

地球科学データと学術論文の関連付け

梅原 佳宏[†] 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: [†]y.umebara@db.soc.i.kyoto-u.ac.jp, ^{††}{tshimizu,yoshikawa}@i.kyoto-u.ac.jp

あらまし 観測機器の発達や情報技術の進歩により、膨大な量の地球科学データが収集、蓄積、解析されており、データの統合利用等による新たな知見の発見が期待されている。しかし、地球科学データは専門性が高く、専門分野の合致しない利用者が適切なデータを検索し、データを理解することは困難である場合が多い。我々は地球科学データに対して関連する学術論文を提示することで利用者に対するデータの理解支援を行うことができ、関連付けられた学術論文の情報を用いることで検索の補助を行うことができると考えた。地球科学データのメタデータおよび学術論文の両者から地球科学に関連する語彙を抽出し、データと学術論文の関連付けを行う手法を提案する。

キーワード 地球科学データ, 学術論文

1. はじめに

近年の観測機器の発達や情報技術の進歩により、膨大な量の地球科学データが収集、蓄積、解析されている。それらのデータは、同分野の専門家によって利用される場合もあるが、異なる分野の専門家によって利用される場合もある。異なる分野のデータの統合的な利用によって新たな知見が得られることが期待される。専門分野の合致しないデータ利用者にとっては、データ自体の理解が困難である場合が多いため、メタデータの整備やデータに対するコメントの付与といったデータの理解支援をする必要がある。さらに、データベース内にどのようなデータがあるのか把握することが困難な場合や、検索要求を適切に表現するのが困難な場合があるため、データを探し出す検索補助のための仕組みが必要である。また、データの提供者が、データがどのように利用され、データを利用した人がそのデータに対してどのような感想を持ったのかというフィードバックを求めている場合もある。

一方、科学者はデータを利用して得られた知見をまとめ、学術的な成果として示すために論文を書く。本研究では、地球科学データに対する学術論文に着目し、地球科学データと学術論文の関連付けを行うことで、データ活用支援ができると考えた。地球科学データベースにおいて地球科学データが特定された際に、関連付けられた学術論文を提示することで、利用者に対するデータの理解支援を行うことができ、関連付けられた学術論文の情報を用いることで検索の補助を行うことができると考えられる。また、学術論文の情報により、データ提供者に対してのデータ利用状況に関するフィードバックもできると考えた。

学術論文を利用する利点としては、既に多くの学術論文が書かれているため、データに対してのコメントのみを目的とした新たな内容の記述が必要とされず、また、一般に記述量が多いため、キーワード検索に利用する際に、キーワード抽出がしやすいことが挙げられる。

既存の地球科学データベースとしては、地球流体データの

ためのデータベースである Gfdnavi(Geophysical fluid data navigator)^(注1)[1] や、アメリカ航空宇宙局 NASA が管理する地球科学データのメタデータベース GCMD(Global Change Master Directory)^(注2)や、多種多様な地球科学データを管理している DIAS(Data Integration and Analysis System)^(注3)[2] などがある。また、FAO GeoNetwork^(注4)などの地球科学データのメタデータポータルでは、空間情報を持つデータのメタデータ編集・検索機能を提供するオープンソースのソフトウェア GeoNetwork^(注5)が利用されている。

地球科学データベースではデータセットに対してメタデータを付与することが一般的であり[3]、例えば、GCMD では、DIF(Directory Interchange Format) というフォーマットで多様な地球科学データのメタデータを収集、分類し、検索機能を提供している。また、DIAS では、専用のウェブツールを用いてデータセット単位で地球科学分野のメタデータ標準である ISO19115 [4] および ISO19139 [5] に準拠したメタデータを作成している。我々はこのようなメタデータを介して地球科学データと学術論文の関連付けを行うことにした。なお、地球科学データに対するメタデータのの一つとして参考文献の情報が格納されていることがあり、その場合は地球科学データと学術論文が明示的に関連付けられているとみなすことができる。しかし、現状、このような手動での作業による明示的な関連付けは十分に整備されているとは言い難い。例えば、DIAS では、地球科学データに対するメタデータのの一つとして参考文献があるが、多くのデータセットに関して参考文献情報の記載がない。また、学術論文中に、データセットに対しての謝辞が記述されている場合があり、そこから地球科学データと学術論文の関連付けを抽出することも考えられるが、我々の調査した範囲では、

(注1): <http://www.gfd-dennou.org/arch/davis/gfdnavi/>

(注2): <http://gcmd.nasa.gov/>

(注3): <http://www.editoria.u-tokyo.ac.jp/dias/>

(注4): <http://www.fao.org/geonetwork/>

(注5): <http://geonetwork-opensource.org/>

現状、データセットに対して直接謝辞を記述しているものは少ない。そのため、メタデータ中の参考文献や学術論文中の謝辞を利用するだけでは、関連付けを得るのに不十分であると考えられる。そこで、メタデータと学術論文の関連付けを行う手法が必要であると考えた。

