重要度と時空間近接度に基づいた地球科学データの推薦

冨田　典也† 清水　敏之† 齊藤　昭則‡ 吉川　正俊†

†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
‡京都大学大学院　理学研究科 〒606-8502 京都市左京区北白川追分町

E-mail: †tomita@db.soc.i.kyoto-u.ac.jp, ‡saitoua@kugi.kyoto-u.ac.jp
††{tshimizu,yoshikawa}@i.kyoto-u.ac.jp

あらまし 観測技術の発達や情報技術の進歩により地球科学データは爆発的に増大しており，データの保存や検索のために地球科学データベースが構築されている．既存の地球科学データベースの使い難い点として，一般にデータに対する知識がなければ適切な検索を行うことができないことが挙げられる．そこで本研究では，データ自体の重要度，およびデータの時空間情報と問合せの時空間条件との近接度から算出したスコアが高いデータを優先して提示する検索システムを提案する．本手法により，厳密には検索条件に一致しないが重要度の高いデータを取得することができ，利用者に対して検索条件をきっかけとしたデータの推薦が可能になる．

キーワード 時空間検索，データ推薦

A Recommendation Method for Geoscience Data Based on Importance and Spatiotemporal Proximity

Fumiya TOMITA† Toshiyuki SHIMIZU† Akinori SAITO‡ Masatoshi YOSHIKAWA†

†Graduate School of Informatics, Kyoto University Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan

‡Graduate School of Science, Kyoto University Kitashirakawa-Oiwakecho, Sakyo-ku, Kyoto, 606-8502 Japan

E-mail: †tomita@db.soc.i.kyoto-u.ac.jp, ‡saitoua@kugi.kyoto-u.ac.jp
††{tshimizu,yoshikawa}@i.kyoto-u.ac.jp

1. はじめに

1.1. 研究の背景

近年，記憶容量やネットワーク容量の増大に伴って，地球科学データの観測と収集が盛んに行われるようになり，その結果，大量の地球科学データが蓄積されている。そのような膨大な観測データを管理，分析及び検索するために，地球科学データベースが開発，運用されている．

これらの地球科学データベースは地球科学の専門家の利用を想定しているのが普通だが，教育目的などで一般人が利用できるようにすれば，環境に対する関心を高めるために利用できる．しかし，地球科学データは専門用語を用いて記述されているため，地球科学の専門家以外の人がキーワード検索やディレクトリ検索を利用して有意な情報を入手するのは困難である．

そのため，ほとんどの地球科学データベースでは，データセットに時間と地理的範囲を付加しておき，時間と空間の条件に含まれるデータのみを取得できるようにしている．しかし，そのような時間，空間検索を利用したとしても，取得したいデータに関する知識を持っていなければ，指定した条件に当てはまるデータセットが存在しないか，大量のデータセットが該当する場合のいずれかであることが多く，有意なデータを探し出すのは困難である．

そこで本研究では,時間と空間の検索条件の緩和を行い，さらにデータの特徴量と緩和の度合いに応じて推薦を行うことで，重要度及び時空間検索条件との近接度が高いデータを推薦するデータベース検索を提案する．

1.2. 既存の地球科学データベース

1.2.1. Gfdnavi

Gfdnavi[3]は，地球流体データのためのデータベースサーバ，及び地球流体データベースの解析，可視化を行うデスクトップツールの総称である．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る．また，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．

検索方法としては，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[4]．ファセット検索とは，データをファセットと呼ばれるいくつかの直交したカテゴリに分類し,ファセットが保持している値(キー)とファセットの組み合わせを選択することでデータの

