様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １５ 日現在

機関番号：８２６２６
研究種目：基盤研究(A)
研究期間：2012～2014
課題番号：２４２４００１５

研究課題名（和文）大規模・異種の時空間データ統合で生じる矛盾を許容するサイエンスクラウド基盤

研究課題名（英文）Science cloud Infrastructure which can support data contradiction in large scale heterogeneous spacio-temporal data.

研究代表者

小島　功（Isao, Kojima）

独立行政法人産業技術総合研究所・情報技術研究部門・総括研究主幹

研究者番号：００３５６９８２

交付決定額（研究期間全体）：（直接経費）　　36,200,000円

研究成果の概要（和文）：本課題は、大規模な時空間データの統合において発生する「矛盾」に多角的に取り組んだもので、要素研究と実証システム開発の課題を行った。
事例として画像データとWeb上のソーシャルデータ等、矛盾を含んだ異種のデータを統合して土地利用等の知見を抽出する研究を進め、衛星画像処理より高精度な結果を実現、実証した。基盤技術としても、異種メタデータを統合して検索する技術や、それを機械学習などで解析するための連携手法等について研究し、それぞれ従来法より優れた結果を得た。原発事故関連の環境モニタリングデータを対象に国際標準に基づく統合検索の実証システムをクラウド上に構築し今後の実用化と標準化へ橋渡しできた。

研究成果の概要（英文）：This research topic challenged the problem of data contradiction which often happens in the integration of large-scale heterogeneous spacio-temporal data.
Major results include: 1) Knowledge extraction method by integrating heterogeneous spacio-temporal data such as photo images, SNS messages which is available to the internet. Several example applications, such as landcover analysis shows the effectiveness of this method over existing methods like satellite image analysis. GPU based outlier detection is also developed.
2) Data querying/managing platform for heterogeneous (Linked Open) data. The results include best-effort federated SPARQL processor under the time constraints, olap mechanism with Linked Data and database based federation method which combines data management and machine learning. 3) In order to show the usefulness of the results, we created a prototype of the integrated cloud-based system for radiation monitoring data related with the Fukushima nuclear accident.

研究分野：データベース

キーワード：データベース　矛盾　データ統合　時空間データ　Linked Data　問合せ処理　クラウド基盤

１．研究開始当初の背景

気候変動など全地球規模の社会的問題に対しては、同じく地球規模での問題解決が求められており、地理的に分散した大規模異種データの統合に基づく、第４の科学研究パラダイムとしての e サイエンス(e-Science)が重要視されている。具体的な応用としては、時期の異なる衛星画像同士の全量比較による災害発生地点の発見といったホットスポットの検出や、衛星画像の解析処理と地上観測データを全量的に突き合わせた土地利用の解析、異なる衛星データの相互全量処理による観測漏れ地点の検出、センサの校正といった異種で大量のデータ同士を様々な条件で結合しつつ、解析処理の結果をも統合するというものがある。災害情報など即応性の高い情報提供も重要であり、基盤に対する機能・性能上の要求は極めて高い。

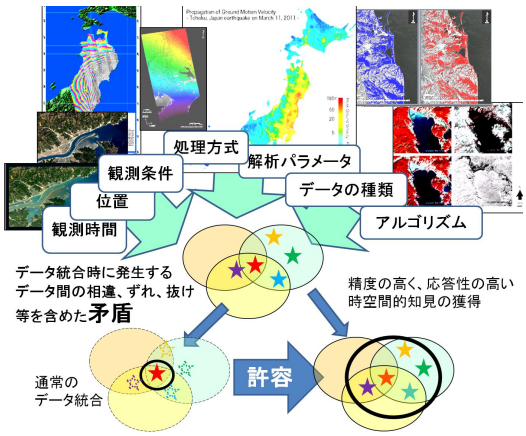
さて、山火事発見のような急を要するホットスポット検出では、異なる衛星に基づいた解析結果を重ね合わせると、多数の地点で一致せず精度の向上が難しい。なぜなら、測定ノイズや誤検出の他にもデータベース処理の結合条件（同一観測時期の許容幅）や解析プログラムのしきい値など、多くのパラメータが複雑に関与しているからである。土地利用解析においても、衛星画像の解析に基づく結果の正確さが高くない（おおむね７割以下程度）ので、複数の解析結果を統合すると結果同士で矛盾するデータが非常に多く発生する。

