文だけでなく，パス上位に存在する ATMOSPHERE，CLOUDS，CLOUDS MICROPHYSICS の定義文の情報も含めたものとして考えるということである．

現在は，あるキーワードの定義文は，そのキーワードの上位に存在するキーワードの定義文も含むと考え，それを拡張定義文と呼ぶ．拡張定義文を用いる手法をここではキーワード定義文を用いた推薦におけるナイーブ手法と考える．

この手法の手順は以下の通りである．

（1） キーワードを一つ選択する．

（2） 選択したキーワードの拡張定義文の特徴ベクトル，推薦対象データセット概要文の特徴ベクトルを作成する．

（3） 作成した 2 つの特徴ベクトルの類似度を計算し，その値をそのキーワードのスコアとする．

（4） 全てのキーワードについて，1〜3 を行い，スコアを算出する．スコアの大きい順にキーワードをランキング形式で並べ替えた後，上位のキーワードを推薦する．

この手法におけるキーワードの拡張定義文の特徴ベクトルの作成には，TF–IDF を用い，TF 式は，文献 [7] で述べられているもののうち LRTF 式を用いている．LRTF は，5 語以上の単語を含んでいる推薦対象データセット概要文をクエリとして扱っている状況において，非常に有効である．IDF 式は一般的なものを用いている．キーワードの拡張定義文の特徴ベクトルと推薦対象データセット概要文の特徴ベクトルの類似度算出には，コサイン類似度が用いられる．

3.2　ナイーブな推薦手法の問題点

キーワード定義文を用いたナイーブな推薦手法において，キーワードのパス上位に存在するキーワードの定義文が推薦対象データセット概要文と非常に類似している際に，同じ上位パスを持つキーワードばかりを推薦してしまう場合がある．図 3 は，ある地球科学データについて，ナイーブな手法を用いてキーワード推薦を行った結果であるが，ATOMSPHERE や ATOMOSPERIC RADIATION の定義文と，推薦対象データセット概要文との類似度が高くなっているために，ATOMOSPHERE > ATOMOSPHERIC RADIATION というパスを持つキーワードばかりが推薦されてしまっている．このような状況では，キーワード自身の定義文と推薦対象データセット概要文との類似度が低い，すなわち推薦対象データセットとの関連性が薄いキーワードが推薦されてしまい，適切なキーワード推薦を行うことができない可能性が高くなってしまう．このような問題点を改善するために，我々はいくつかの手法を検討した．次節ではそれらの手法について，詳しく説明する．

- Atmosphere > Atmospheric Radiation > Absorption
- Atmosphere > Atmospheric Radiation > Albedo
- Atmosphere > Atmospheric Radiation > Heat Flux
- Atmosphere > Atmospheric Radiation > Sunshine
- Atmosphere > Atmospheric Radiation > Atomospheric Emitted Radiation
- Atmosphere > Atmospheric Radiation > Anisotropy
- Atmosphere > Atmospheric Radiation > Solar Radiation
- Atmosphere > Atmospheric Radiation > Net Radiation
- Atmosphere > Atmospheric Radiation
- Atmosphere > Atmospheric Radiation > Reflectance

図 3　同じ上位パスを持つキーワードが推薦される事例

4　階層を考慮したキーワード推薦手法

3.2 節で議論した問題点を踏まえ，本論文ではナイーブ手法を改善する，階層を考慮した手法として KR (Keyword Removal) 手法および WT (Weight Tuning) 手法を考案した．さらに，これらの手法の組合せに関しても検討した．

4.1　KR 手 法

KR (Keyword Removal) 手法は，キーワード自体の定義文と推薦対象データセット概要文の関連性が低いと思われるキーワードを推薦キーワードの候補から排除した上で，ナイーブ手法を適用するような提案手法である．具体的な手法については次の 2 つを考えた．

4.1.1　KR–S 手法

3.2 節で述べた問題が起こっている場合では，キーワード自体の定義文と推薦対象データセット概要文の類似度が極端に低いキーワードも推薦結果に現れる可能性が考えられる．キーワード自体の定義文と推薦対象データセットの類似度が低いということは，そのキーワードと推薦対象データセットの関連性はほとんどないと考えられるので，このようなキーワードが推薦結果に現れることは好ましくないと考えられる．KR–S (Keyword Removal based on Similarity) 手法は，推薦対象データセット