我々は、地球科学データと学術論文との間には、直接的な関係と間接的な関係があると考え、それぞれに基づいた関連付け手法について検討を行った。本研究の手法の概要としては、直接的な関係に基づく関連付けについては、地球科学データのメタデータに記述されているデータセット名、データセット名略称および概要と学術論文の書誌情報および本文から地球科学に関連する語彙を抽出し、地球科学データと学術論文の関連付けを行う。抽出した語彙を用いて、地球科学データのメタデータおよび学術論文のそれぞれでベクトルを作り、コサイン類似度を計算する。そして、コサイン類似度の高い地球科学データのメタデータと学術論文の組を関連するものと考えた。ここで、地球科学に関連する語彙として、Wikipedia の地球科学カテゴリに属するページタイトル一覧を用いて、語彙の抽出を行った。これによりできた関連付けを用いて、地球科学データベースにおいて、データセットが特定された際に、関連する学術論文を提示することができる。

本研究の手法による、地球科学データと学術論文の関連付けの取得、および関連付けの精度の評価を行うために、DIAS [2] 内の地球科学データと DIAS プロジェクトでの成果論文を対象に実験を行った。その際、語彙の抽出に、Wikipedia から抽出した地球科学に関連する語彙集合を利用する場合の他に、地球科学に関連する語彙を考慮せずに全ての名詞を利用する場合についても、実験を行った。実験のそれぞれの結果に共通してコサイン類似度の高かった地球科学データと学術論文の組は、地球科学データと学術論文との間の関連性が高い組であった。地球科学に関連する語彙集合を用いた関連付けにおいてのみコサイン類似度の高かった地球科学データと学術論文の組は、地球科学の複数の分野で利用されるような一般的な語彙によって関連付けがされたと考えられた。また、名詞を用いた関連付けにおいてのみコサイン類似度の高かった地球科学データと学術論文の組は、地球科学の特定の分野で利用される語彙によって関連付けがされたと考えられた。

2. 関連研究

地球科学データに対する理解支援としては、1 章で述べたとおりデータに対するメタデータの付与が一般化してきている。それに伴い、メタデータを利用した地球科学データの検索に関する研究が行われている。本研究の応用として、地球科学データに対する検索について考えている。そこで、まず、現在実際に運用されている地球科学データベースにおける検索システム

図 1 DIAS の検索フォーム

を示す。

DIAS での検索フォーム [3] を図 1 に示す。DIAS では、データセットに対しメタデータを作成し、それを利用した俯瞰・検索システムを開発している。基本的な検索機能としては、キーワード検索、空間検索、時間検索の機能がある。キーワード検索機能では、メタデータに記述されたデータセットの概要やドキュメント作成者等の情報から文字列マッチによる検索を行うことが可能である。また、キーワード入力フォームはいくつかに分かれており、キーワードをデータセット名のみでマッチさせる指定なども可能である。空間検索では、地図インタフェースを用いて範囲を指定して検索を行うことが可能であり、時間検索では、開始日と終了日を指定することが可能である。また、利用分野や作成手段といったデータセット分類軸が整備されており、二次元の表上にデータセットの分類を行っている。これらの機能により、DIAS に登録されているデータセット全体を俯瞰し、検索することが可能になっている。

アメリカ航空宇宙局 NASA が管理する、衛星観測や現場観測によって得られる地球科学データのメタデータベース GCMO2 での検索フォームを図 2 に示す。登録されているデータは農業、大気、海洋などの 13 の分野に分かれており、分野ごとにさらに細かいカテゴリに分類され、階層型の検索インタフェースを提供している。さらに、メタデータとして各データに説明文が付与されているため、それを利用したキーワード検索も可能である。また、時空間情報が付与されているデータならば、地図上の範囲指定による空間検索と期間の入力による時間検索も可能である。

地球科学データのデータ見本集である、Dagik^(注6) [6] では、登録されているデータセットを kml 形式で記録することにより、Google Earth 上に概要を表示できる。また、データ検索ではデータセット名と時間範囲を指定可能である。

最後に、地球流体データのためのデータベースである Gfd-navi [1] では、データベースに登録したデータの検索、解析、共有、メタデータの付与が可能であり、解析結果や可視化された図をデータに付与することもできる。検索システムとしては、カテゴリ検索とファセット検索、及び時空間検索が可能である。

以上の現在実際に運用されている地球科学データベースにお

(注 6): <http://www.dagik.net/>



図 2 GCMD の検索フォーム

ける検索システムに対して、さらに検索を容易にする手法について、いくつかの関連研究がある。衛星画像の検索に関するシステムを提案している岡本らの研究 [7] や、データ自体の重要度および問合せとの時空間近接度に基づく検索を提案している富田らの研究 [8] などがある。