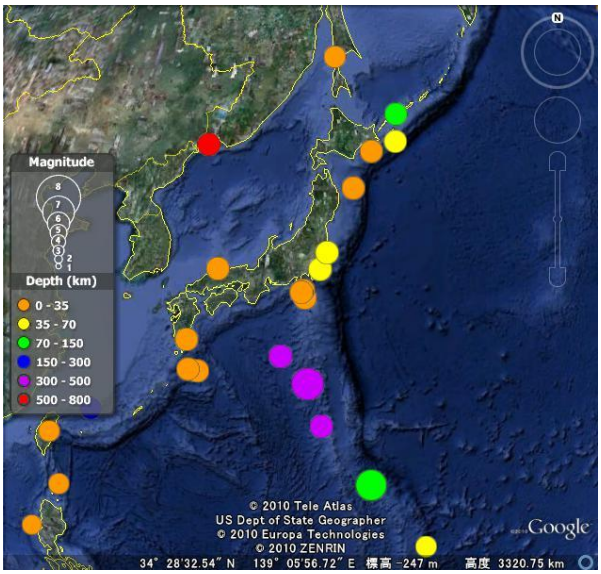

図 1　2000 年の M6 以上の地震の震源地（色は震源の深さを表す）

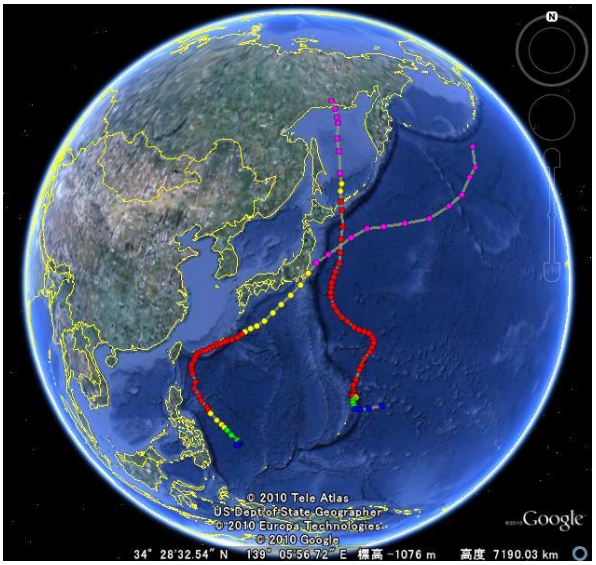

図 2　1992 年の台風の中心軌道例（色は規模の大きさを表す）

絞り込みを行う方法である．なお，時空間検索を行うためには，時空間情報をメタデータとして付与する必要がある．

1.2.2. GCMD

GCMD[5]は，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であり，アメリカ航空宇宙局 NASA が管理している．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ている．さらに，各データに説明文がメタデータとして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，地図の範囲指定による地理検索と日付の範囲指定による時間検索も可能である．ただし，条件として指定した範囲とデータに付与された範囲に重なりがあるようなデータは全て取得するため，大量のデータセットが結果として表示されることが多い．

1.3. データ例と既存の検索の問題点

例として，Dagik [1][2]の災害に関する地球科学データを検索する場合を考える．Dagik では，kml 形式のデータを Google Earth 上に表示する．例えば，図 1 のような日本で起こった地震のデータを取得する場合，日本の国土内に与えた被害が大きくなると考えられる地震を取得したい．

この場合，規模が大きく日本の国土に近い地震ほど被害が大きくなると考えられるが，そのような地震の震源だけを含む範囲を設定するのは難しい．そこで，日本の国土を含む最も小さな長方形を空間範囲として，その空間範囲と震源との距離によるスコアと，マグニチュードに応じたスコアを総合してデータ推薦をする．これによって，空間範囲の緩和と同様の効果が得られ，空間範囲から少し外れていても規模が大きい地震を取得できる．

他の災害として，台風のデータを取得する場合も，図 2 のように台風の中心軌道を空間条件とすることができる．ここで日本に被害を与えた台風を調べるとすると，日本に上陸した台風だけでなく，図 2 において日本の東を通過した台風のような，上陸はしなかったが規模の大きい台風も取得したい．台風は時間によって移動するため，日付ごとに空間条件となる地点が異なるデータとなる．この場合，最も日本に接近した時の日本との距離，最大風力，そして暴風範囲によって有意性を評価する．

