Vol.2014-DPS-161 No.13
Vol.2014-EIP-65 No.13
2014/9/19

異種情報融合フレームワークの提案

鈴木吉輝[†1]　澤本潤[†1]　瀬川典久[†1]　杉野栄二[†1]
佐藤裕幸[†1]　和田雄次[†2]

ユビキタスコンピューティング社会では，コンピュータ，スマートフォンはじめ様々な機器から大量のデータが収集される．今日では，これらのデータを集約し，知識や傾向を発見し意思決定などに役立てることが重要になっている．しかし，収集されたデータは多種多様なデータベース上に置かれており，情報の収集には膨大な時間を要する．本稿では，異種データベースの統合的な情報の抽出及び異種情報群同士の結合機能を提供することで，あたかも一つの仮想データベースから情報を抽出する異種情報融合フレームワークを実現する．

Proposal of a Framework for Integrating Multiple Heterogeneous Databases

YOSHIKI SUZUKI[†1]　JUN SAWAMOTO[†1]　NORIHISA SEGAWA[†1]
EIJI SUGINO[†1]　HIROYUKI SATO[†1]　YUJI WADA[†2]

In ubiquitous society, a lot of data is acquired from various devices such as computers, smart phones, etc. Today, it is becoming extremely important to integrate these data, to discover knowledge and social trend and to utilize them for decision-making. However, this kind of data is placed on a variety of distributed databases, and the acquisition of information takes huge time for the user. In this paper, a method for the uniform acquisition of the information from heterogeneous databases and the extended join procedure of data of heterogeneous databases are proposed. This realizes a framework for integrating multiple heterogeneous databases exactly like extracting data from a huge single virtual database.

1. はじめに

ユビキタスコンピューティング社会では大量のデータが収集利用されるが，こうしたデータの中に隠された知識や傾向を発見し意思決定や新サービス開発などに役立てることが重要となっている．こうしたデータベース（以下 DB）はいわゆるビッグデータと呼ばれ，データマイニング技術への利用拡大が期待されている．しかしながら，ユビキタスコンピューティング社会の発達とともに情報の多種多様化への対応と利用目的の多様化に伴い，世の中にはリレーショナルデータベース（RDB）だけではなく，新しい概念を実現した XML DB，NoSQL DB や Linked Open Data(以下 LOD)のような，オープンデータ同士を関連づけ，人間が Web サイトをハイパーリンクで巡るように，コンピュータも相互に関連付けられたオープンデータを辿ることができるアーキテクチャを利用した情報群など，多種多様な DB が存在し分散配置されている．情報分析者は，本来的にはルール発見・解釈の作業や新サービス開発の作業に集中したいにも関わらず，準備過程であるデータ収集作業に膨大な時間を割かねばならず，こうした作業負担を軽減する技術が必要となってきた．我々の研究は，種々の DB 視点を利用者に提供する機能を研究し，利用目的に応じて DB 内データの取捨選択や，それぞれの DB からデータマイニングした結果から得られる複数のルールを調停･統合する「知識融合技術」を実現することである．本稿では，データの解析や融合情報を用いたサービス開発を行う技術者のために MySQL，MongoDB，LOD を擬似融合し，あたかも一つの仮想データベースから情報を抽出する環境を実現する異種情報融合フレームワークの提案を行う．

2. 異種情報利用における問題点

異種情報同士を結合し新しい結果を示す研究や[1]，近年 LOD 同士を繋げて新しい情報を提供する試みが行われている[2].

しかし，RDB，NoSQL や LOD などの複数の異なる DB を融合して利用する場合，表 1 のように，各情報にアクセスするための問い合わせクエリや，データモデルに差異がある．利用者は問い合わせクエリの学習と，出力結果の差異の吸収に多くの時間を割く必要がある．

表 1 問い合わせクエリの違い
Table 1 Different query languages.

DB	MySQL	MongoDB	LOD
問い合わせクエリ	SQL 文	Mongo クエリ	SPARQL

例として，以下のような表 2 から点数と氏名を MySQL，MongoDB，SPARQL 及び XML それぞれから取り出す場合

†1 岩手県立大学大学院
Iwate Prefectural University Graduate School
†2 東京電機大学
Tokyo Denki University

の問い合わせクエリの差異は表 3 のようになる.

表 2 成績データ
Table 2 Score data.

name	gender	score
山田太郎	男	55
鈴木太一	男	80
吉田香織	女	78

表 3 問い合わせクエリの差異
Table 3 Different query sequences.

