連想検索エンジン GETAssoc の 超高層物理学におけるメタデータ・データベースへの適用

八木 学† 小山 幸伸 阿部 修司 梅村 宜生 新堀 淳樹 堀 智昭

田中 良昌 上野 悟 佐藤 由佳 谷田貝 亜紀代 Bernd RITSCHEL

†東北大学大学院理学研究科惑星プラズマ・大気研究センター
〒980-8578 宮城県仙台市青葉区荒巻字青葉 6 番 3 号

E-mail: †yagi@pparc.gp.tohoku.ac.jp

あらまし 地球規模の物理現象を解明する為には、様々なデータを用いた分野横断的な研究推進が重要である。これには、様々なデータを直ちに取得する為のインフラストラクチャが 1 つの鍵となる。超高層物理学を対象とした IUGONET プロジェクトにおいては、複数の研究機関において分散管理されている様々なデータの所在情報等を提供するメタデータ・データベースを構築することで、この課題に対処した。しかしながら、広範な研究領域を対象としている本メタデータ・データベースにおいて、検索語句の選択が専門分野外のユーザーにとって難しいことが指摘された。そこで、本メタデータ・データベースと連想検索エンジン GETAssoc を連携させ、ユーザーによって入力された検索語句の関連語を用いた、再クエリ文字列の自動生成を検討した。そこで、Wikipedia のデータベースを元に、自動的に辞書を作成する 2 種の試みを行った。一方はその膨大なデータ量の取り扱いが原因で、他方は対象ドメインを超高層物理学周辺に絞り込む為の情報の欠落により、期待した関連語が得られなかった為、いずれも断念した。最終的に、超高層物理学分野周辺の、別々の研究背景を持つ研究者らの人力による辞書作成を行い、期待した連想検索結果を得ることが出来た。

キーワード データベース, メタデータ, 連想検索, 情報検索, 学際的研究, 超高層大気

Application of the associative searching system GETAssoc to the metadata database of the upper atmosphere

Manabu YAGI†, Yukinobu KOYAMA, Shuji ABE, Norio UMEMURA, Atsuki SHINBORI, Tomoaki HORI, Yoshimasa TANAKA, Satoru UeNo, Yuka SATO, Akiyo YATAGAI, and Bernd RITSCHEL

†The Planetary Plasma and Atmospheric Research Center,
Graduate School of Science Tohoku University

Aoba 6-3, Aramaki, Aoba, Sendai, Miyagi, 980-8578, Japan

E-mail: †yagi@pparc.gp.tohoku.ac.jp

Abstract Multidisciplinary researches using many kinds of data are crucial to studies on global-scale phenomena in the upper atmosphere. An infrastructure to access to many kinds of data is one of the keys on the multidisciplinary researches. The Inter-university Upper atmosphere Global Observation NETwork (IUGONET) project solved this problem by developing a metadata database to provide information such as whereabouts of data, which are dispersion-managed by several institutes. Because the metadata database covers the wide scientific field, it is pointed out that researchers from other scientific fields can not easily give a right query keyword. We implemented associative search engine GETAssoc into the metadata database, in order to get the related words and to create re-query strings by using themautomatically. We tried to create dictionary from two types of Wikipedia database automatically. Since handling of the huge data was difficult, one side was given up. For another side, the expected related term was not obtained by lack of information to narrow down target domain to upper atmospheric research field. Finally, the dictionary was manually created by the domain researchers of astronomy, meteorology and upper atmospheric research, we could get the expected related term to use re-query.

keywords Database, Metadata, Associative Search, Information Retrieval, Interdisciplinary Study, Upper Atmosphere