2. 関連研究

地球科学データを決定するためには，データセット，時間，空間の 3 要素を指定する必要がある．立床らの研究[6]では，これらの 3 つの要素の一つであるデータセットを，Wikipedia と地球科学のドメインオントロジーを利用してキーワード検索をする手法が提案されている．

この手法では，まず指定されたキーワードに対応する Wikipedia の記事に張られたリンクを利用して，キーワードと関連の強い概念を抽出する．次に，これらの概念のうち，ドメインオントロジーに存在しないものを除外する．その後，ドメインオントロジーの語彙

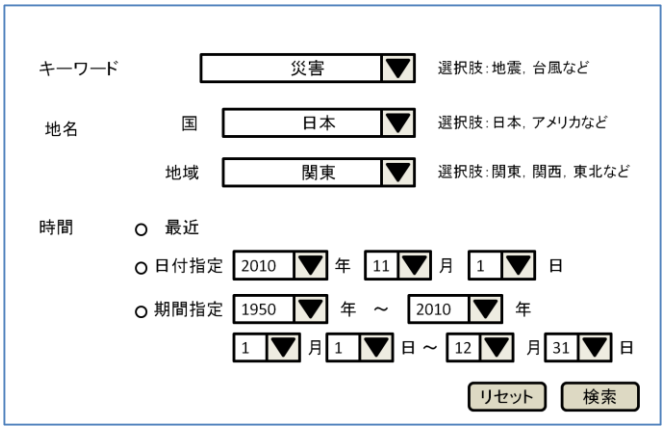

図 3　検索フォーム

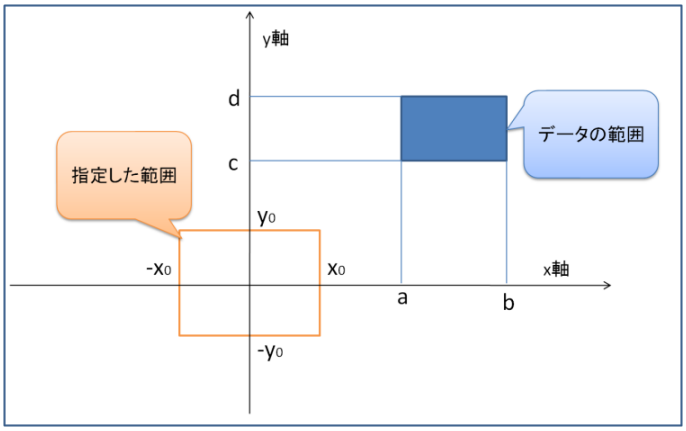

図 4　条件とデータの時空間範囲

とデータセットの対応付けから，抽出した概念に関係するデータセットを取り出す．この研究では，NASA が開発した地球科学オントロジーである SWEET をドメインオントロジーとして使用している．

石川の研究[7]では，位置が確定していないオブジェクトから，周辺にあるオブジェクトを検索するための空間問合せ処理方法を考えている．特に，その位置が正規分布で表されるオブジェクトから，位置が確定しているオブジェクトを，距離に基づく範囲問合せによって検索する場合の処理方法を提案している．

具体的には，その位置が正規分布で表されるオブジェクトを q，q が(x，y)に存在する確率を $p_q(x, y)$とする．そして，q との距離が δ 以下である確率が θ 以上であるようなオブジェクトの集合を PRQ(q,δ, θ)で表し，この集合に含まれるオブジェクトを全て取得するのが目的である．検索対象のオブジェクトを a とし，それを中心とする半径 δ の円を R とすると，

$$\iint_{(x,y)\in R} p_q(x,y)dxdy \leq \theta$$

ならば a は PRQ(q,δ, θ)に含まれる．

以下の研究では，本研究と同様に，地球科学データの予備知識がなくてもデータの検索を可能にするシステムを提案している．

高橋らの研究[8]では，地球科学データとメタデータを関連付けるアノテーションシステムを提案している．このシステムには，地球科学データの粒度に応じた柔軟なアノテーションを提供し，さらにアノテーションされた情報を自動的に関連するデータに伝播させる機能がある．これによって，データ利用者による自由なデータのアノテーションおよび共有を可能にするとともに，データ提供者のメタデータ構築作業の軽減を実現している．