このように、大規模で異種の情報を統合した場合にはデータ間での相反、ずれ、抜けなど、様々な形の「矛盾」が多数発生する。これは本来、解析プログラムなど応用の課題であり、手法を改良して矛盾を解消、一貫したデータを作る必要があった。しかし応用研究者にとってはデータベース操作も含め統合に関わるパラメータが多すぎることや、全量的なデータ処理自体が大規模で時間がかかるなどから、現実には経験に基づく単純な方法等で処理（解像度の高い衛星の結果を真とする、など）されることも多い。しかしこの手の単純な方法では、例えば消火に有効な初期の山火事を検出するには精度が低すぎて結果の信頼性に欠け、逆に矛盾を解消して一貫性のあるデータを作ると手間が大きく災害などで即応できない。

そこで、これら異種・大量のデータ同士の結合や解析に基づく応用において、必ず発生するデータの「矛盾」をそのまま基盤で許容して統合することが応用の発展に必須と考え、これを効果的に発見できるよう可視化したり、データベース結合やプログラム解析の多数のパラメータ設定を柔軟にするなどの機能によって矛盾を「許容」し、複数の情報源を無理なく統合して新たな時空間的知見を得るための基盤技術が求められている。

２．研究の目的

そこで本研究課題では、図１のようなデータ統合において「矛盾」を許容して知見を獲得することを目的とした。



（図１　矛盾の許容による知見の獲得）

この目的の実現のために既存の大規模データ統合システムやデータアーカイブの研究開発を評価し、以下のような課題にまとめ、これらの解消を目的とした。

①時空間データの矛盾の発見や判断には、多角的な可視化を含めたモデリングと対話的インターフェイス基盤が不可欠である。

②矛盾するデータを含む異種の大規模アーカイブ同士での結合・解析を柔軟に行う検索・解析基盤が必要である。

③可視化・検索・解析の柔軟な連携による即応性の高い情報生成・提供が重要である。特にこれら要素技術を統合して実証できるサービスの実現が重要である。

３．研究の方法

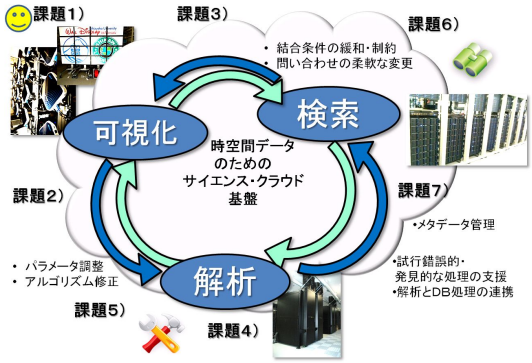
本研究は３年計画として、先の課題設定に対応して、矛盾したデータを許容できるデータ統合のための要素研究（①、②）と、それらを統合するシステム構築と応用サービスの構築・提供による実証（③）を行うものとした。

①　<u>モデル化・可視化</u>：多角的な可視化と対話性の向上により矛盾の発見と許容を扱うモデルとその可視化手法を研究する。具体的には、矛盾を許容するモデルやデータ統合結果データの多角的な可視化、解析ワークフローと可視化との連携、ファセット検索による可視化との連携などの研究を行う。

②　<u>検索・解析基盤</u>：矛盾を許容する柔軟な結合や、探索的な検索・解析処理を支援するデータベース基盤を研究開発する。具体的には、ダイジェストデータや差分に基づいた柔軟な検索、解析処理と並列データベースとの効果的な連携手法、時空間イベントの検出手

法や GPU を使った解析、メタデータによる矛盾や許容のためのパラメータを発見する方法などを研究開発する。

③　統合応用サービス：時空間ホットスポットの検出等地理空間応用を支援するクラウドサービスの研究開発。これを産総研 GEO Grid アーカイブを中心として各システムを連携してクラウドとして実現し、手法の実用性を検証する。