1. はじめに

高度数十 km より上空の大気は「超高層大気」と呼ばれており、中性大気と電離大気が混じり合う「熱圏・電離圏」領域から、太陽風と地球磁場の相互作用によって生じ、無衝突プラズマが物理を支配する「磁気圏」領域までを含む。超高層大気は、下層大気からの波動や対流による運動量・エネルギーの流入に加え、太陽放射や太陽風の影響、そして化学反応・光化学反応といった物理過程が複雑に絡み合う領域であることが知られている。従って、超高層物理学の研究においては、全球規模の観測データを用いた多角的なデータ解析が必要であり、磁力計、レーダー、そして太陽望遠鏡など、様々な観測器を用いた地上観測が継続的に行われてきた。こうした観測データは、観測を行った各研究機関ごとにデータベース化され公開されてきたが、分散管理されたこれらの観測データを横断的に検索するシステムは過去に存在しなかった為、多くの種類のデータを必要とする地球規模の学際的研究推進に多大な労力を要した。この様な背景から、大学間連携プロジェクト「超高層大気長期変動の全球地上ネットワーク観測・研究（IUGONET: Inter-university Upper atmosphere Global Observation NETwork）」においては、観測データに関するメタデータ・データベースを構築し、各機関において分散管理されている観測データに関するメタデータを、横断的に検索できるシステムを構築した[1][2][3][4][5][6][7]。

図 1 に示した Web ベースの IUGONET メタデータ・データベース[8]は、時刻検索、領域検索、そして単語検索が可能である。IUGONET メタデータ・データベースによって、各機関が管理する観測データに関するメタデータを横断的に検索出来る仕組みを構築したものの、超高層物理学を中心としつつ、隣接する天文学や気象学に渡る、広範な研究領域を対象にしているが故に、単語検索時における検索語句の選択が専門分野外のユーザーにとって難しいことが指摘された。そこで、連想検索エンジン GETAssoc[9]を導入することによって、ユーザーが入力した検索語句の関連語を複数個提示し、そのリンクをクリックすることにより、再検索を行う機能を実装し、メタデータ・データベースに組み込むことを検討するに至った。

2. IUGONET メタデータ・データベース

太陽地球系物理学分野の主に衛星データ向けに開発された Space Physics Archive Search and Extract (SPASE)メタデータ・フォーマット[10][11]をベースとして、地上観測データに関する要素を拡張したメタデータ・フォーマットが、IUGONET 共通メタデータ・フォーマット[12]である。

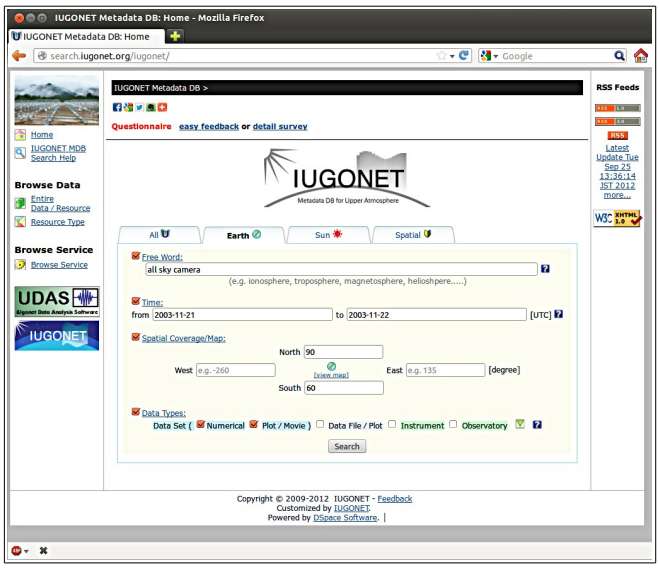

図 1: IUGONET メタデータ・データベース。時刻検索、領域検索、そして単語検索が可能である。

このメタデータ・フォーマットに則って作成された XML 形式のメタデータは、機関リポジトリ[13]等で利用される DSpace[14]をベースにカスタマイズした IUGONET メタデータ・データベースにインポートされる。標準で Dublin Core メタデータ・フォーマット[15]を取り扱う DSpace からの主要なカスタマイズ箇所は、1. 観測日時や緯度経度の範囲を取り扱う点、2. 機関リポジトリにおいては PDF ファイル等のデジタルコンテンツが保持される Bitstream[16]に、XML 形式のメタデータを保持する点、である。観測データの所在情報がメタデータ内に含まれているため、分散管理されている観測データに到達可能である。2012 年 11 月時点において、IUGONET メタデータ・データベースに 700 万件を超えるメタデータが登録されている。

3. IUGONET メタデータ・データベースと連想

検索エンジン GETAssoc の連携

図 2 に IUGONET メタデータ・データベースと連想検索エンジン GETAssoc の連携システム図を示す。GETAssoc は、国立情報学研究所連想情報学研究開発センターで開発された連想検索エンジンである。