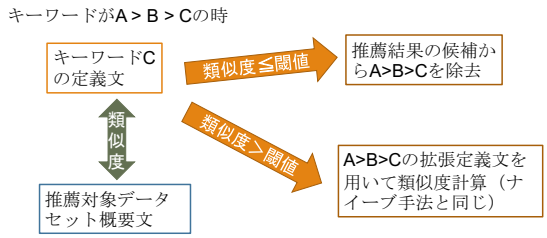

図 4　KR–S 手法

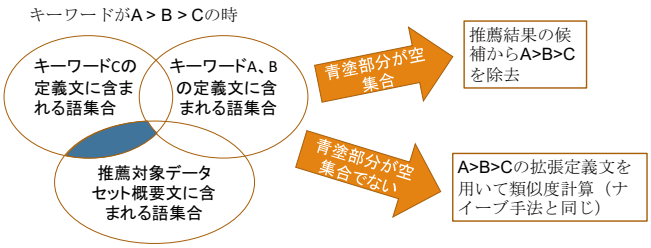

図 5　KR–U 手法

との関連性がほとんどないキーワード，すなわちキーワード自体の定義文と推薦対象データセット概要文の類似度がある閾値以下のキーワードを事前に推薦キーワードの候補から除去し，その後残ったキーワードを推薦キーワードの候補として，3.1 節で述べたナイーブ手法を適用する手法である．図 4 はキーワード A ＞ B ＞ C を例に，この手法を図で表したものである．

4.1.2　KR–U 手法

キーワード自体の定義文と推薦対象データセット概要文の類似度が 0，すなわちマッチしている語が存在しない場合，そのキーワードと推薦対象データセットの関連性は全くないと考えられるので推薦キーワードの候補から除去するのが適切であると考えた．更にキーワード自体の定義文と推薦対象データセット概要文が何らかの語でマッチしている場合でも，その語が様々な地球科学分野において使われるような抽象的な語である場合もそのキーワードと推薦対象データセットの関連性は薄いと考えられるので推薦キーワードの候補から除去する必要があると考えた．我々はあるキーワードについて，キーワードパス上位に存在するキーワードはより抽象的な分野を表しており，キーワードパス上位に存在する語は抽象的な語が使われていると考えた．KR–U (Keyword Removal based on term Uniqueness) 手法はキーワードのパス上位の定義文には含まれないかつ，キーワードの定義文と推薦対象データセット概要文どちらにも含まれている語が存在しない場合に，そのキーワードを推薦キーワードの候補から除去し，残ったキーワードについて，3.1 節で述べたナイーブ手法を適用する手法であると説明できる．KR–U 手法は，推薦対象データセット概要文とキーワードが，キーワードが表す分野特有の語でマッチしている場合のみ推薦キーワードの候補として残す手法であるとも言える．図 5 はキーワード A ＞ B ＞ C を例に，この手法を図で表したものである．

4.2　WT 手 法

あるキーワードにおいて，パスの上位になるにつれて，キーワードの重要度は減少していくと思われる．したがって，あるキーワードにおいて，キーワード自身から遠ざかるにつれ，キーワードの重みを減らしていく必要があると考えられる．また，兄弟の数が多いキーワードの方が兄弟の少ないキーワードよりも細分化されており，似通ったものが多いと考えられるため，兄弟キーワードが共有している上位パスの定義文の重みを減らすことで，キーワード自体の定義文を重要視する必要があると考えた．更にキーワードの兄弟の数が多ければ多いほど，3.2 節で述べた問題が起こりやすくなると考えられるため，兄弟キーワードが多いキーワードについては，共有している上位パスの定義文の重みを減らし，この問題を緩和する必要があると考えた．以上の理由から兄弟キーワードが多い場合は，あるキーワードについて，重みを兄弟の数が多いければ多いほどその親となるキーワードの重みを減らすべきであると考えた．WT (Weight Tuning) 手法は，兄弟キーワードが多いほど親のキーワードの重みが小さくなるように，キーワードパス上位になればなるほどキーワードの重みが減少するように，親キーワードに兄弟キーワードに応じた重みを伝搬していくような手法となっている．3.1 節同様，WT 手法におけるキーワードの拡張定義文の特徴ベクトルの作成には，TF–IDF を用いる．IDF 式は一般的なものを用いている．TF 値計算は 3.1 節の LRTF を少し改良したものとなっている．この手法における TF 値計算はアルゴリズム 1 の通りである．アルゴリズム 1 のそれぞれの記号の説明は，表 1 の通りである．