MySQL	SELECT name, score FROM 成績データ
MongoDB	成績データ.find({},{_id:0, name:1, score:1})
SPARQL	PREFIX score: 成績データエンドポイント SELECT ?name, ?score WHERE {?s score:name ?name . ?s score:score ?score}

以上のように，それぞれに対応するクエリに応じて結果が出力されるが，上述のように問い合わせクエリが異なることと，出力結果についても，個々のデータベースで独立しており出力形式も様々である.

3. 関連研究

仮想的に異種データベースへの統合検索を行う技術は, 連邦データベース[3]，メディエータ技術[4]，クエリ統合技術[5]などがある．特に XML メディエータ技術に関しては，今日様々な研究が行われており, XML メディエータを用いて仮想的に XML 上の異種情報群への問い合わせの最適化を行う研究[6]や，異種データベース間で異なるデータ表現方法の異種性の解消[7]などがある．これらの研究では，データの表現方法を柔軟性の面から XML としている. また，XML 以外の表記方法としてグラフ探索を用いる研究も存在している[8]．本研究では，データの表現方法を XML から冗長性を省きデータ量の削減に期待できる JSON 形式とし，従来のリレーショナルデータベースに加え MongoDB や SPARQL のクエリの統一を行う.

4. 先行研究

我々は異種情報融合技術の実現のために，分散データベースを P2P ネットワーク上で融合し単一のデータベースとしてユーザに提供する研究[9]を行った．また，RDB と XMLDB の構造を共通の XML に変換し，XQuery を用いて問い合わせを行えるようにすることで，利用者に対して共通のクエリで双方のデータベースを扱うことのできるスキーマ統合型の仮想化技術の研究[10]を行った.

これら先行研究の問題点としては，前述の P2P による分散データベースの融合は，利用者に対して提示されるデータベースは，一つのデータベースの見え方のみを提供するため，個々のデータベースに対する柔軟な見え方を提供することができない. 後述の XML に変換し, 融合の後 XQuery による問い合わせクエリの統一方法は，前述の問題と同じ問題点である，複眼的なデータベースの扱いができないことと, XML に変換するため全てのデータを分散データベースから取り出す必要がある。これは情報量が増加したときに XML データを保持する場所の確保が問題になる可能性がある.

5. 異種情報融合フレームワーク

5.1 概要

今回開発した異種情報融合フレームワークとは，前述の問題を解決するために様々な異種 DB から情報を抽出する際の統一した問い合わせクエリを提供し，本フレームワーク内で，統一クエリを各 DB の問い合わせクエリに変換することで，ユーザからはあたかも 1 つのデータベースで管理されているように見えるようにするものである．対応 DB としては，MySQL，MongoDB，LOD とした.

出力結果としては，個々のデータベースの結果を表示または，複数データベースの情報を仮想的に融合した結果の出力機能を提供することで，利用者に対して個々のデータとして扱うものと結合して扱いたいものを，柔軟に分けて対応することができる.

本フレームワークは，異種情報を融合したアプリケーションや，情報を解析するユーザ向けであり，データベースに精通したデータベースエンジニア向けではない．そのため，本フレームワークが提供するのはデータの呼び出しのみを提供する. 実際の運用例を図 1 に記載する. 利用者は，異種情報を利用するアプリケーションや情報解析に必要な異種データベース上の情報の取得に本フレームワークを介して行う.

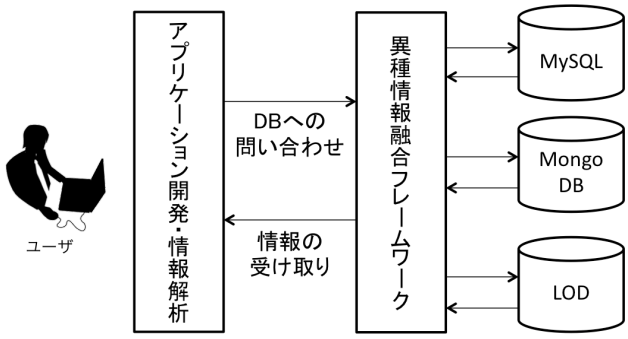

図 1 本フレームワークの運用例
Figure 1 Example usage of the proposed framework.

