様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ８ 日現在

機関番号：１７１０２
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ１１３１５

研究課題名（和文）問合せに着目したデータの理解支援に関する研究

研究課題名（英文）Research on data understanding support through queries

研究代表者
清水　敏之（SHIMIZU, Toshiyuki）

九州大学・附属図書館・准教授

研究者番号：６０４０２４６８

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：データ管理者がデータに対する理解を深めつつ効率よくデータ管理を行うことを考え、具体的にはデータクリーニングとデータ共有に関する研究に取り組んだ。データクリーニングに関する研究としては、データ中の不整合な値の候補を検出し、不整合な値の候補を含む部分データを、問合せを利用したビューを用いてデータ管理者に提示することを考え、そのための不整合候補検出手法およびビュー提示手法について提案した。データ共有に関する研究としては、複数のデータ管理主体が自律分散的にデータを管理している状況において、問合せによって表現される部分データを適切に共有・更新するための仕組みを考案し、その実装方法についても提案した。

研究成果の学術的意義や社会的意義
近年、機械学習を用いたデータ分析が盛んに行われているが、適切な分析のためにはデータを整備し、理解して手法を適用することが重要だと思われる。分散的にデータが生成され、管理されている場合など、一人のデータ管理者が関連するデータの全体像を把握するのが困難な状況も多いと思われる。本研究では、問合せを用いて部分データを切り出して扱うことで、データ管理者によるデータの理解を助けながらデータ管理を行うことを考え、そのための手法の提案と実装を行った。

研究成果の概要（英文）：We considered how data managers can efficiently manage data while deepening their understanding of the data, and specifically worked on research related to data cleaning and data sharing. As for the study on data cleaning, we proposed a method for detecting candidates of inconsistent values, and also a method for showing partial data containing the candidates of inconsistent values to data managers using a view-based approach using queries. As for the study on data sharing, we devised a mechanism for appropriately sharing and updating partial data expressed by queries in a situation where multiple data management entities are managing data in an autonomous and decentralized manner, and proposed an implementation method for this mechanism.

研究分野：データベース

キーワード：データベース　問合せ　データクリーニング　データ共有

１．研究開始当初の背景

データの大容量化・多様化および計算機資源の発達により、大規模データの分析には強い需要がある。しかし、大規模なデータではデータ保有者自身であってもデータの内容を把握することが困難な場合があり、高度なデータ利用のためには、単純な検索技術だけでなく、検索補助、利用補助、理解補助などのための技術が重要になっている。

２．研究の目的

データの理解支援に際し、問合せを用いることで、利用者が気付きにくいデータに関する知見を利用者が解釈可能な形で得ることができると考えた。データ利用者が着目する部分データを問合せで表現して扱うことでデータに対する理解を深めつつ効率よくデータ管理を行う。

３．研究の方法

主な対象データとして科学データに対するメタデータ（科学メタデータ）を実際のデータとして想定して研究を遂行した。具体的な実データを想定することで、実際のニーズに基づく議論を行うことができた。研究の進め方としては、共同研究者や指導学生と定期的に打合せを行って議論を進め、研究会での発表を行うことで研究成果の完成度を高めてきた。

４．研究成果

データ管理者がデータに対する理解を深めつつ効率よくデータ管理を行うための研究を実施し、具体的にはデータクリーニングとデータ共有に関する研究に取り組んだ。

(1) データクリーニングに関する研究

データ分析のために機械学習の利用が盛んに行われているが、適切な分析のためには空値の補填や値に一貫性をもたせるなどのデータ整備が重要となる。本研究で実データとして想定した科学メタデータは専門性が高い語を多く含み、複数人が自由記述形式で入力を行う場合もあるため、表記ゆれや誤記、記入漏れが多いデータになる傾向があるが、専門性の高いデータでは機械的な修正が困難であり、人が確認して修正する必要がある。しかし、大量のデータを全て閲覧するのは現実的ではないため、不整合な値や不適切な値を含む部分を問合せで表現し、関係データベースにおけるビューとして着目すべき部分を切り出して提示することで、データ管理者による修正を補助する仕組みを提案した。部分データを問合せで表現することでデータに対する理解を深めながら修正作業を行うことができると考えられる。地球科学分野のデータセットに対する実際の科学メタデータを観察し、実例に基づいた有用事例の議論を行った。

図１はデータ統合・解析システム DIAS (Data Integration and Analysis System)で管理されている地球科学分野のデータに対するメタデータの一部を用い、ビューで部分データを切り出して観察した例である。「著者名=Hiromichi Igarashi」の問合せ条件によりビューを作成した結果、所属機関や作成機関の項目で「DrC/JAMSTEC」と「JAMSTEC/DrC」などの表記の異なる記述が散見された。このように何らかの条件で部分データを切り出すことで、データの把握と修正が容易になると思われる。