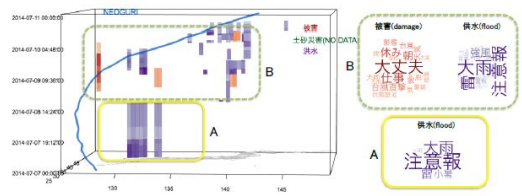
（図２　研究開発の概要）

４．研究成果

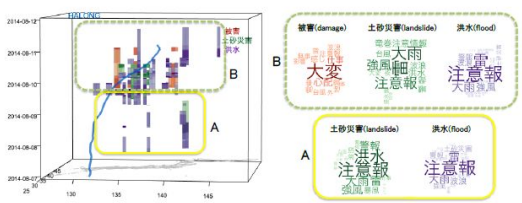
①　モデル化と可視化：

モデル化と可視化においては、特にソーシャルデータと地理空間データのデータ統合とその過程で発生する矛盾を解消するため、柔軟なデータ統合を行って可視化する研究開発を行った。これらの研究の主眼は、データ矛盾や不均質さを大きく含むため通常の結合演算等では結果が意味を持たないいデータについて、解析やクラスタリング、可視化を組み合わせることで知見を抽出しようとするものである。

特に（成果７）では、Twitter の時空間的な動きを解析し、さらに台風の進路など地物データとを統合し、可視化することで、概念間の関連を発見することができた。



(a) NEOGURI [2014/07/07-2014/07/11]



(b) HALONG [2014/08/07-2014/08/12]

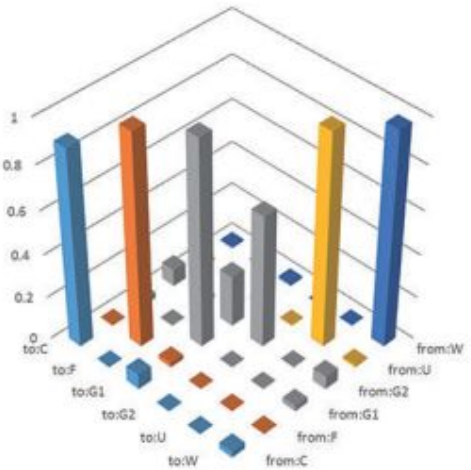
（図３　台風の進路と Tweet の動き解析）

また（成果１）では、同様のソーシャルデータである Flickr についてそのキーワードやジオタグ等を解析して地物を検出する手法を考案し、例えば海岸線において実際の海岸線と統合したところ、64~82%の正しさで検出できていることが示された。Flickr などは与えるメタデータが十分な情報がない場合があるが、この場合のクラスタリングの方法を（成果 11）で提案している。

応用事例の一つとして土地利用解析を取り上げており、（成果３）では Degree Confluence Project と呼ばれる校正プロジェクトにおける写真の画像解析による土地利用パターンの抽出を試みた。

（図４　土地利用を撮影した写真）

これらを教師データとして画像解析を行ったところ、基本的な 6 分類において図 5 のような結果を得て、Kappa と呼ばれる評価指標も 0.924 と高い結果となった。衛星画像解析による土地利用解析よりも高精度な結論を得た。



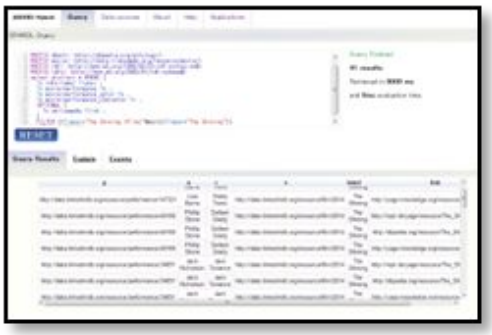
（図５　6 分類（実質は 5 分類）における評価結果。対角線上にないデータが誤り）