データに対して関連する情報をひも付けて検索に利用する研究に立床らのものがある [9]。現在実際に運用されている地球科学データ検索システムは、データタイプ、時間、空間のいずれか一つ、あるいはそれらの組み合わせを指定することでデータセットを特定する。このうち、時間、空間の指定については専門領域以外の分野においても比較的容易に指定可能であるが、データタイプの指定に関しては専門領域以外の分野のデータセット名だけでは自分の求めているデータかどうか判断するのは困難であると考え、専門知識を地球科学のドメインオントロジーで補うことで、一般的な用語から関連するデータセットを推薦する手法を提案している。この研究では、ドメインオントロジーとして、NASA が開発した地球科学オントロジーである SWEET (Semantic Web for Earth and Environmental Terminology) を使用している。この手法では、まず、検索語に対応する Wikipedia の記事を見つける。そして、Wikipedia のリンク構造を用いて関連する Wikipedia の記事集合を取得する。同時に、それらの記事集合と検索語に対応する記事との関連度が得られる。検索語に対応する記事とリンク先の記事のタイトルを用いて、SWEET の語彙との文字列照合を行い、文字列照合に成功した場合は、地球科学関連記事とみなす。SWEET の語彙にはデータセットとの対応付けがなされており、地球科学関連記事から SWEET の語彙を仲介して対応するデータセットを取得できる。

本研究で得られた地球科学データと学術論文の関連付けを検索に応用する際に、学術論文と地球科学のドメインオントロジーの対応付けを行い、複数のデータセットの対応付けを行うなど、立床らの研究が利用できると考えている。

3. 関連付け手法

本研究では、DIAS 内の地球科学データと成果論文を対象に関連付けを検討する。DIAS 内の地球科学データのメタデータとしては、具体的な項目として、以下がある [3]。

- データセット名、バージョン、データセット名略称
- 問合せ先
- メタデータ作成者、作成日
- データセット作成者、作成日
- データセット概要
- 時間情報
- 地理的範囲
- グリッド
- キーワード
- データセットに関するオンライン情報
- データ周辺情報
- データ配布情報
- データセット作成に関する加工過程や履歴の説明
- 統合データの品質に関する問題点
- データセット利用制約条項
- 参考文献

そこで、地球科学データについては、データセット名、データセット名略称および概要がメタデータに記述されていることを前提として、関連付けに利用できる情報とする。また、学術論文については、書誌情報および本文が、関連付けに利用できる情報とする。

我々は、地球科学データと学術論文の関係として、直接的な関係と間接的な関係があると考えた。直接的な関係は、学術論文の中で関連付けの対象となる地球科学データそのものが言及されているものである。間接的な関係は、データセットと学術論文が地球科学に関連する語彙、分野、人・組織等、何らかの要素を介して関連しているものである。例えば、DIAS のデータベース中に“ マルハナバチモデリングデータ ”というデータセットがあり、DIAS の研究成果論文として“ 野付半島におけるセイヨウオオマルハナバチの定着状況と在来マルハナバチ相 ”[10] という論文がある。論文中において、データセット名およびデータセット名略称が直接参照されていないため、直接的な関係とみなすことはできないと考えられるが、“ マルハナバチ ”という語彙によって関連していると考えられる。そのため、この二つの論文は間接的な関係を持つと考えられる。

関連付けのための手法については、上述の直接的な関係と間接的な関係それぞれに対して手法を考察する必要がある。関連付けの利用例としては、検索等の手段によってデータセットが特定された際に、そのデータセットに関連する学術論文を提示することが考えられる。図 3 に地球科学データベースからのデータ検索と本論文で議論する関連付けとの関係を示す。地球科学データベースにおいて、検索等によりデータセットが特定された際に、直接的な関係と間接的な関係に基づく関連付けにより学術論文を併せて提示する。以下に、直接的な関係と間接的な関係に基づく関連付けについての詳細を示す。