5.2 システム構成

異種情報融合フレームワークの構成は図 2 に記載する. 利用者は，アプリケーション上やターミナル上で統一クエ

リを発行する．フレームワーク内では，融合データベース情報の管理ファイルから登録されている DB の種類を抽出し, 種類に応じて各クエリへの変換を行う. クエリ変換後，融合データベース情報の管理ファイルから各データベースへのアクセス情報を取得し，変換したクエリを用いて各データベースに並列にアクセスを行う．利用者が，データの結合を行う場合，結合するデータ群同士を結合させた結果を JSON 形式に変換して出力し，結合しない場合は，各データベースから返ってきたデータを JSON 形式に変換して出力する．

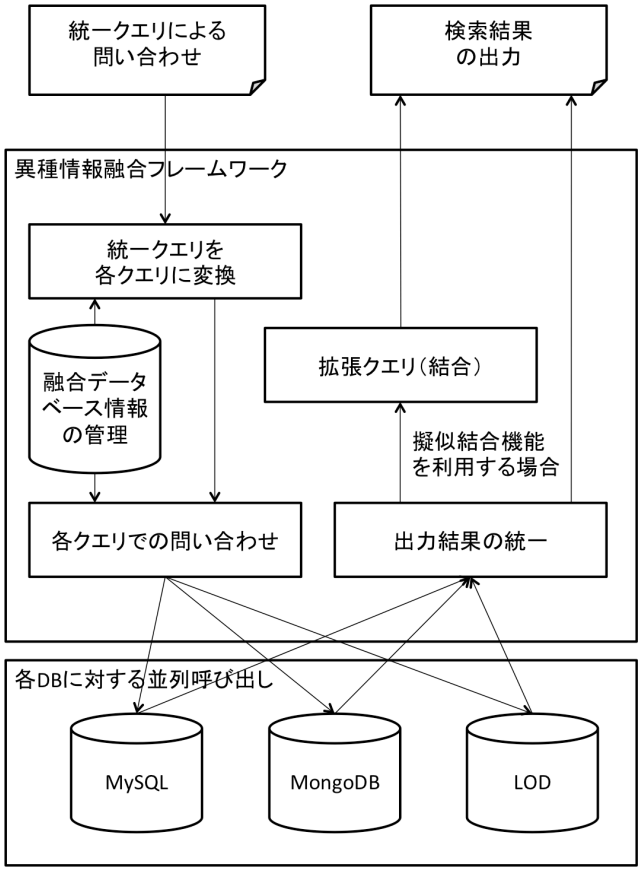

図 2 異種情報融合フレームワークの構成
Figure 2 Configuration of the framework for integrating multiple heterogeneous databases.

5.3 データベース情報の管理

本フレームワークでは，サーバ設定ファイルで各データベースの情報を管理しており，利用者が統一クエリによる問い合わせを行った際に管理されているデータベースに対して呼び出しを行う．問い合わせには表 4 に書かれている項目を必要に応じて記述する必要がある．

5.4 統一クエリの提供

様々な情報源から情報を抽出する際に，問い合わせクエリが異なり個々の問い合わせクエリの学習と記述量の増加してしまう．本フレームワークでは，１つの問い合わせクエリで，異なる情報源から情報を抽出することができる統一クエリを提供する．統一クエリの候補としては，SQL，Mongo，SPARQL，独自クエリなどが考えられる．今回は文献の多い SQL の SELECT 文に準拠しつつ, 他の DB の差異を埋めるために独自クエリを構築した．

5.4.1 各データベースの差異の吸収

MySQL，MongoDB，LOD のクエリの差異を吸収するために，本フレームワーク内で統一クエリとの対応付けを表 6 のように行っている． SPARQL においては，書き方が柔軟であるため，主語，述語，目的語の抽出対象が入れ替わることがあるが，今回は他のデータベースとの対応の関係上，目的語の抽出のみの対応とした．

表 4 融合データベース情報管理ファイルの項目
Table 4 Items in the integrated database management file.

引数	説明	対応 DB
[DB 名]	統一クエリで利用する DB 名	MySQL,MongoDB,SPARQL
type	DB の種類	MySQL,MongoDB,SPARQL
address	DB アドレス	MySQL,MongoDB
port	DB ポート番号	MySQL,MongoDB
db	DB 名	MySQL,MongoDB
user	DB ユーザ名	MySQL,MongoDB
password	DB パスワード	MySQL,MongoDB
table	テーブル名	MySQL
collenction	コレクション名	MongoDB
endpoint	SPARQL エンドポイント	SPARQL
subject	SPARQL 主語	SPARQL

実際の記述例を表 5 に記す．この例では，[db1]に mysql，[db2]に mongoDB，[db4]に LOD を設定し，問い合わせに必要なデータベースの情報を記述している．

表 5 融合データベース情報管理ファイルの記述例
Table 5 Examples of the description of integrated database management file.