さらに、これをジオタグ付きの写真やキーワードを使って解析する手法と組み合わせることで、より精度の高いデータ統合が可能なことが示された。

その他にも、Web 情報と衛星データを使った建築物発見等のホットスポット解析を研究した。

② 検索・解析基盤：

検索においては、特に分散した Linked Open Data など RDF のデータベース処理についての研究成果を得た(成果９)。まず、膨大なインターネット上の SPARQL エンドポイントの分散検索システムについて、データが膨大になることから分散検索が現実的でないことから、時間制限下でベストエフォートな検索を行う手法とシステムを構築した(Aderis-Hybrid)。これは、問い合わせ処理を動的に行う（プランの動的な変更）ことで時間内にできるだけ多くの答えを得ると共に、その段階での答え全体の予測とその確度を予想して示す点で特徴がある。



（図 6　Aderis-Hybrid の処理画面）

同時に、こうして検索された RDF の集合を解析するための ETL(Data Extraction、Transformation, Loading)フレームワークを提案し（成果５)、あわせて Linked Open Data の検索と解析のフレームワークとできた。
また、ダイジェストデータを用いた結合演算の効率化については、RDF の結合が低選択率であることに注目した効率化手法を考案した（成果６）

解析基盤との連携については、Hadoop 上のデータベース基盤である Hive に機械学習の処理を組み合わせ、あわせてスケーラビリティを実現する Hivemall と呼ばれるシステムを構築した(成果２)。これは、User Defined Table Function と呼ばれる機能を活用することでデータ量に対する高いスケーラビリティを実現することができた。

集合的なデータにおける矛盾の現れとして「外れ値」を考え、この計算の高速化を試みた（成果４)。同時に、GPGPU による高速化を行い（成果１０)、最大２桁の高速化に成功した。

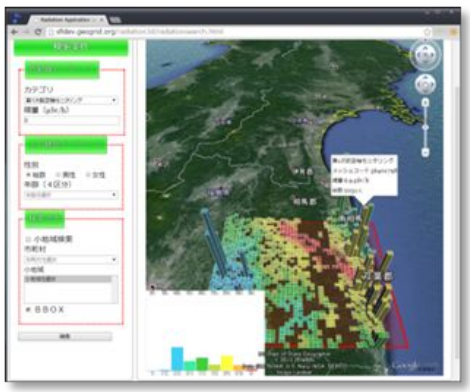
③ 統合応用サービス：要素技術のいくつかは図で示すようにプロトタイプシステムとして実現したが、特に応用利用者にサービス提供して実証するために、環境モニタリングデータの統合環境をクラウド上に構築した。

これは、福島原子力発電所の事故以来各組織が広く測定、公開している多様な放射線モニタリングデータベースを統合するものであり、以下のような問題が考えられた。

1. 各組織や測定等でボトムアップ的にデータが作成されたため、統合で矛盾がじじる可能性が高い。
2. データの不均質さに伴う矛盾の発生：年数回の離散的な測定から１０分に一回と言ったリアルタイムに近いデータ生成といった時間的なばらつきや、面的な測定や数点の測定ポイント、移動体による測定と言った空間的なばらつきがあり、統合が容易でないのと同時に、測定条件の違いなどからの矛盾が発生しうる。

これらを吸収するためのプロトタイプを産総研の仮想クラスタ上に構築し、クラウド的にサービス提供することを試みた。

これを基礎に、データ統合として環境モニタリングデータと、政府統計など外部のデータを統合する場合の問題に注目し、これらを共に Linked Open Data/RDF として実現することで統合的に連携することを試みた。応用例として集団線量を測定するサービスを構築して提供し、これは LOD チャレンジ 2014 において環境 LOD 賞を受賞している(成果８）
なお本サービスの実現にあたっては、原子力規制庁および日本原子力研究所の協力と支援を得ている。



(図 7　放射線 LOD システムの操作画面)

まとめ：様々な「矛盾」を扱う要素技術と統合システムの構築を試みた。いくつかの応用で十分な結果が得られたものの、矛盾そのものは必ずしも十分な解決が得られたとは言えず、今後の課題と言える。一方、時空間基盤とその上での高度解析、可視化はビッグデータの基礎技術として極めて有効であり、人工知能など高度な解析や、Cyber-Physical System など、高度な時空間データ基盤への発展が期待できる。