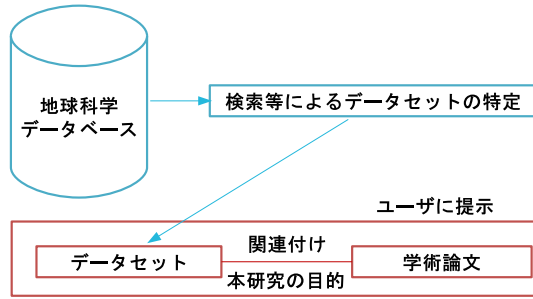


図3 データ検索と関連付け手法との関係

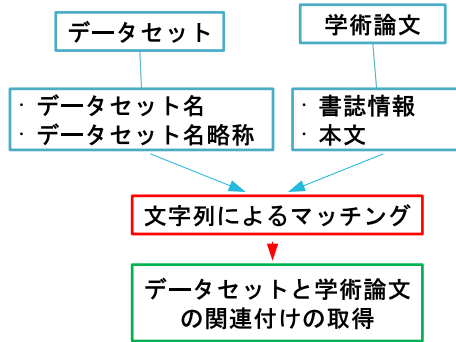


図4 直接的な関係に基づく関連付け

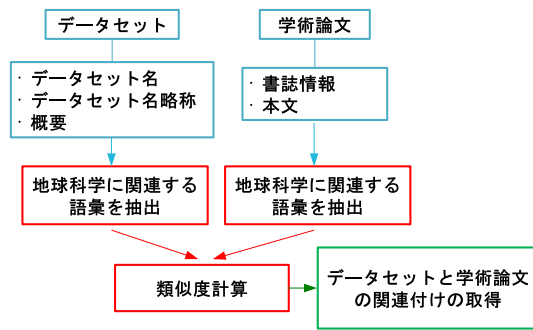


図5 間接的な関係に基づく関連付け

3.1 直接的な関係

図4に、直接的な関係に基づく関連付けの流れを示す。地球科学データのメタデータに記述されているデータセット名およびデータセット名略称と、学術論文の書誌情報および本文を用いて、文字列によるマッチングを行う。学術論文の書誌情報もしくは本文にデータセットのデータセット名もしくはデータセット名略称が出現する場合、それらの間には直接的な関係があるとみなす。

データセット名およびデータセット名略称は、そのデータセットを特定できると考えられるため、文字列のマッチングにより高精度に関連が取得できると考えられる。

3.2 間接的な関係

図5に、間接的な関係に基づく関連付けの流れを示す。地球科学データのメタデータに記述されているデータセット名、データセット名略称および概要と、学術論文の書誌情報および本文から地球科学に関連する語彙を抽出し、類似度を計算し、類似度が高いものに関して間接的な関係があるとし、関連付けを行う。

つまり、具体的なステップとしては、大きく語彙の抽出に関するステップと類似度の算出に関するステップに分けられる。以下に、詳細を示す。

(1) 語彙の抽出

地球科学データのメタデータに記述されているデータセット名、データセット名略称および概要と、学術論文の書誌情報および本文に含まれる地球科学に関連する語を抽出する。その際、事前に語彙集合を定義し、地球科学に関連する語の判定に利用する。地球科学に関連する語彙集合を得る際には、日本語では、リモートセンシング用語集^(注7)、JAXA 衛星用語集^(注8)、地理情報標準専門用語集^(注9)などの利用が考えられる。我々は4章で述べる実験においては Wikipedia を利用して地球科学に関連する語彙を取得した。また、英語では、GCMD サイエンスキーワード、SWEET などの利用が考えられる。

(2) 類似度の算出

語彙の抽出に関するステップにおいて、地球科学データのメタデータと学術論文から抽出した語彙を用いて、tf-idf によって語に重み付けを行い、地球科学データのメタデータと学術論文の特徴ベクトルを作り、コサイン類似度を計算する。

(1) 式のように定義される文書 D_i における単語 t_j の出現頻度 tf_{ij} と、(2) 式のように定義される全文書数中で単語 t_j を含む文書数の逆数の対数 idf_j を求める。

$$tf_{ij} = \frac{\text{文書 } D_i \text{ に含まれる単語 } t_j \text{ の数}}{\text{文書 } D_i \text{ に含まれる全単語数}} \quad (1)$$

$$idf_j = \log \frac{\text{全文書数}}{\text{単語 } t_j \text{ を含む文書数}} \quad (2)$$

(3) 式のように定義される文書 D_i における単語 t_j の重み w_{ij} を求め、(4) 式のような文書 D_i の特徴ベクトルを作る。

$$w_{ij} = tf_{ij} \times idf_j \quad (3)$$

$$D_i = (w_{i1}, w_{i2}, \dots, w_{in}) \quad (4)$$

地球科学データのメタデータと学術論文のそれぞれで特徴ベクトルを作り、式(5)のように定義されるコサイン類似度を計算する。ここで、 M_i は地球科学のメタデータの集合の内、 i 番目のメタデータの特徴ベクトルで、 T_j は学術論文の集合の内、 j 番目の学術論文の特徴ベクトルである。

$$\text{sim}(M_i, T_j) = \frac{(M_i, T_j)}{\sqrt{M_i} \sqrt{T_j}} \quad (5)$$

コサイン類似度の高い地球科学データのメタデータと学術論文の組を、間接的な関係に基づく関連付けとする。

関連付けられた地球科学データと学術論文は、地球科学に関連する語彙を介して関連していると考え、関連付けを提示する際に仲介する語も合わせて提示する。地球科学データのメタデータと学術論文に共通して出現する語の中で、tf-idf 法における重みの値が大きい語を、関連付けを仲介している語と考えた。