表 1　アルゴリズム 1 の記号説明

変数名	変数の説明
$Word$	TF 値を求める際に対象となる語
$keyDefWordList$	キーワードのパス上の キーワードの定義文に含まれる語を 階層ごとに格納した二重配列
$keyDefAveLength$	全てのキーワードの拡張定義文に 含まれる語数の平均
$TfValue$	求める TF 値
$currentWordValue$	キーワードの拡張定義文における 語 $Word$ の重みの総和
$currentKeyWeight$	現在のキーワード階層における重み
$keyListWordLength$	キーワードの拡張定義文の語の数
$currnetSibWeight$	現在のキーワードの重みに 掛け合わせる重み
$sibling(t)$	キーワード t の兄弟キーワードの数

4.3　KR–S & WT 手法

4.1.1 節で述べた KR–S 手法と 4.2 節で述べた WT 手法を組み合わせた手法である．ある閾値を設定し，キーワードの定義文と推薦対象データセット概要文の類似度が閾値以下だった場合に，キーワード推薦の候補からそのキーワードを除去し，この操作の後に候補として残っているキーワードについて，WT 手法を適用する手法である．

Algorithm 1 calculateTfValue

Input: $Word, keyDefWordList, keyDefAveLength$

Output: $TfValue$

$currentWordValue \leftarrow 0$
$currentKeyWeight \leftarrow 1$
$keyListWordLength \leftarrow 0$
for $i \leftarrow keyDefWordList.length$ to 0 **do**
  for $j \leftarrow 0$ to $keyDefWordList.[i].length$ **do**
    if $keyDefWordList.[i][j] = Word$ **then**
      $currentWordValue \leftarrow currentWordValue + currentKeyWeight$
    end if
  end for
  $currnetSibWeight \leftarrow sibling(keyDefWordList[i])$
  for $j \leftarrow 0$ to $keyDefWordList.length - i - 1$ **do**
    $currentSibWeight \leftarrow \log_2(currentSibWeight + 1)$
  end for
  $currnetKeyWeight \leftarrow \frac{currentKeyWeight}{currnetSibWeight}$
end for
for $i \leftarrow 0$ to $keyDefWordList.length$ **do**
  $keyListWordLength \leftarrow keyListWordLength + keyDefWordList[i].length$
end for
$TfValue \leftarrow \frac{currentWordValue}{keyListWordLength} \cdot \log_2\left(1 + \frac{keyDefAveLength)}{keyListWordLength}\right)$

4.4 KR–U & WT 手法

4.1.2 節で述べた KR–U 手法と 4.2 節で述べた WT 手法を組み合わせた手法である．あるキーワードについてキーワードのパス上位の定義文には含まれないかつ，キーワードの定義文と推薦対象データセット概要文どちらにも含まれている語が存在しない場合にキーワード推薦の候補からそのキーワードを除去し，この操作の後に候補として残っているキーワードについて，WT 手法を適用する手法である．

5 評価実験

5.1 実験概要

実験に用いる推薦対象データセットのメタデータとしては，Global Change Master Directory (GCMD)[2]に存在する地球科学メタデータのうち，キーワードが 10 個以上ついているものを利用し，更に地球科学メタデータの概要文の語数が 1〜50 語であるもの，51〜100 語であるもの，101〜150 語であるもの，151〜200 語であるもの，201 語以上であるものの 5 つの集合に分けて観察した．5 つの地球科学メタデータ集合それぞれについてメタデータの数は表 2 に示すとおりである．改善手法として提案した KR-S 手法，KR-U 手法，WT 手法，KR-S & WT 手法，KR-U & WT 手法のとナイーブ手法に対する評価結果の比較を行い，ナイーブ手法に比べて改善手法より推薦精度の向上が見られた事例がどれくらいあるかを観察した．また，それぞれの改善手法において 5 つの地球科学メタデータ集合の評価結果の比較を行い，概要文の語数の違いで推薦結果にどのような影響があるかを観察した．ある地球科学データのキーワード推薦結果に対する評価を行う際は，地球科学データに既に付与されているキーワードを正例として，文献 [4] で用いられている 4 つの評価指標のうち，$nhCG_{DSreduce}@n$ を用い，評価スコアを算出した．