５．主な発表論文等

〔雑誌論文〕（計 23 件）

1. 大森正巳、廣田雅春、石川博、横山昌平、”ソーシャルメディア上から収集したジオタグに基づく地理的特徴の抽出と評価“、情報処理学会論文誌データベース、Vol.8,No.1,1pp1-15,2015
2. 油井誠、小島功、“Apache Hive を用いたスケーラブルな機械学習機構の構築”、情報処理学会論文誌データベース,Vol8.No.1,pp73-87,2015.
3. 尾崎竜史、岩田健司、岩男弘毅、小島功、”風景画像データのためのカテゴリー推定“,精密工学会誌、Vol.80, No.12, pp1189-1193,2014.
4. S.Shaikh,H.Kitagawa,”Top-k Outlier Detection from Uncertain Data”,InternationalJournal of Automation and Computing, Vol.11. pp 128-142,2014.
5. 井上寛之、天笠俊之、北川博之、“LOD の OLAP 分析を可能にする ETL フレームワークの提案”日本データベース学会論文誌 Vol.12, No.1,pp79-84,2013.06

〔学会発表〕（計 69 件）

6. A.Matono, H.Ogawa and I.Kojima, Improvement of Join Algorithms for Low-Selectivity Joins on Mapreduce” ,The 26th Australasian Database Conference, 2015.06.
7. K.Kim, H.Ogawa, A.Nakamura and I.Kojima,”Sophy; A Morpohological Framwork for Structuring Geo referenced Social Media”,ACM LBSN workshop, 2014.11.
8. I.Kojima,Y.Tanaka,A.Matono and A.Nakamura, ”Implementation of the Fukushima Radiation LOD Framework”, W3C Linking Geospatial Data Workshop, 2014.03.
9. S.Lynden, I.Kojima, A.Nakamura and M.Yui, “A Hbrid Approach to Linkd Data Query Processing with Time Constraints”, WWW Linked Data on the Web Workshop, 2013.04.
10. Y.Kozawa,T.Amagasa and H.Kitagawa, “GPU acceleration of probabilistic frequent itemset mining from uncertain databases”, 21st ACM CIKM, 2012.10
11. M.Hirota, N.FUjikta, S.Yokoyama and H.Ishikawa, “A Robust Clustering Method for Missing Metadata in Image Search Results”, Journal of Information Processing, Vol.20, No.3, pp537-547, 2012.07.

〔図書〕（計 1 件）

〔産業財産権〕なし

〔その他〕 なし

６．研究組織

(1)研究代表者

小島 功（KOJIMA ISAO）
独立行政法人 産業技術総合研究所
情報技術研究部門 総括研究主幹
研究者番号：00356982

(2)研究分担者

北川 博之(Kitagawa Hiroyuki)
筑波大学
システム情報工学研究科（系）教授
研究者番号: 00204876

石川 博(Ishikawa Hiroshi)
首都大学東京
システムデザイン学部 教授
研究者番号：60326014

天笠 俊之(Amagasa Toshiyuki)
筑波大学
システム情報工学研究科（系）准教授
研究者番号：70314531

横山 昌平(Yokoyama Shohei)
静岡大学
情報学研究科 講師
研究者番号：20443236

川島 英之
筑波大学
システム情報工学研究科（系）講師
研究者番号：90407148

的野 晃整（Matono Akiyoshi）
独立行政法人 産業技術総合研究所
情報技術研究部門 研究員
研究者番号： 10443227

Steven Lynden（ Steven Lynden ）
独立行政法人 産業技術総合研究所
情報技術研究部門 研究員
研究者番号： 30528279

油井 誠（Yui Makoto）
独立行政法人 産業技術総合研究所
情報技術研究部門 研究員
研究者番号： 10586712

金 京淑（Kim Kyoungsook）
独立行政法人 産業技術総合研究所
情報技術研究部門 研究員
研究者番号： 20738728

岩田 健司（Iwata Kenji ）
独立行政法人 産業技術総合研究所
情報技術研究部門 研究員
研究者番号：80549890

中村 章人（Nakamura Akihito）
独立行政法人 産業技術総合研究所
情報技術研究部門 主任研究員
研究者番号： 70357664

(3)連携研究者 なし