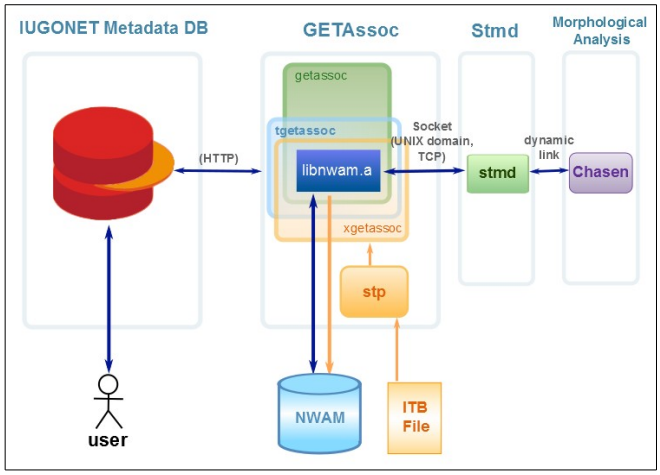

図 2: IUGONET メタデータ・データベースと連想検索エンジン GETAssoc の連携システム図。形態素解析システムには CHASEN を用いた。

IUGONET メタデータ・データベースと GETAssoc は、HTTP プロトコルを用いて接続される。GETAssoc のシステムの構築には、GETAssoc 以外に外部ツールがいくつか必要であるが、それらの外部ツールは、相互にバージョン依存している為、インストール手順が複雑であった。そこで、外部ツールを含めた GATAssoc インストール手順を自動化する為に、Ant のビルドファイルである build.xml を作成した。このビルドファイルは、Scientific Linux 6.3 (32bit)上で動作確認済みであり、GitHub 上の IUGONETAssociativeSearch リポジトリ[17]において公開している。

図 3 に IUGONET メタデータ・データベースと GETAssoc の連携システムにおける検索シーケンス図を示す。検索シーケンスは、1. ユーザーが入力した検索語句を読み込み、2. 通常のメタデータ検索を行い、3. ユーザーが入力した検索語句を元に、メタデータの該当有無に関わらず GETAssoc を用いて連想検索を行い、4. メタデータならびにユーザーが入力した検索語の関連語を表示する、5. ユーザーは、必要に応じて関連語を用いて再検索を行う、という流れを想定している。ブラウザ上で表示された関連語は、HTTP で記述されたクエリ文字列へのリンクになっている為、関連語をブラウザ上でクリックすると、直ちに再検索される。

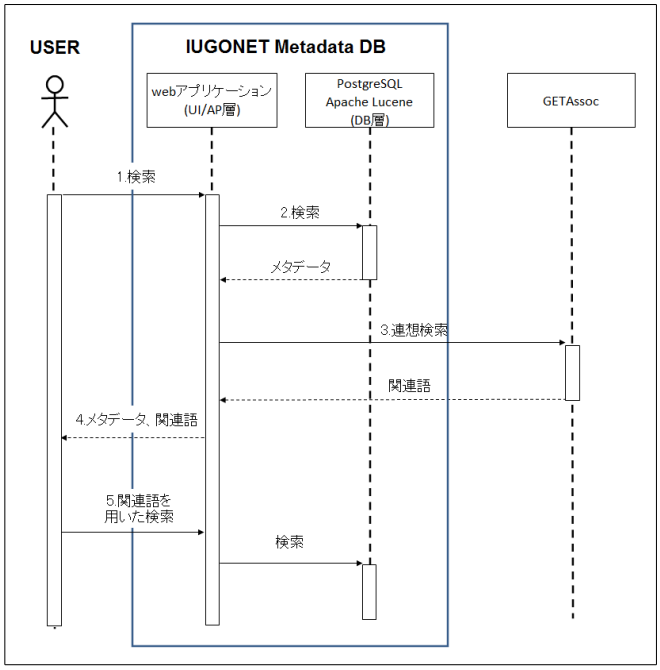

図 3: IUGONET メタデータ・データベースと GETAssoc の連携システムにおける検索シーケンス図。

4. 辞書の作成

GETAssoc においては、図 4 に例示した ITB ファイル形式によって関連語を結びつける為の辞書を記述し、インデックス作成コマンドである stp を用いて、連想計算に必要なインデックス（NWAM）ファイルに変換する必要がある（図 2）。つまり、超高層物理学分野のメタデータ・データベースにとって、有効な ITB ファイルを如何にして作成するかという問題に帰着したことになる。