著者名	所属機関	作成者名	作成機関
Hiromichi Igarashi	JAMSTEC/DrC	Sugiura, Nozomi, Dr.	**JAMSTEC/DRC**
Hiromichi Igarashi	JAMSTEC/DrC	Dr. Nozomi Sugiura	**JAMSTEC/DRC**
Hiromichi Igarashi	**Japan Agency for Marine-Earth...**	Kazuo Umezawa	Japan Aerospace Exploration...
Hiromichi Igarashi	**Japan Agency for Marine-Earth...**	Kazuo Umezawa	Japan Aerospace Exploration...
Hiromichi Igarashi	**Japan Agency for Marine-Earth...**	Kazuo Umezawa	Japan Aerospace Exploration...
Hiromichi Igarashi	**Japan Agency for Marine-Earth...**	Kazuo Umezawa	Japan Aerospace Exploration...
Hiromichi Igarashi	**Japan Agency for Marine-Earth...**	Kazuo Umezawa	Japan Aerospace Exploration...
Hiromichi Igarashi	JAMSTEC/DrC	Hiromichi Igarashi	JAMSTEC/DrC
Hiromichi Igarashi	JAMSTEC/DrC	Hiroshi Kawamura	Center for Atmospheric and...
Hiromichi Igarashi	JAMSTEC/DrC	Sugiura, Nozomi, Dr.	**JAMSTEC/DRC**
Hiromichi Igarashi	**DrC/JAMSTEC**	Remote Sensing Systems	NULL
...	...	...	...

図 1　ビューの実例

不整合な値を含む部分データを取得するにあたり、まずは、関係データ中の不整合な値の候補を得ることが重要であると考え、エンティティ解決手法を応用した不整合検出手法を提案した。本研究で実データとして想定した科学メタデータにおける不整合は、データの専門性の高さやデータ量の少なさの問題から機械学習に基づく判定が困難であり、ルールベース手法の考え方

に基づきつつも単語の分散表現を用いることで柔軟な判定が可能になると考えた。単語の分散表現を用いた既存のエンティティ解決手法を応用し、着目した値の組合せが同一のエンティティと予測されるが、実際の値は異なる場合を不整合候補として検出することを考えた。

実際の科学メタデータを対象として提案した不整合検出手法を適用し、不整合な値の候補が抽出できることを確認した。評価では既存のデータクリーニング手法と比較して再現率の向上を確認した。再現率の向上は、専門性の高い科学メタデータなどに対して、データ管理者の確認を踏まえて値の修正を行う枠組みを考えている本研究において重要であると考えている。

さらに、検出された不整合候補を利用し、ビュー提示を行う具体的な手法について研究を推進し、検出された不整合候補から類似した不整合をグルーピングし、トピックに基づいてタプルを分類することでデータ管理者に提示するビューを作成する手法を考案した。提案した手法により、精度を大きく悪化させることなくビュー提示が可能になることを確認した。

(2) データ共有に関する研究

近年、膨大なデータが分散的に生成されており、関連するデータの全容を把握することが困難になっている。組織や事業によってデータが収集され、それぞれのデータベースによって管理されている場合があるが、類似するデータでも管理主体（以後、ピアと呼ぶ）によって異なる形式で管理されている場合が多いと思われる。関連するデータが複数のデータベースに分散しており、複数のデータベース間でやり取りしながらデータの管理を行う状況を想定してデータ共有に関する研究を推進した。

科学メタデータの共有を行う際、利用規約の問題や管理上の問題から自身が管理するメタデータの中で特定のメタデータのみ共有したい（特定のメタデータのみを輸出したい）場合や、自身のデータベースの特性を考慮して特定のメタデータや特定の相手からのみメタデータを輸入したい場合がある。データを理解し、適切な管理を行うために、データ共有の際に条件を設定して輸出入するデータをフィルタリングできることが望ましいと考え、問合せによって表現される部分データを適切に共有・更新するための仕組みが必要だと考えた。科学メタデータのような専門性の高いデータでは、データの追加・削除や、値の更新が行われた際に、その更新を他のピアに輸出すべきか、また、他のピアから輸入すべきか判断が難しい場合があるため、各ピアがそれぞれのポリシーに基づいて更新を受け入れるかどうかを定義できるべきだと考えた。

このような状況に対して適切かつ柔軟に対応するために、自律分散型データ共有・更新システムである SKY を考案した。SKY の構成イメージを図 2 に示す。図 2 では P1、P2、P3 の三つのピアがデータ共有を行っている状況を示している。さらに、各ピアは複数の共有に参加する場合もあり、P2 は P4 とも別のデータ共有を行っている。SKY では共有に参加するピアが共通してアクセス可能なピアとして共有リポジトリ SR (Shared Repository)を導入し、各ピアが共有したいデータは SR 中の共有テーブル ST (Shared Table)に格納される。各ピアが管理する情報は基礎テーブル BT (Base Table) に格納されており、各ピアは制御テーブル CT (Control Table)を通して輸出したいデータおよび輸入したいデータを制御する。