地球科学データの中でも，衛星画像の検索を容易にするシステムを提案しているのが，岡本らの研究[9]である．この研究では，イベント名などの検索クエリに対応した，Wikipedia のページ内の infobox テンプレートとジオタグに含まれる情報と，記事内容のパターンマッチングにより，検索したいイベントに対応した時空間情報を取得する．そして，取得した時空間情報を利用して既存の Web GIS システムから衛星画像の検索，閲覧を行っている．

3. 提案手法

本研究では，関連研究[6]の手法を利用して，指定したキーワードからデータセットの絞り込みを行う．その後，指定した時空間範囲とデータの時空間条件との近似度を関連手法[7]を応用して計算し，近似度と重要度が高いデータを優先して提示する．

3.1. 検索インタフェース

本研究で提案するデータベース検索では，一般的な地球科学データベースを踏襲し，図 3 のようにキーワード，時間範囲，空間範囲を選択あるいは指定することを想定している．空間条件は，国名と地域名を地名として選択できる．その地名に対応する長方形の範囲を指定した空間範囲とする．

時間条件としては，開始年月日と終了年月日をそれぞれ指定し，年と月日の 2 次元空間上の長方形を時間範囲とする．1 つの年月日だけを指定することもでき，その場合は時間範囲を点で表す．開始年と終了年は空白にすることも可能であり，その場合は範囲無制限とする．