(注7): <http://www.science.aster.ersdac.or.jp/jp/glossary/index.html>

(注8): <http://www.satnavi.jaxa.jp/basic/glossary/index.html>

(注9): <http://www.gsi.jp/GIS/stdind/jsi-yogo-keyword.html>

表 1 実験に用いたデータ数

データセット数	学術論文数	利用した Wikipedia の ページタイトル数
119	52	3902

表 2 文字列によるマッチング

データセット名	学術論文タイトル	マッチング した文字列
SSM/I 海洋プロダクト	GSMaP_TMI プロダクト における陸上降水強度の評価	SSM/I

4. 実験

実験では、地球科学データとしては、DIAS のデータセットを対象とした。また、DIAS では研究成果も管理しており、我々は DIAS の成果物である学術論文の一覧を入手して実験に用いた。また、今回の実験では、DIAS のデータセットと成果論文の内、日本語で記述されているデータセットメタデータと日本語で記述されている論文を対象とした。実験に用いたデータセットのメタデータと学術論文の数を表 1 に示す。

語の抽出には、MeCab^(注10)を利用して形態素解析を行った。英語で記述されている GCMD サイエンスキーワードといった語彙体系や、SWEET といったオントロジーは、地球科学分野全体を網羅している一方で、3.2 節で述べた日本語で利用できる辞書は、特定の分野を専門的に扱っている。そこで、地球科学分野全体を網羅する日本語の語彙集合として、Wikipedia の地球科学カテゴリ下に含まれるページタイトルの一覧を利用した。利用した語の総数は、表 1 に示す。現在、Wikipedia の地球科学カテゴリ下に含まれるページタイトルの内、著者が主観的に選んだページのタイトルを地球科学に関連する語彙として利用して実験を行ったが、語彙の選択手法によって関連付けの精度が変わると思われるため、適切な語彙集合の選択に関して考察が必要であると考えている。

4.1 直接的な関係

データセットのメタデータに記述されているデータセット名および略称と学術論文の書誌情報および本文を用いて、文字列マッチングを行った結果を表 2 に示す。本実験で対象とした地球科学データおよび学術論文から、直接的な関係として抽出できたものは 1 例のみであった。

表 2 より、マッチした文字列である“SSM/I”とは、マイクロ波放射計の名称である。“GSMaP_TMI プロダクトにおける陸上降水強度の評価”^[11]という学術論文では、TRMM(熱帯降雨観測衛星)搭載のマイクロ波放射計 TMI を利用して全球の降水マップを作成するプロダクト GSMaP_TMI について、陸上を対象とした評価結果について記述してある。その中で、「また、TMI 用のアルゴリズムは、AMSRE、SSM/I といった他のマイクロ波放射計用としても基本的に受け継がれているため、これを詳細に評価しておくことは、GSMaP 全体にとっても有

用である」というように、マイクロ波放射計のその他の例として SSM/I が挙げられている。そのため、“SSM/I 海洋プロダクト”というデータセットのデータセット名略称“SSM/I”と直接的な関係に基づく関連付けが行われた。

他の例としては、“西スマトラレーダー・雨量計統合降水量データ第 1.1 版”というデータセットのメタデータにおいて、データセット名略称として“ WeSRI-1.1 ”が記述されている。“西スマトラレーダー・雨量計統合降水量データの試作と検証”^[12]という学術論文の本文中で、「インドネシアの西スマトラを対象に、気象レーダーと地上雨量計の観測値を組み合わせ、2006 年 10 月 28 日から 11 月 27 日までの、時間間隔 30 分、空間間隔緯度経度 0.025 度 (約 2.8km) の格子型降水量データ (第 1 版、WeSRI-1) を試作し、ウェブサイトから公開した。本発表では、この WeSRI-1 の作成方法と検証結果について報告する」という記述がある。学術論文では、“西スマトラレーダー・雨量計統合降水量データ第 1.1 版”とはバージョンが異なるデータセットが利用されたことが考えられるが、文字列によるマッチングでは関連付けとして取ってることが出来なかった。ここで、学術論文中にバージョンが異なるデータセットが記述されている場合についても直接的な関係とみなすことを考えた。そこで、このように、バージョンが異なるデータセットが学術論文で記述されている場合でも、関連付けが取得できるような手法が必要であると考えられる。また、地球科学データのメタデータと学術論文との間で表記ゆれが見られる場合でも文字列によるマッチングでは関連付けが取得できないため、併せて手法を考察する必要があると考えられる。