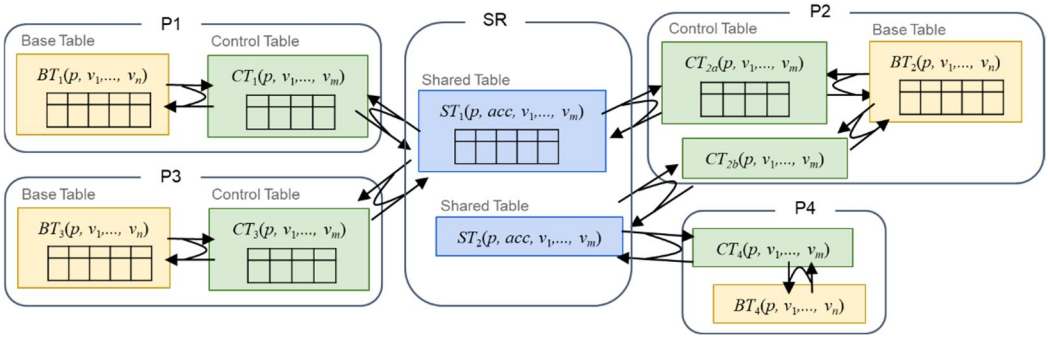

図 2　SKY の構成

提案した SKY の実装方法として、双方向変換を利用したデータ共有アーキテクチャである Dejima を用いることを考え、プロトタイプシステムの開発を行った。Dejima は Dejima テーブルの単純同期によってデータ共有を行う仕組みであるが、図 3 のように、SR も一種の特殊なピアであると考え、CT が Dejima テーブルを用いて実装可能であることを示した。

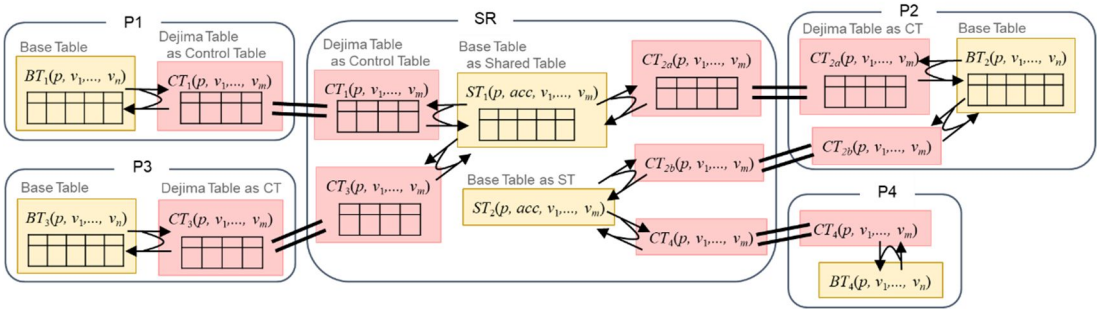

図 3　Dejima を用いた SKY の構成

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 中林 和也，清水 敏之，大手 信人
２．発表標題 科学メタデータにおけるデータクリーニングのための不整合検出の効率化
３．学会等名 第14回データ工学と情報マネジメントに関するフォーラム
４．発表年 2022年

１．発表者名 清水 敏之，加藤 弘之，吉川 正俊
２．発表標題 Dejimaを用いた自律分散型データ共有・更新システムの実装
３．学会等名 第14回データ工学と情報マネジメントに関するフォーラム
４．発表年 2022年

１．発表者名 大森 弘樹，清水 敏之，吉川 正俊
２．発表標題 制約脆弱データに対するデータクリーニングのための不整合候補検出
３．学会等名 第13回データ工学と情報マネジメントに関するフォーラム
４．発表年 2021年

１．発表者名 沖野 雄哉，若林 勇弥，清水 敏之，加藤 弘之，吉川 正俊
２．発表標題 自律分散型データ共有更新システムにおける共有ポリシー実装のためのアーキテクチャの検討
３．学会等名 第13回データ工学と情報マネジメントに関するフォーラム
４．発表年 2021年

１．発表者名
Toshiyuki Shimizu, Hiroki Omori, and Masatoshi Yoshikawa

２．発表標題
Toward a view-based data cleaning architecture

３．学会等名
Third Workshop on Software Foundations for Data Interoperability（国際学会）

４．発表年
2019年

１．発表者名
大森 弘樹，清水 敏之，吉川 正俊

２．発表標題
エンティティ解決手法を応用したデータクリーニングのための不整合検出

３．学会等名
第12回データ工学と情報マネジメントに関するフォーラム

４．発表年
2020年

１．発表者名
大森 弘樹，清水 敏之，吉川 正俊

２．発表標題
ビューに基づくデータクリーニング方式の提案

３．学会等名
第11回データ工学と情報マネジメントに関するフォーラム

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関