KR–S 手法や KR–S & WT 手法で用いる閾値に関しては予備実験として 0 から 0.1 まで 0.01 刻みで 11 個の候補について，4690 個ある推薦対象データセットのメタデータそれぞれの評価スコアの平均を求め，予備実験の結果，最も評価スコアの平均が高かった閾値が 0.09 だったため，今回の実験の KR–S 手法や KR–S & WT 手法における閾値には 0.09 を用いた．

推薦対象データセットのメタデータ集合 5 つについて，それぞれのメタデータについて改善手法の推薦結果における評価スコア，ナイーブ手法の推薦結果における評価スコアを計算し，改善手法の方が評価スコアが高いものの数，ナイーブ手法の方が評価スコアが高いものの数，どちらの手法についても評価スコアが等しいものの数をそれぞれ算出し，それらのメタデータ集合中の割合を図にまとめた．図 6〜図 11 に結果を示す．

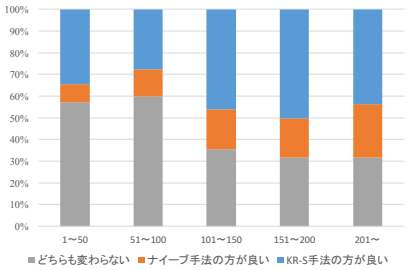

図 6　KR–S 手法とナイーブ手法の評価結果比較

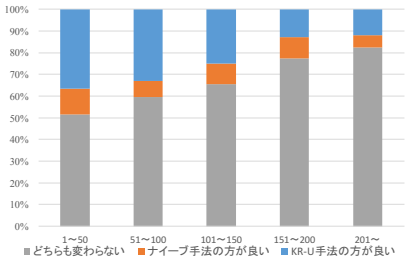

図 7　KR–U 手法とナイーブ手法の評価結果比較

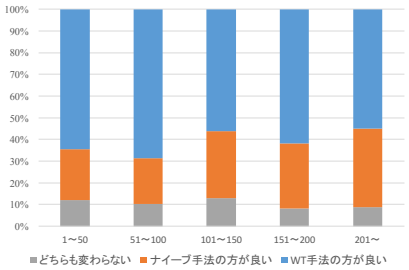

図 8　WT 手法とナイーブ手法の評価結果比較

2 : http://gcmd.nasa.gov/

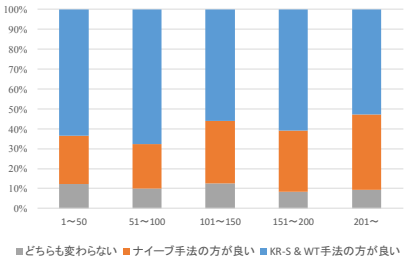

図 9　KR–S & WT 手法とナイーブ手法の評価結果比較

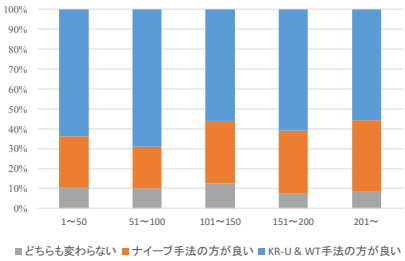

図 10　KR–U & WT 手法とナイーブ手法の評価結果比較

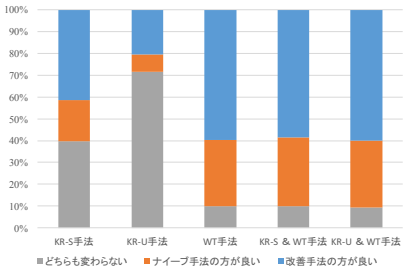

図 11　5 つの改善手法とナイーブ手法の評価結果比較

表 2　5 つに分けた地球科学メタデータ集合それぞれの地球科学メタデータの数

地球科学メタデータ集合 (概要文の語数)	地球科学メタデータ集合のメタデータ数
1～50 語	401
51～100 語	893
101～150 語	807
151～200 語	774
201 語以上	1815
合計	4690