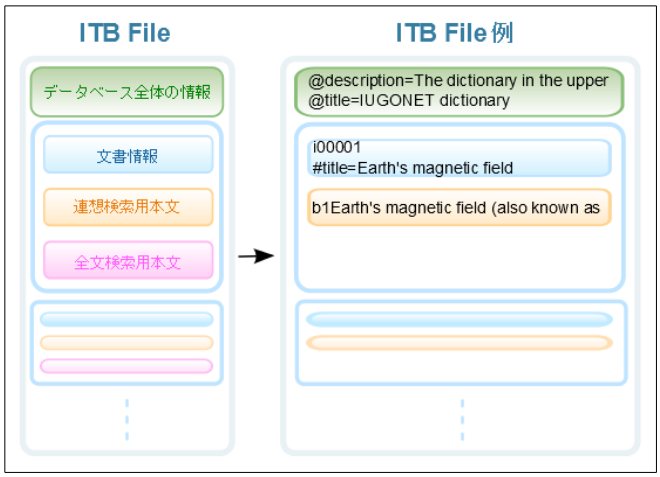

図 4: ITB ファイルの概要と例。連想検索のみに着目している為、本件では全文検索用本文は作成しない。

4.1 Wikipediaのノートページ、利用者ページを除くデータを用いた辞書作成

GETAssocのWebページ[8]において、オンラインで提供されているWikipediaのダンプデータ[18]の利用が例示されていたので、これを辞書として利用することを検討した。対象は、enwiki-20121201-pages-articles¥d{1}|¥d{2}.xml-*.bz2の全27ファイルであり、bz2を展開して得られる全XMLファイルの合計サイズが約40GBに及ぶ。このXMLファイルから、ITB形式が要求するtitle要素(#title)と連想検索用本文(b1)を抽出する必要がある。mediawiki/page/text内にITB形式の両要素に対応する要素がWiki記法で押し込められており、これを円滑に抽出する上で、Wikipedia Parserの利用が有用である。本稿執筆時点においてはWikipedia Parserを使用せず、上記XMLファイルのごく一部に対し、XSLTとAWKを用いて手動で要素を抽出した上で、インデックス作成をテスト的に行なった。このテストの結果、上記の全27ファイルを対象としたITBファイルへの整形、ならびにインデックス作成に大量の時間、計算機資源が必要であることを確認した。4GB程度のメモリ、そして1コアのCPUで運用可能な我々のメタデータ・データベースの補助的な役割として期待している連想検索のシステムが、本体以上に大掛かりになると運用上都合が良くないので、この辞書作成を断念した。

4.2 Wikipediaの要約を用いた辞書作成

次に、Wikipediaの全ページの要約を提供するenwiki-20121201-abstract¥d{1}|¥d{2}.xmlの利用を検討した。XMLファイルの合計が約3.6GBであり、このサイズは前者より圧倒的に小さい。このWikipediaの要約が保持している内容は、レコードのタイトルを示すfeed/doc/title、本文のURLを示すfeed/doc/url、Wikipediaコンテンツの最初の1文を保持したfeed/doc/abstract、関連コンテンツへのリンクを示したfeed/doc/linksのみで構成される。図 5に示す様に、1.Wikipediaのデータベースをダウンロードし、2.XSLTを用いて整形されたITBファイルに変換し、3.AWKを用いて明らかに不要な文字の排除や辞書項目のナンバリングなどの処理、を行った後にGETAssocが取り扱える妥当なITBファイルを作成した。この一連の作業は、前述のビルドファイルとAntを用いることによって、自動的に変換される。

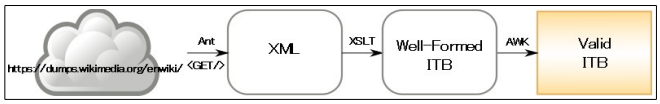

図 5: ITBファイル作成の流れ。

実際には、XML形式で記述されたWikipediaのfeed/doc/title要素とfeed/doc/abstract要素を、ITB形式のtitle要素(#title)と連想検索用本文(b1)として変換することで、ITB形式の辞書ファイルを作成した。この辞書を用いて連想検索を行ったところ、一般用語が関連語として上位に挙げられる等、期待した結果を得ることが出来なかった。連想検索対象のドメインを超高層物理学分野に絞り込むことが出来れば、連想検索結果が改善すると期待されるがWikipediaの要約にはドメインを絞込む為の要素が無い為、それを絞り込むことが出来なかった。

