

自己記述的多次元数値データ間の演算での次元対応の自動化の試行 - ドラッグアンドドロップであらゆるデータ演算ができるか?

How can we build dimension matchings in automatic manipulation of self-describing multi-dimensional numerical data?

豊田 英司[1], 地球流体電脳倶楽部 Davis Project 林 祥介

Eizi TOYODA[1], GFD Dennou Club Davis Project Hayashi Yoshi-Yuki

[1] 気象庁 数値

[1] JMA NPD

<http://www.gfd-dennou.org/arch/davis/>

多次元数値データの自己記述化を実用化するために、操作の自動化の要件を吟味している。

次元順が必ずしも同じでない複数のデータを用いる演算においては、次元の対応をどうとるかが問題となる。

ここでは gtool4 netCDF 規約を想定して、対応手順を提案する。

* はじめに

本学会の前回大会でも紹介したように、地球流体電脳倶楽部 davis プロジェクトでは気象を含む地球流体現象に関するあらゆる多次元数値データを自己記述的に格納できるメタデータ形式と処理系を開発している。キャッチフレーズは「クリックするだけで絵が描けるデータ」である。

現在下位データ形式(NetCDF)を抽象化するインターフェイスとそれを用いたメタデータ形式案 (gtool4 netCDF 規約,

<http://www.gfd-dennou.org/arch/gtool4/conventions/>) の開発は一段落し、さまざまな応用を想定

しながらメタデータ形式を再吟味している。今回はその一環として、複数のデータによる演算に際して次元の対応を機械的に判定するための試みについて述べる。

* 問題設定

多次元数値データを複数用いる単純な演算を行うことを考える。単純とは、格子配置が一致していれば多次元配列の要素ごと演算に帰着するもの(たとえば四則演算)とする。

次元とはたとえば時空間4次元のことであるが、数値モデルのパラメタなど、その他さまざまなものを考える。格子配置も一致しないかもしれない。

信頼のおけるデータ処理環境は、このような演算を任意のデータについて正しく行うか、それが不可能である場合はせめて格子配置が一致しない旨の報告ができなくてはならない。そのための最初の問題が格子配置を「同じ次元」について比較するため

の次元の対応付けである。

* 外延的対応の限界

問題を低レベルに「多次元配列どうしの次元組み合わせの決定」ととらえると、「格子数の一致する次元どうしを対応させる」という方式が考えられる。

この方式は格子配置が一致する場合だけを想定すれば有効かつ明解である。しかし、格子配置が一致しないが同じ空間にあるデータは補間によって対応をとって演算したいという用途には無力である。

* メタデータを解釈する手順案

メタデータのうち次元を特定するために最も重要なのが単位である。ある次元についての格子位置を比較できるためには、その次元のもつ単位が加算可能でなければならぬ。単位を比較し、加算不可能な単位を対応させないようにしなければならない。東経の度と北緯の度のような単位の対は加算可能ではあるが、できることならば組み合わせたくない。幸い、NetCDF ユーザの中で標準的に用いられる Uduunits ライブラリの書式で東経と北緯の度は Degree_E, Degree_N として区別されているので、単位の表現が一致するかどうかで判定するのがよいであろう。

これらの条件について次元名全体や次元名の最後のワードなどを比較していけば、かなりの確度で「同じ次元」を判定できるものと予想している。

* 実装の取り組み

以上見てきたように、メタデータ解釈は文字列の解析の比重が大きい。特に、単位の加算可能性の判定は構文解析を必要とする。これらを移植性良く Fortran で実現するため、現在単位解析ライブラリとその中で用いる正規表現パッケージを Fortran で書く作業を進めている。

これらをふくむ地球流体電脳倶楽部の活動については <http://www.gfd-dennou.org/arch/davis/> を参照されたい。