手法の改良としては、語の近接性の利用が考えられる。語の近接性を考慮した直接的な関係に基づく関連付けの概要を図 6 に示す。地球科学データのメタデータに記述されているデータセット名およびデータセット名略称のそれぞれから地球科学に関連する語彙を抽出し、学術論文の書誌情報および本文中で、抽出した語が出現した位置をそれぞれ取得していく。そして、取得したそれぞれの語の出現した位置についての近接度を算出する。算出した近接度が閾値を上回っている場合、直接的な関係に基づく関連付けとする。算出した近接度が閾値を上回っているということは、もとのデータセット名およびデータセット名略称と類似した文字列が学術論文の書誌情報および本文中に記述されているということであると考えられる。これにより、単純な文字列によるマッチングでは関連付けを行うことができなかった、地球科学データのメタデータと学術論文との間で表記ゆれが見られる場合や、学術論文で記述されている地球科学データとバージョンが異なる地球科学データについても対応できるようになると考えられる。

今後、近接度の算出についての検討を行い、地球科学データのメタデータと学術論文との間で表記ゆれが見られる場合や、学術論文で利用されている地球科学データとバージョンが異なる地球科学データについて、関連付けが行うことができるのか実験を行っていきたいと考えている。

4.2 間接的な関係

語の抽出の際に用いる語彙集合を変えて、実験を行った。一

(注 10): オープンソースの形態素解析エンジン <http://mecab.sourceforge.net/>

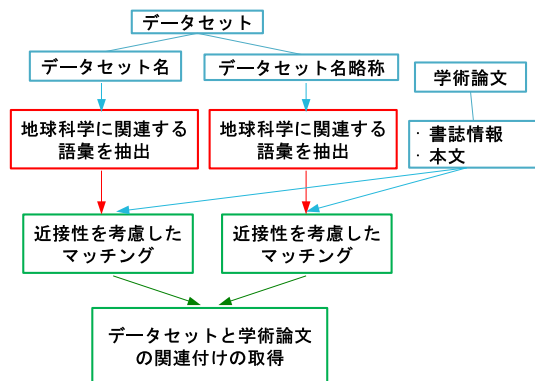


図 6 語の近接性を考慮した直接的な関係に基づく関連付け

つ目の語彙集合として、Wikipedia から抽出した地球科学に関連する語彙集合を MeCab の辞書に追加し、Wikipedia の語彙集合に含まれると判定された語を関連付けに利用する語として用いた。MeCab の辞書は、大量の一般的な語彙によって構成されている。そこで、二つ目の語彙集合として、形態素解析を行い、名詞として判定された語を関連付けに利用する語として用いた。それぞれの語彙集合で地球科学データと学術論文の関連付けが 10 程度になるように、Wikipedia の語彙集合を用いた場合には、コサイン類似度の閾値を 0.6 に設定し、名詞として判定された語を用いた場合には、コサイン類似度の閾値を 0.15 に設定した。

関連付けられた地球科学データと学術論文は、地球科学に関連する語彙を介して関連していると考えたが、地球科学データと学術論文に共通して出現する語の中で、tf-idf 法における重みの値が大きい語を 3 程度提示し、その地球科学に関連する語彙とした。その結果をそれぞれの表の“ 仲介する語 ”の列に示す。それぞれの結果には、データセットのメタデータと学術論文の内容を見て、以下のような 4 段階の基準で関連付けの評価を行った。

- (1) データセットと学術論文の間に高い関連性がある学術論文で、データセットを利用・作成している等の場合
- (2) 語を介して強く関連している学術論文で、データセットを利用・作成していると判断することは難しいが、同様のデータセットを利用・作成しているか、利用・作成しているデータセットの一例と考えられる等の場合
- (3) 語を介して弱く関連している学術論文はあるプロジェクト自体について記述していて、データセットがそのプロジェクトで作成されたデータである等の場合
- (4) データセットと学術論文の間に関連性はほとんどない

それぞれの結果で、共通してコサイン類似度が閾値を上回った地球科学データと学術論文の組を表 3 に示す。共通してコサイン類似度が閾値を上回った組については、関連付けの評価は 1 となっており、地球科学データと学術論文が介して関連している地球科学に関連する語彙についても、それぞれ類似した語彙が提示されている。

表 3 共通してコサイン類似度が閾値を上回った組

データセット名	学術論文タイトル	Wikipedia の語彙集合を用いた場合の仲介する語	名詞として判定された語を用いた場合の仲介する語
1945-47 年撮影の航空写真を基にして作成された地形図より抽出を行った氷河分布域 (サンプル版)	モンゴル氷河インベントリ	氷河・地形図	氷河・モンゴル・記載
西スマトラレーダー・雨量計統合降水量データ第 1.1 版	西スマトラレーダー・雨量計統合降水量データの試作と検証	雨量計・地形・気象レーダー・地上・降水量	レーダー・雨量・スマトラ

表 4 Wikipedia の語彙集合を用いた場合のみ関連付けが取得された組の内、関連付けの評価が高かった組

データセット名	学術論文タイトル	仲介する語	評価
植生指数と気温、降水量、光合成有効放射量との経年変化の関係を示す全球マップ	4 倍大気 CO2 時の植生分布変化と施肥効果が大気と炭素収支に及ぼす影響	気温・降水量	(2)