5.2　考　　察

まず，KR–S 手法，KR–U 手法，WT 手法の三つの手法を比較する．WT 手法は改善手法が良くなるものの数も多いが，ナイーブ手法が良くなるものの数も多く，ナイーブ手法とは異なった傾向の推薦ができることが分かる，KR–U 手法は，ナイーブ手法と推薦の傾向があまり変わらないものの，ナイーブ手法が良いものの数を抑えつつ，結果の改善を行うことができることが分かる．次に WT 手法と KR–U & WT 手法，KR–S & WT 手法を比較する．図 11 を見てみると，それぞれの手法について違いがほとんどないことが分かる．これは，WT 手法は，キーワード自身の定義文と推薦対象データセット概要文の類似度を重視している手法であるため，KR–S 手法や KR–U 手法において除去されてしまうような，キーワード自身の定義文と推薦対象データセット概要文の類似度の低いキーワードは WT 手法において推薦キーワードになりにくいからであると考えられる．KR 手法と WT 手法を組み合わせて利用する手法の改善については今後の課題である．

最後に，推薦対象データセット概要文の語数によって，結果がどのように変化しているのかを考察する．図 6 を見ると，KR–S 手法では推薦対象データセット概要文の語数が大きくなるにつれて，KR–S 手法の方が良いメタデータの割合，ナイーブ手法の方が良いメタデータの割合が増加し，どちらも変わらないメタデータの割合が減少している．また図 7 を見ると KR–U 手法において，推薦対象データセット概要文の語数が大きくなるにつれて，改善手法の方が良いメタデータの割合やナイーブ手法が良いメタデータの割合が減少し，どちらも変わらないメタデータの割合が大きくなっている．これは，推薦対象データセット概要文の語数が大きい方が，推薦対象データセット概要文とキーワードの定義文が何らかの語でマッチする可能性が高くなるために，KR–U 手法とナイーブ手法での結果が変わりにくくなるからであると考えられる．文献 [4] では，直接手法は，推薦対象データセット概要文の語数が大きければ大きいほど，キーワード推薦の精度が上がると述べられているが，KR–U 手法では，推薦対象データセット概要文の語数が少ない場合のキーワード推薦の精度を向上することができると考えられる．図 8，図 9，図 10 を見ると，WT 手法，KR–S & WT 手法，KR–U & WT 手法は，推薦対象データセット概要文の語数が大きくなるにつれて，改善手法の方が良いメタデータの割合がわずかに減少し，ナイーブ手法が良いメタデータの割合がわずかに増加している．これは，さきほど述べたようにナイーブ手法は語数が大きくになるにつれてキーワード推薦の精度も向上していくため，その分相対的に改善手法の評価結果が悪くなっているように見えるからだと思われる．

6　おわりに

6.1　ま　と　め

本研究では，階層構造となっている統制語彙を持つ GCMD サイエンスキーワードに注目し，階層を考慮したキーワード推薦を行う手法を検討した．既存メタデータ集合を用いる間接手法はデータベースに存在するメタデータの質に依存するため，質の高いメタデータが豊富に存在しない状況では用いることが難しい．本研究ではこのような状況にも対応できるキーワード定義文を用いる直接手法に焦点を絞り，様々な手法を提案した．まずは，直接手法のおける基礎的な手法となるナイーブ手法を定義し，その手法における問題点を述べた．そして，その問題点を改善できるような手法を考案し，実験を行い，ナイーブ手法と改善手法の比較を行った．また，データセット概要文の語数によって結果がどのように変化するかも実験において観察した．ナイーブ手法と比較して改善手法はキーワード推薦精度が向上する傾向が見られた．

6.2 今後の課題

KR–S & WT 手法や KR–U & WT 手法は，ほとんど WT 手法と結果が変わらず，二つの手法を上手に組み合わせることが出来ていない．これは，WT 手法において，キーワード自体の重みの影響が強く，KR 手法におけるキーワード除去の影響が弱くなってしまうからであると考えられる．KR 手法と WT 手法を組み合わせて利用する際には，WT 手法における重みの算出方法を工夫する必要があると思われる．また，WT 手法において，本論文ではキーワードの階層を考慮した TF 値の計算方法を考えたが，兄弟キーワードの定義文情報なども利用することで，より適切な重みの調節ができる可能性があると思われる．また，今回 KR–S 手法や KR–S & WT 手法で用いた閾値はいくつか試して最も良いものを選んだが，閾値の適切な設定方法も検討していく必要があると考えている．