[db1] type : mysql address : 127.0.0.1 port: db : lab user : root password : root table : score [db2] type : mongo address : 127.0.0.1	port : 27017 db : labdb user : root password : root collection : users [db4] type : sparql endpoint : http://ja.dbpedia.org/sparql subject :<http://ja.dbpedia.org/resource/東京都>

表 6 統一クエリと各問合せクエリの対応付け

Table 4 Mapping from integrated query to each query.

統一クエリ	MySQL	MongoDB	SPARQL
テーブル	テーブル	コレクション	主語
レコード	レコード	ドキュメント	述語
カラム	カラム	フィールド	目的語

統一クエリの構文解析図は図 3 のようになっている．

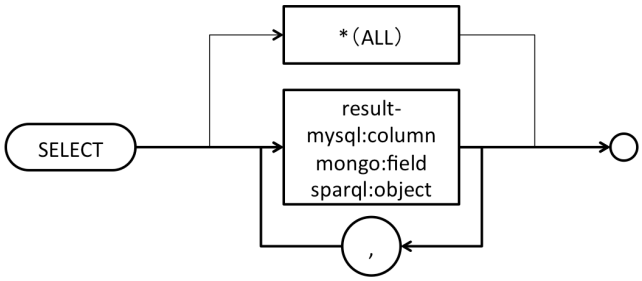

図 3 統一クエリ構文解析図

Figure 3 Syntax diagram for integrated query.

現在実装している機能としては，融合データベース情報の管理ファイルに登録されているデータベースの情報の全件取得，特定列名指定による取得機能である．SELECT 文が呼ばれた時の構文解析の手順は以下のようになる．

1. SELECT の次の区がアスタリスク（*）の場合，全件取得として扱う．その後３．を実行する．
2. SELECT の次の区がアスタリスク（*）でない場合，取得する列名として扱う. 例:SELECT name の場合，各データベースから name の列を取得するために抽出列名リストに追加する．

2.1. ２．の次の区が，カンマ（,）の場合，複数列名取得となり，カンマの後の区を列名として扱う．例：SELECT name, address の場合，各データベースから name と address の列を取得し，抽出列名リストに格納する．

2.2. 列名の後ろにカンマ（,）が存在する場合 2.1.を繰り返し，実行する．

3. 構文解析が終了場合は，各クエリへの関連付けを行う．

5.4.2 各クエリとの関連付け

4.4.1 で統一クエリを解析した結果得られた列名は表 7 のように各クエリと関連付けられる．その際に，MySQL であれば融合データベース情報管理ファイルに記述されている table 名，MongoDB の collection 名，SPARQL の subject 名に対して問い合わせを行う．

表 7 各クエリと列名の関連付け

Table 7 Relation of column names and each query.

MySQL	SELECT [列名 1], [列名 2]… FROM [table]
MongoDB	[collection].find({},{_id:0, [列名 1]:1, [列名 2]:1…})
SPARQL	PREFIX [subject] SELECT ?name, ?score WHERE {?s [subject]:[列名 1] ?[列名 1] . ?s [subject]:[列名 2] ?[列名 2]…}

5.5 DB 間での結合出力

前述の統一クエリにより抽出されたデータは各情報源の個々のデータとして出力される．融合情報利用アプリケーションの開発や解析をするためには，個々の抽出した情報を結合する必要がある．しかし，結合するために必要な JOIN 文は個々の DB 内でしか行えないため，異なる情報源同士の結合は前述のクエリ統合のみでは実現することができない．本フレームワークでは，各データ同士の共通項目同士で結合し，１つのデータリストとして提供する機能を実装した．結合する際には独自実装の JOIN 文を用いる．JOIN 文では，結合元データベースの列名と，結合するデータベースの列名を指定することにより，結合結果を出力する．MySQL と MongoDB の内容が図 4 の時，式（１）のクエリを実行すると表 8 のように MySQL の生徒成績に対して，キーである gender が一致する MongoDB の全国平均点が結合された結果を出力することができる．

JOIN db1 gender db2 gender (1)

MySQL:生徒成績

name	gender	score
山田太郎	男	55
鈴木太一	男	80
吉田香織	女	78

MongoDB:全国平均点

prefecture	gender	score
愛知	男	70
愛知	女	72
青森	男	73
青森	女	75

図 4 MySQL と MongoDB の情報内容

Figure 4 Data of MySQL table and MongoDB collection.