Wikipedia の語彙集合を用いた場合の関連付けにおいてのみコサイン類似度が閾値を上回った地球科学データと学術論文の組の内、関連付けの評価が高かった組を表 4 に示す。表 4 にある地球科学データと学術論文の組は、“ 気温・降水量 ”という地球科学に関連する語彙の中でも一般的な語彙によって関連付けが行われたと考えられる。一方、名詞として判定された語を用いた関連付けでは、地球科学とは関連しない一般的な名詞も抽出されるため、それらの語彙が影響してコサイン類似度が閾値を上回らないという結果になったと考えられる。そこで、抽出の際に利用する語彙集合を地球科学の分野に限定することによって、地球科学の複数の分野で利用されるような一般的な語彙によって関連している地球科学データと学術論文の関連付けを取得できると考察した。

次に、名詞として判定された語を用いた場合の関連付けにおいてのみコサイン類似度が閾値を上回った地球科学データと学術論文の組の内、関連付けの評価が高かった組を表 5 に示す。名詞として判定された語を用いた場合の関連付けにおいては、地球科学データのメタデータと学術論文に含まれる大量の一般的な名詞が、抽出した語として扱われるため、全ての地球科学データと学術論文の組にある程度のコサイン類似度が出た。表 5 にある地球科学データと学術論文の組は、“ TRMM ”という Wikipedia の語彙集合に含まれていない地球科学に関連する語彙によって関連付けが行われたと考えられる。“ TRMM ”という語彙は MeCab の辞書にも含まれていないため、形態素解析において未知語として扱われるが、未知語は名詞に含まれる語として判定されるため、“ TRMM ”という地球科学に関連する

表 5 名詞として判定された語を用いた場合のみ関連付けが取得された組の内、関連付けの評価が高かった組

データセット名	学術論文タイトル	仲介する語	評価
CEOP 衛星データセット (TRMM > TMI > 3 PRECI)	GSMaP.TMI プロダクトにおける陸上降水強度の評価	降雨・TRMM・強度	(1)

表 6 名詞として判定された語を用いた場合のみ関連付けが取得された組の内、関連付けの評価が低かった組

データセット名	学術論文タイトル	仲介する語	評価
DIAS 衛星データセット	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	DIAS・アーカイブ・ユーザ	(4)

表 7 Wikipedia の語彙集合を用いた場合での関連付けの評価

	閾値:0.5	閾値:0.3	閾値:0.1
評価 (1)	2	2	4
評価 (2)	5	12	20
評価 (3)	11	35	135
評価 (4)	19	117	622

表 8 名詞として判定された語を用いた場合での関連付けの評価

	閾値:0.1	閾値:0.075	閾値:0.05
評価 (1)	5	5	5
評価 (2)	9	15	27
評価 (3)	25	31	94
評価 (4)	33	96	310

語彙が関連付けに利用することができた。そこで、一般的な名詞によって全ての地球科学データと学術論文の組にある程度のコサイン類似度が出る中で、地球科学の特定の分野で利用される語彙によって関連している地球科学データと学術論文の関連付けを取得できると考察した。

一方で、名詞として判定された語を用いた場合の関連付けにおいてのみコサイン類似度が閾値を上回った地球科学データと学術論文の組の内、関連付けの評価が低かった組を表 6 に示す。表 6 にある地球科学データと学術論文の組は、“DIAS”という Wikipedia の語彙集合に含まれていない語彙によって関連付けが行われたと考えられる。この語彙は、地球科学の分野におけるプロジェクトの名称ではあるが、データセットとの関連は弱いと考えられ、地球科学データと学術論文の関連付けとして取得するべきと考えることは難しい。

最後に、Wikipedia の語彙集合を用いた場合と名詞として判定された語を用いた場合のそれぞれについて、閾値を動かし、関連付けの評価を行った。それぞれの閾値において、関連付けられた地球科学データと学術論文の組で、評価 (1) ~ 評価 (4) となった組の数をまとめたものを表 7 と表 8 に示す。表 7 より、Wikipedia の語彙集合を用いた場合では、評価 (3) と評価された地球科学データと学術論文の組を多く取得できている。特に、閾値を下げていった場合に、評価 (3) の組が名詞を用い

た場合よりも多く取得できている。一方で、評価 (4) の組も多く取得されている。また、表 8 より、名詞として判定された語を用いた場合では、評価 (1) や評価 (2) の組が多く取得できている。

語の抽出に利用する語彙集合を変えて、実験を行ったが、地球科学に関連する語彙集合を用いた場合の関連付けでは、語彙集合を地球科学の分野に限定することによって、地球科学の複数の分野で利用されるような一般的な語彙によって関連している地球科学データと学術論文の関連付けを取得できた。名詞として判定された語を用いた場合の関連付けでは、地球科学の特定の分野で利用される語彙によって関連している地球科学データと学術論文の関連付けを取得できた。