4.3 人力による辞書作成

超高層物理学に関連する辞書をゼロから作成することは、人的資源的に高負荷であると考え、Wikipediaを用いた上記2案を試みたが、結果が芳しくなかった為、最終的に人力による辞書作成に帰着した。

最初に、Googleスプレッドシート上で辞書[19]を用意し、そのスプレッドシートの第1カラムならびに第2カラムを、各々ITBファイル形式におけるtitle要素(#title)ならびに連想検索本文(b1)と定義した。次に、メタデータ・データベースの単語検索において検索されるであろう単語の語呂出しを行った。その単語もしくは近い単語に関するWikipedia記事のtitleを、スプレッドシートの第1カラムに入力した。その単語に関する記事の先頭文章を、スプレッドシートの第2カラムに手動でコピーした。この先頭文章を取り出すアイデアは、Wikipediaの要約と同様である。

Googleスプレッドシートは共有可能な為、複数の辞書作成者によって辞書作成を行うことが出来、高負荷な辞書作りの負荷分散が可能となる。さらに、辞書作成には、天文学、超高層物理学、気象学の各々の研究背景を持つIUGONETプロジェクト開発者が携わるので、IUGONETメタデータ・データベースの対象分野をおおよそ覆うことが出来た。

上記の手法によって作成したスプレッドシートをITB形式に変換する為に、スプレッドシートを外部からクエリーを受け付け可能な公開状態にし、スクリプトによって外部からクエリーを発行してレコードを取り出し、ITB形式のファイル

を生成した。このスクリプトの作成には、Ruby と google-spreadsheet-ruby ライブラリを用いた。ITB ファイルが 1000 件程度であれば、1 分とかからないうちに変換作業が終了する。その為、cron 等で数分置きに ITB ファイルへ変換する様に設定しておき、スプレッドーシート更新後、数分で連想検索に反映される様に設定した。辞書編集結果の速やかな反映により、高負荷な辞書作成者による人力の辞書作りのモチベーションが保たれると考える。この超高層物理学分野周辺にドメインを絞り込んだこの辞書を用いた連想検索結果は、図 6 に例示するように、適切な連想検索結果が得られた。

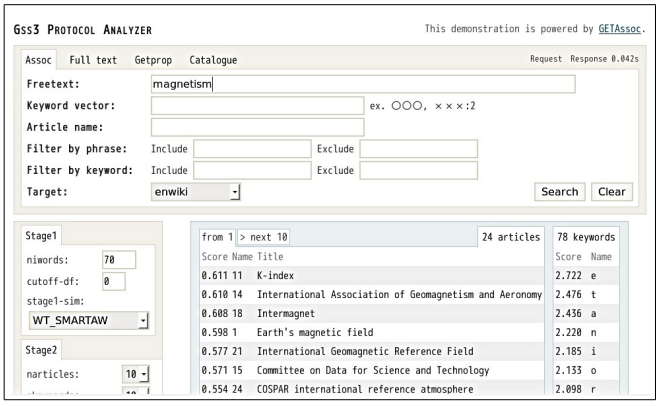

図 6: 作成した辞書ファイルを用い、gss3 protocol analyzer を用いて連想検索した例。

5. まとめ

超高層物理学を対象にした IUGONET メタデータ・データベースは、検索語句の選択が専門分野外のユーザーにとって難しいことが指摘されていた。そこで本件では、GETAssoc を用いた連想検索を導入することにより、ユーザーによる入力検索語句の関連語を用いた再クエリ文字列の自動生成を検討した。最終的に帰着した、超高層物理学分野のメタデータ・データベースにとって、有効な辞書を如何にして作成するかという問題に対して次の 3 つの取り組みを試行した。Wikipedia のノートページ、利用者ページを除くデータを用いた辞書作成は、その大量のデータ量、計算機リソースから利用を断念した。Wikipedia の要約は、取り扱い上適切なファイルサイズであるものの、超高層物理学分野というドメインへ絞り込む為の要素が欠落していたので、期待した連想検索結果が得られなかった。超高層物理学周辺の、しかしながら別々の研究背景を持つ研究者らの人力による辞書作成においては、辞書作りの負荷を軽減する環境を整備した上で辞書作りに取り組んだ結果、期待した連想検索結果を得ることが出来た。