今回は，キーワード定義文を用いた直接手法について階層構造を考慮した様々な手法を検討したが，キーワード及びデータセット概要文などのメタデータがデータセットに付与されている既存メタデータ集合を用いた間接手法においても，キーワードの階層構造を考慮した手法を検討する余地があると考えている．このような手法については，既存メタデータ集合中のメタデータにおけるキーワードはパス上位のキーワードも付与されていると考え，キーワードを展開する手法があると考えている．この場合，展開後のキーワードに重複が存在する場合にどのように扱っていくか，キーワードを展開すると階層上位のキーワードが付与されているメタデータが非常に多くなってしまい階層上位のキーワードが推薦されやすくなってしまう問題をどのように緩和していくかなど様々な議論が必要であると考えている．

文献 [4] では，直接手法は階層下位のキーワードを推薦する事ができ，間接手法は階層上位のキーワードを推薦する傾向にあると述べられている．このような傾向から直接手法と間接手法を組み合わせる手法は非常に効果的であると考えられるので，階層を考慮した直接手法，間接手法を組み合わせた手法を検討することが今後の研究における最終目標であると考えている．

文　献

[1] Wei Bi and James T Kwok. Multi-label classification on tree-and DAG-structured hierarchies. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 17–24, 2011.

[2] Danish Contractor, Kashyap Popat, Shajith Ikbal, Sumit Negi, Bikram Sengupta, and Mukesh K Mohania. Labeling educational content with academic learning standards. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pp. 136–144. SIAM, 2015.

[3] Global Change Master Directory (GCMD). GCMD Keywords, Version 8.6, 2018. Greenbelt, MD: Global Change Data Center, Science and Exploration Directorate, Goddard Space Flight Center (GSFC) National Aeronautics and Space Administration (NASA). URL (GCMD Keyword Forum Page): https://earthdata.nasa.gov/gcmd-forum.

[4] Youichi Ishida. A keyword recommendation method independent of metadata richness and its application to earth science data. Master's thesis, Kyoto University, 2016.

[5] Akiyuki Kawasaki, Akio Yamamoto, Petra Koudelova, Ralph Acierto, Toshihiro Nemoto, Masaru Kitsuregawa, and Toshio Koike. Data integration and analysis system (DIAS) contributing to climate change analysis and disaster risk reduction. *Data Science Journal*, Vol. 16, , 2017.

[6] Smitha Milli and Marti A Hearst. Augmenting course material with open access textbooks. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 229–234, 2016.

[7] Jiaul H Paik. A novel tf-idf weighting scheme for effective ranking. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 343–352. ACM, 2013.

[8] António Paulo Santos and Fátima Rodrigues. Multi-label hierarchical text classification using the ACM taxonomy. In *14th Portuguese Conference on Artificial Intelligence (EPIA)*, pp. 553–564, 2009.

[9] Suppawong Tuarob, Line C Pouchard, Prasenjit Mitra, and C Lee Giles. A generalized topic modeling approach for automatic document annotation. *International Journal on Digital Libraries*, Vol. 16, No. 2, pp. 111–128, 2015.

[10] 戴憶菱, 浅野泰仁, 吉川正俊. Wikipedia 構造分析による科目シラバスと専門分野知識の関連付け. In *DEIM Forum*, pp. G5–4, 2017.

[11] 宮脇克典, 白松俊, 水野創太, 福本加奈恵, 池田雄斗. 科目区分ダイアグラム検索システムにおけるテキスト類似度に基づく科目推薦機構の試作. 人工知能学会全国大会論文集 2017 年度人工知能学会全国大会 (第 31 回) 論文集, pp. 3N22–3N22. 一般社団法人 人工知能学会, 2017.

[12] 西田京介, 藤村考. 階層的オートタギングによる Q & A コミュニティの知識整理. In *DEIM Forum*, pp. D3–4, 2010.