3N-04

分野特徴語を利用した地球科学データに対するキーワード推薦手法

石田陽一† 清水敏之† 吉川正俊†

京都大学大学院情報学研究科†

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で，異分野間でのデータの統合的利用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．海外ではGCMD(Global Change Master Directory)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などが存在する一方で，国内では，地球環境情報統融合プログラム(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．本研究では，この DIAS-P に注目する．DIAS-P では，異分野間でデータを相互運用するためのデータ基盤の構築を目指しており，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-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．しかし，GCMD が提供している科学キーワード集 GCMD Science Keywords[1]などは，2000 語以上もの語彙が階層化されて管理されており，その統制語彙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，本研究では，データ提供者が各メタデータを記述する際に，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た．なお，今回推薦を行うキーワードは，上述した GCMD Science Keywords に収録されている語彙とする．

我々は地球科学分野特徴語を用いてキーワードを推薦する手法を提案しているが[3]，本論文ではその有効性を確認するために，提案手法と TF-IDF を用いた手法を比較して検証を行った．以下 2 節で文献[3]で提案した推薦手法の概要を述べ，3 節で評価実験について述べる．

2. キーワードの推薦手法

本研究では，メタデータ中の概要文の情報と，各キーワードの定義文の情報を利用，解析することで，各データセットに対してキーワードを推薦する手法を提案する．メタデータにおける概要文とは，そのデータセットの意味や大まかな内容を説明したものである．また，キーワード定義文とは，前述のGCMD Science Keywords に収録されたキーワードに付与されている，そのキーワードの意味を説明した文を指す．我々は，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ることを考えた．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，地球科学分野特徴語を定義し，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．今回は地球科学分野特徴語を導出するために，地球科学分野以外の他分野と比較し，地球科学分野における相対的出現頻度が高い単語を利用した．

提案手法の概要としては，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，定義文中やキーワード名中に含まれている抽出単語の数や，その抽出単語に付与される重みの値を利用し，各データセットに対す

Keyword Recommendation Method for Earth Science Data using Domain-Specific Words : Youichi Ishida, Toshiyuki Shimizu, Masatoshi Yoshikawa(Kyoto Univ.)

る各キーワードの適合度を算出する．抽出単語に付与される重みの値としては，各単語の分野間での相対的出現頻度値を利用した．なぜなら，その単語の相対的出現頻度値は，いかに地球科学に特徴的な単語であるかの度合を表すためである．

地球科学分野特徴語リストの作成過程や，キーワード適合度算出式の詳細などは，[3]を参照されたい．

3．評価実験

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や各単語への重み付与の有用性を確かめるため，TF-IDF を用いた手法と比較した．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平均を算出した．

3-1．比較手法

我々は，概要文中に存在する単語を多く含むような定義文を持つキーワードを推薦することを考えていた．

そこで，提案手法と比較するために，比較手法として，各データセットの概要文とキーワードの定義文との類似度を測定する手法を検討した．クエリとなる概要文を特徴ベクトルで表現し，特徴ベクトルの各要素には，その単語が出現すれば 1，出現しなければ 0 を設定した．そして，各キーワード定義文も特徴ベクトル化し，各単語の TF-IDF 値を特徴ベクトルの各要素の値とした．この際，TF 値として，[4]や[5]などで利用されている LRTF(Length Regularized TF)を使用した．LRTF は，TF 値を平均文書長で正規化するような式となっており，例えば 5 単語以上で構成されるような文書長の長いクエリに対して特に有効に働く[4]．そのため，概要文をクエリと考える場合はこの LRTF を利用するのがふさわしいと考えた．また，IDF 値としては，全定義文数 2017 を各単語の DF 値で割った値に対数を取るという，最も標準的な式を利用した．そして，作成した二つの特徴ベクトルのコサイン類似度を測定し，類似度上位 10 件中の適合率平均を算出した．

3-2．実験結果及び考察

提案手法と TF-IDF を用いた手法を比較した結果を表 1 に示す．

表 1 提案手法と比較手法の適合率平均

手法	適合率平均
提案手法	28%
TF-IDF 利用の比較手法	24%

表 1 の結果より，僅かながらではあるが，提案手法により，適合率平均が向上した．

その中でも大きく適合率が向上したデータセットの概要文には，”degree”や”minute”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含まれていた．比較手法では，それらの単語の重みが比較的大きな値を示していたことが，推薦精度の悪化に起因したものと考えられる．一方，提案手法では，地球科学分野特徴語リストを作成したことで，これらの単語を抽出単語から排除できたことが，適合率向上につながったと考えられる．

4．まとめと今後の課題

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では，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た．しかし，ある一部の単語に過度な重みを付与したことで，推薦精度が悪化するデータセットも存在したので，今後は単語の重み付与の仕方について再考したい．また，現在は，分野間での相対的出現頻度を測定する際に２群の差の比率の検定式[6]を利用しているが，自己相互情報量，対数尤度比など，他にも相対的出現頻度を測定する指標は存在する．それらの指標による比較実験を行い，最も適する指標を選択する必要がある．

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，小池 俊雄，"DIAS におけるデータ公開と課題，" 情報知識学会誌，Vol. 24, No. 3, pp. 254-274, 2014 年 10 月.

[3] 石田陽一，清水敏之，吉川正俊：地球科学データに対するキーワード推薦手法，第 5 回 Web インテリジェンスとインタラクション研究会(ARG SIG-WI2)，2014.

[4] PAIK, Jiaul H. 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. ACM, 2013. p. 343-352.

[5]AMATI, Gianni; VAN RIJSBERGEN, Cornelis Joost. Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS), 2002, 20.4: 357-389.

[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.