表 8 JOIN 文出力結果

Table 8 Output result in JOIN format.

{“db”:”db1,db2”, “name”:” 山 田 太 郎 ”, “gender”:” 男 ”, “score”:55, “join1”:{“prefecture”:” 愛 知 ”, “gender”:” 男 ”, “score”:”70”}, “join2”:{ “prefecture”:”青森”, “gender”:”男”, “score”:”73”}…}
{“db”:”db1,db2”, “name”:” 鈴 木 太 一 ”, “gender”:” 男 ”, “score”:80, “join1”:{“prefecture”:” 愛 知 ”, “gender”:” 男 ”, “score”:”70”}, “join2”:{ “prefecture”:”青森”, “gender”:”男”, “score”:”73”}…}
{“db”:”db1,db2”, “name”:” 吉 田 香 織 ”, “gender”:” 男 ”, “score”:78, “join1”:{“prefecture”:” 愛 知 ”, “gender”:” 女 ”, “score”:”72”}, “join2”:{ “prefecture”:”青森”, “gender”:”女”, “score”:”75”}…}

6. 評価と応用

実験内容に対応した評価の基準については以下の通りである. 評価の基準対象として, MySQL, MongoDB, SPARQL の個々のデータベースからそれぞれにアクセスした場合との比較を行う.

- 本フレームワークを利用した場合の学習量の比較
- コード記述量の比較
- 独自実装の JOIN 文 MySQL の JOIN を利用した際のトラフィックの計測

また，本フレームワークを活用した応用アプリケーションを構築し，本フレームワークの有用性を確認する予定である.

7. まとめ

本稿では，異種情報源から情報を取得しアプリケーション開発や情報解析をするための準備過程である情報抽出部分の作業量削減のために，統一クエリによる MySQL，MongoDB，SPARQL に対する情報取得機能と，異種情報源同士の結合機能をもつ異種情報融合フレームワークの提案を行った．これにより，データマイニング技術の研究や知識融合アプリケーション開発の準備過程である，異種情報群からの情報抽出の負荷を軽減することができる.今後は，統一クエリと結合機能の有用性の評価を行っていく．その評価を踏まえ，本フレームワークの拡張と改良を行う．さらに，本フレームワークを用いた応用システムを考え，実際に運用実験を行っていくことによりその有用性の検証も進めていく予定である.

謝辞

本研究は JSPS 科研費 24500122 の助成を受けたものです.

参考文献

1) 佐藤大介,河本穣, 清木康, 異種のデータベース統合による複数情報源の統合的な閲覧環境の実現, DEWS2004 I-5-01
2) 森田武史, 山口高平, Linked Data を利用した情報統合, 人工知能学会誌 27(2), 189-199, 2012-03-01
3) Sheth A.P., Larson J.A. : Federated database sys-tems for managing distributed, heterogeneous, and au-tonomous databases. ACM Computing Survey, Vol. 22, No.3 , pp.183-236 (1990)
4) Wiederhold, G: Mediators in the architec-ture of future infometion systems, IEEE Com-puter, Vol.25, pp.38-49(1992)
5) Batini, C.Lenzarini, M.Navathe,S.B.:A Compara-tive analysis of methodologies for database schema inte-gration, ACM Computing Survey, Vol.18, No. 4, pp.323- 364 (1986).
6) 林孝志, 小西一也, 堀口恭太郎, 綱川光明, 鈴木源吾, 芳西崇, 異種情報源統合のための XML 問い合わせ最適化と情報源問合せ能力管理,情報処理学会論文誌 Vol.44 No.SIG12(TOD 19) pp.1-10 (2003).
7) 出口彰, 小西修, 異種情報源統合のための XML メディエータ,高知大学理学部紀要 情報科学 23, 49-55, 2002-03
8) 鈴木源吾, 鬼塚真, 榎本俊, 小林伸幸, 文制約つきグラフ探索を実現する異種データベース統合技術, 情報処理学会研究報告, Vol.2013-DPS-154 No.11, Vol.2013-CSEC-60 No.11(2013).
9) 佐々木拓也, 澤本潤, 加藤貴司, 和田雄次, P2P ネットワークにおけるコンテンツのグループ化に関する研究情報処理学会論文誌 52(2), 359-367, 2011-02-15
10) 渡辺裕太, 菖蒲佳右, 和田雄次, 澤本潤, 加藤貴司, 異種データベースの仮想化技術, 情報科学技術フォーラム講演論文集 8(2), 211-214, 2009-08-20