3.2. スコアの計算

キーワードで絞り込みを行った後のデータについて，以下の 3 つの指標によるスコアを付け，それらをもとに計算した総合スコアが大きい順に検索結果の表示を行う．

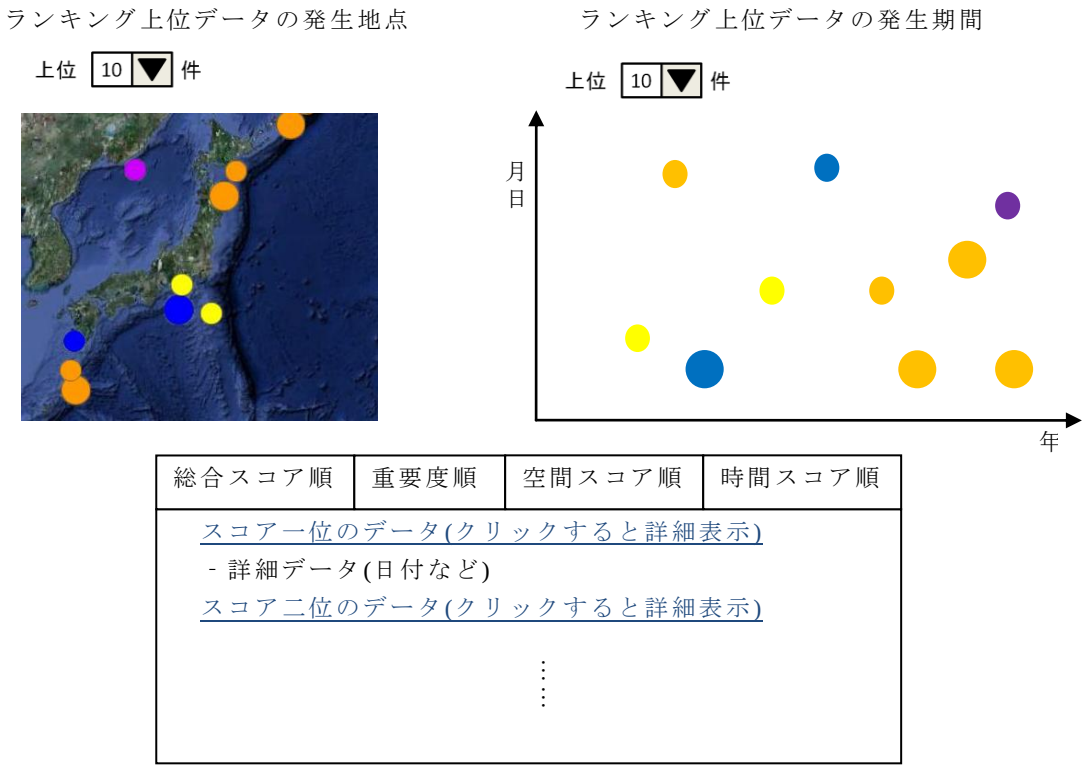

図 5　推薦データ表示

- データ自体の重要性
- 空間検索条件とデータの観測地との距離
- 時間検索条件とデータの観測日の差

以下にそれぞれのスコアの計算方法案を示す.

3.2.1. データ自体の重要度によるスコア

地震のマグニチュードと震度，台風の風速と大きさなどの現象量があるデータの場合は，現象量が大きいほど有意であるとして，スコアを大きくする．そうでないデータは，データの種類と観測地点と観測日ごとに有意か否かを指定する．また，スコアの計算にかかるコストを減らすため，現象量が閾値以下のデータは合計のスコアを 0 とすることも考えられる.

3.2.2. 空間条件と観測地の比較によるスコア

関連研究[7]では，検索を行うのが点で表されるオブジェクトであったが，本研究ではそのオブジェクトを長方形の範囲に置き換えたものとみなせる．そして，正規分布による確率の大きさに応じてスコアを付ける．x 軸を経度，y 軸を緯度として，図 4 のように検索範囲とデータの範囲を定めた場合，スコアの計算式は以下のようになる.

$$\frac{\int_a^b \int_c^d (p_q(x) + p_q(y))dydx}{(b-a)(d-c)}$$

ただし，

$$p_q(x) = \begin{cases} 1\ (-x_0 \le x \le x_0) \\ \exp\left(-\frac{(|x| - x_0)^2}{2}\right)(x < -x_0, x_0 < x) \end{cases}$$

とする.

データの範囲は，図 4 においては長方形で表されているが，地震は震源位置の点，台風は 1 日ごとの中心位置の点の集合で表される.

3.2.3. 時間条件と観測日の比較によるスコア

時間条件については，図 4 において年を x 軸，月日を y 軸とすることにより，空間条件と同様にスコアを決める．データの範囲は地震や台風が発生してから消えるまでの期間とする．時間を 2 次元で表すことにより，季節が同じで年が異なるデータを優先して取得できる.

3.3. 検索結果表示

検索，推薦結果の表示方法は，図 5 のようなものを想定している．総合スコアの高いデータから順番に概要を表示し，興味を持ったデータについてその詳細を表示できるようにする．また，個別のスコアが高い順にも表示できるようにする．さらに，総合スコアが上位のデータについて，地図上あるいは時間平面上に観測地点及び観測期間を表示することで，例えば大きな地震や台風がどの場所や時期に集中しているのかわかるようにする.

年	月	日	緯度	経度	マグニチュード	時間スコア	空間スコア	総合スコア
2000	10	6	35.456	133.134	6.7	1	2	0.605
2000	7	30	33.901	139.376	6.5	1	2	0.587
2000	6	9	30.491	137.73	6.3	1	2	0.569
2000	10	3	40.282	143.124	6.3	1	2	0.569
2000	8	6	28.856	139.556	7.4	1	1.519	0.562
2000	6	3	35.552	140.464	6.2	1	2	0.560
1999	4	8	43.607	130.35	7.1	0.607	2	0.557
2000	7	15	34.319	139.26	6.1	1	2	0.551
2000	7	1	34.221	139.131	6.1	1	2	0.551
2000	6	6	29.424	131.421	6.4	1	1.847	0.549