これらの結果から、事前に定義する地球科学に関連する語彙集合について、特定の分野で利用される語彙を全て網羅するには限界があるため、語彙集合に含まれていない地球科学に関連する語彙を関連付けに利用するために、語彙集合の組合せを行う等の工夫が必要であると考えられる。また、評価 (4) と評価された地球科学データと学術論文の組を取得しないように判別する工夫も検討していきたい。

今回の実験では、語の抽出には、地球科学分野全体を網羅する日本語の語彙集合として、Wikipedia の地球科学カテゴリ下に含まれるページタイトルの一覧を利用したが、地球科学に関連する語彙集合の選択についてもさらに考察が必要であると考えている。

5. おわりに

我々は地球科学データに対して関連する学術論文を提示することで利用者に対するデータの活用支援ができると考えた。地球科学データと学術論文の関係として、直接的な関係と間接的な関係があると考え、地球科学データのメタデータおよび学術論文の書誌情報、本文を用い、地球科学に関連する語彙を抽出して関連付けを行った。

今後の課題として、直接的な関係に基づく関連付けでは、文字列によるマッチングでは関連付けができなかった場合に対応するために、地球科学データのメタデータに記述されているデータセット名およびデータセット名略称のそれぞれから地球科学に関連する語彙を抽出し、学術論文中の出現した位置の近接度によって、直接的な関係に基づく関連付けを行うことを考えている。今回の研究の対象に含めなかったが、学術論文中にデータセットの URL が記述されている場合もあり、直接的な関係に基づく関連付けとして利用できると考えられるため、検討していきたいと考えている。

特定の著者が同じ分野において複数の学術論文を書く場合が多くあると考えられるため、学術論文の著者や共著者のその他の学術論文や参考文献に記述されている学術論文の情報などが、間接的な関係に基づく関連付けに利用できると考えている。さらに、同じプロジェクト内の他の地球科学データセットや学術論文の情報なども間接的な関係に基づく関連付けに利用できる

と考えている。

今回の実験では、学術論文に、DIAS の成果論文を対象として関連付けを行った。今後、対象とする学術論文の拡大を考えており、一般の論文データベースからの取得を検討していきたい。

文 献

- [1] Takeshi Horinouchi, Seiya Nishizawa, Chiemi Watanabe, Akinori Tomobayashi, Shigenori Otsuka, Tsuyoshi Koshiro, Yoshiyuki Hayashi, GFD Dennou Club, " Gfdnavi, Web-Based Data and Knowledge Server Software for Geophysical Fluid Sciences, Part I: Rationales, Stand-Alone Features, and Supporting Knowledge Documentation Linked to Data ", Lecture notes in computer science, Vol.6193, pp. 93-104, 2010.
- [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, DE2010-33, pp. 45-50, 2010 年.
- [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, DE2010-32, pp. 39-44, 2010 年.
- [4] International Organization for Standardization: " ISO19115:2003, Geographic information Metadata ".
- [5] International Organization for Standardization: " ISO/TS 19139:2007 Geographic information Metadata XML schema implementation ".
- [6] Akinori Saito, Daiki Yoshida, " Dagik:A Data-Showcase System For The Geospace ", Data Science Journal, Vol.8, 2009 年.
- [7] 岡本章裕, 横山昌平, 福田直樹, 石川博, " Wikipedia を用いた衛星画像検索システムの開発 ", DEIM Forum 2010 F4-4
- [8] 富田典也, 清水敏之, 齊藤昭則, 吉川正俊, " 重要度と時空間近接度を統合した地球科学データのランキング手法 ", WebDB Forum 2011
- [9] Masashi Tatedoko, Toshiyuki Shimizu, Akinori Saito, and Masatoshi Yoshikawa, " A Retrieval Method for Earth Science Data Based on Integrated Use of Wikipedia and Domain Ontology ", DEXA 2010, pp. 277-284, 2010.
- [10] 井上真紀, 菊池玲奈, 石川聖江, 横山潤, 鷺谷いづみ, " 野付半島におけるセイヨウオオマルハナバチの定着状況と在来マルハナバチ相 ", 保全生態学研究, Vol.12, No.2, pp. 172-175, 2007 年.
- [11] 瀬戸心太, 久保田拓志, " GSMaP-TMI プロダクトにおける陸上降水強度の評価 ", 水文・水資源学会 2006 年度総会・研究発表会, 2006 年.
- [12] 上米良秀行, 増田耕一, 服部美紀, 森修一, 濱田純一, 櫻井南海子, 松本淳, 山中大, " 西スマトラレーダー・雨量計統合降水量データの試作と検証 ", 水文・水資源学会 2008 年度総会・研究発表会, 2008 年.