参 考 文 献

[1] 林 寛生, 小山 幸伸, 堀 智昭, 田中 良昌, 新堀 淳樹, 鍵谷 将人, 阿部 修司, 河野 貴久, 吉田 大紀, 上野 悟, 金田 直樹, 米田 瑞生, 田所 裕康, 元場 哲郎, 大学間連携プロジェクト「超高層大気長期変動の全球地上ネットワーク観測・研究」, JAXA-RR-11-007, pp. 113-120, 2012.

[2] Hayashi, H., Y. Koyama, T. Hori, Y. Tanaka, M. Kagitani, A. Shinbori, S. Abe, T. Kouno, D. Yoshida, S. UeNo, N. Kaneda, M. Yoneda, H. Tadokoro, T. Motoba, and IUGONET project team, ”Inter-university Upper atmosphere Global Observation NETwork (IUGONET)”, The First ICSU World Data System Conference, 2011.

[3] 小山 幸伸, 河野 貴久, 堀 智昭, 阿部 修司, 吉田 大紀,林 寛生, 田中 良昌, 新堀 淳樹, 上野 悟, 金田 直樹, 米田 瑞生, 元場 哲郎, 鍵谷 将人, 田所 裕康, 超高層物理学分野の為のメタデータ・データベースの開発, JAXA-RR-11-007, pp. 91-98, 2012.

[4] 堀 智昭, 鍵谷 将人, 田中 良昌, 林 寛生, 上野 悟, 吉田 大紀, 阿部 修司, 小山 幸伸, 河野 貴久, 金田 直樹, 新堀 淳樹, 田所 裕康, 米田 瑞生, IUGONET 共通メタデータフォーマットの策定とメタデータ登録管理システムの開発, JAXA-RR-11-007, pp. 105-111, 2012.

[5] 小山 幸伸, 河野 貴久, 林 寛生 , 堀 智昭, 田中 良昌, 鍵谷 将人, 吉田 大紀, 上野 悟, 阿部 修司, 三好 由純, 金田 直樹, 能勢 正仁, 岡田 雅樹, 超高層物理学分野におけるメタデータ・データベースの構築 , DEIM Forum 2010 F4-3.

[6] 河野 貴久, 小山 幸伸, 堀 智昭, 阿部 修司, 吉田 大

紀, 林 寛生, 新堀 淳樹, 田中 良昌, 鍵谷 将人, 上野 悟, 金田 直樹, 田所 裕康, DSpace を用いた超高層物理学のためのメタデータ・データベースの構築, DEIM Forum 2011 C8-5.

[7] 梅村 宜生, 小山 幸伸, 堀 智昭, 阿部 修司, 林 寛生, 新堀 淳樹, 田中 良昌, 上野 悟, 米田 瑞生, 金田 直樹, 元場 哲郎, 超高層物理学のための分野横断型メタデータ・データベースの構築, DEIM Forum 2012 A7-1.

[8] http://search.iugonet.org/iugonet/

[9] http://getassoc.cs.nii.ac.jp

[10] Todd King, James Thieman, and D. Aaron Roberts, SPASE 2.0: a standard data model for space physics, Earth Sci Inform, 3:67-73, 2010.

[11] Thieman, J. R., D. A. Roberts, T. A. King, C. C. Harvey, C. H. Perry, and P. J. Richards, SPASE AND THE HELIOPHYSICS VIRTUAL OBSERVATORIES, Data Science Journal, 25 February, 2010.

[12] http://www.iugonet.org/data/schema/iugonet.xsd

[13] http://www.nii.ac.jp/irp/list/

[14] http://www.dspace.org/

[15] http://dublincore.org/

[16] http://www.dspace.org/1_7_0Documentation/Architecture.html

[17] https://github.com/iugonet/IugonetAssociativeSearch

[18] http://dumps.wikimedia.org/enwiki/

[19] https://docs.google.com/spreadsheet/pub?key=0AgbaOOcOsZncdDFYa2xHd0puZnVBVUVwakp0bzJuVkE&output=html