表 1　スコア上位 10 位までの地震データ

3.4. スコア付けの例と考察

ここで，実際の地震のデータについてスコアを計算した．データは Dagik の地震データの提供元である USGS(U.S. Geological Survey)のウェブサイト[10]より取得した．時間条件は 2000 年 1 月 1 日，空間条件は日本を想定して，北緯 30～45 度，東経 130～145 度の長方形の範囲とした．あらかじめ 1997 年から 2003 年に起こったマグニチュード 6 以上の地震に絞り込んでからスコア計算を行った．

この条件で総合スコアが高かった地震を表 1 に示す．ただし時間スコアは震源の緯度と経度を，空間スコアは地震が発生した年と月日を，それぞれ x,y とした時の

$$p_q(x)+p_q(y)$$

であり，それぞれ最大値は 2 となる．総合スコアは

((時間スコア)＋(空間スコア))÷4×(マグニチュード)
÷(マグニチュードの最大値)

とした．

この例では上位 10 位に入るデータは，震源が時間条件の範囲に含まれる上に発生したのが 2000 年の地震であることが多かった．その原因は，スコアの計算に正規分布を利用したため，指定した条件とデータの距離あるいは発生年月日の差が大きくなると，スコアが極端に小さくなるためである．そのため，指定した時間や空間の条件から外れているが重要度が高いデータのスコアを上げるためには，スケーリングを広くする，あるいは正規分布以外の関数を $p_q(x)$に適用する必要があると考えられる．

また，上記の計算式では総合スコアがマグニチュードに比例するようになっているが，マグニチュードと地震のエネルギーの関連を考えると，スコアはマグニチュードに応じて指数的に増加するほうが適切と考えられる．

4. まとめ

本論文では，データの重要度及びデータと検索条件の時空間近接度をもとに算出したスコアが高いデータを推薦することで，有意なデータの取得を容易にするシステムを提案した．

今後は，3 つの指標によって算出したスコアをどのように組み合わせて総合スコアを計算するかを決めることが課題となる.今回の例以外の地球科学データに，今回の例とは異なるスコアの計算式を適用し，それらの計算式の有用性を示す指標を明らかにする予定である．特に，重要度は現状では検索条件によらず一定となっているが，ユーザの求めるデータ内容は一定ではないと考えられるので，条件指定による変数を含んだ計算式にするつもりである..

さらに，地震や台風といったイベントのデータだけでなく，震度や降水量などの観測データもデータベースに含める予定である．イベントデータについては 1 つのイベントを 1 つのデータとしていたが，観測データについては，1 か所の観測地点で 1 日に観測した内容を 1 つのデータとすることが考えられる．

参　考　文　献

[1] A. Saito and D. Yoshida, “Dagik: A Data-Showcase System for the Geospace”, Data Science Journal, 8, S92-S95, doi:10.2481/dsj.8.S92, 2009.

[2] Dagik http://dagik.org

[3] Gfdnavi: Geophysical Fluid Data navigator http://www.gfd-dennou.org/arch/davis/gfdnavi

[4] 諫本 有加，渡辺 知恵美，堀之内 武，西澤 誠也, “Gfdnavi における対話的横断検索の実現”, 電子情報通信学会技術研究報告，データ工学, vol.109, no.186, pp.21-26, 2009-08-31.

[5] GCMD: Global Change Master Directory http://gcmd.nasa.gov

[6] 立床 雅司，齊藤 昭則，清水 敏之，吉川 正俊，“Wikipedia とドメインオントロジーの統合利用による地球科学データ推薦手法”，電子情報通信学会技術研究報告，データ工学 vol.109, no.186,

pp.39-43, 2009-08-31.

[7] 石川　佳治, “曖昧な位置情報に基づく空間問合せの処理手法”, DBSJ Letters Vol. 6, No. 2.

[8] 高橋　慧, 立床　雅司, 絹谷　弘子, 吉川　正俊, “地球科学データアノテーションシステムの構築”, 電子情報通信学会 第13回Webインテリジェンスとインタラクション研究会, 2008

[9] 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

[10] USGS Earthquake Search
http://earthquake.usgs.gov/earthquakes/eqarchives/epic/