2N-07

知識ベースを対象にした異種データ統合

佐藤祥吾 † 天笠俊之 ‡

† 筑波大学情報学群情報科学類 ‡ 筑波大学計算科学研究センター

1 はじめに

多様な情報を表現する技術として RDF (Resource Description Framework) が注目されている. RDF は, 主語, 述語, 目的語からなるトリプルの集合で, 実世界のエンティティの間をグラフ構造として表現することができ, リンクトオープンデータ (LOD) の基盤としてさまざまな情報の蓄積・交換に利用されている. その一方で, 多くの情報が CSV, JSON, XML など 従来のフォーマットを利用して記述されており, それらを RDF と統合して利用することが期待されている。それを目的として, RML (RDF Mapping Language) が提案されている。RML によって, 多様なフォーマットのデータから RDF への変換を記述することができるが, データセット全体を一括して変換する必要がある. このため, あらかじめ利用する可能性のあるデータを全て RDF として格納しておく必要がある. この場合, すべてのデータが利用されるとは限らず、 また RDF は冗長のフォーマットであるためストレージのコストが膨大になってしまうと言う問題がある.

そこで本研究では、CSV や JSON など 多様なフォーマットのデータを RDF として透過的に利用できるフレームワークを提案する.

2 関連研究

Antonino らの手法 [1] では, SPARQL, SPIN (SPARQL Inferrencing Notation) を用いてオープンデータの統合を行う手法が提案されている. Antonino らの手法は, CSV, XML, リレーショナルデータベースといった異なるフォーマットに, それぞれ対応した異なる方法で RDF に変換し, SPARQL, SPIN を用いることで, 変換した RDF と既存の LOD を統合している.

また, 特定の分野において異種データ統合, 知識ベースの構築を行った研究として, 化石や地層をはじめとする地球科学分野では [2], 地理空間データの分野では [3] がある.

Heterogeneous data integration for knowledge bases
Shogo SATO†(s.sato@kde.cs.tsukuba.ac.jp),
Toshiyuki AMAGASA‡(amagasa@cs.tsukuba.ac.jp)
†College of Information Sciences, University of Tsukuba
‡Center for Computational Sciences, University of Tsukuba

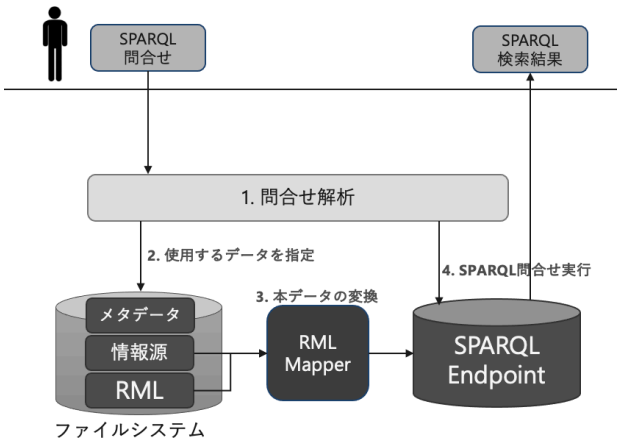

図 1: 提案手法の概要図

3 RML (RDF Mapping Language)

RML[4] は異種のデータ構造やシリアライズ形式から RDF を生成するためのマッピングルールを記述するための言語である. RML を利用することで構造データ, 半構造データから RDF を生成することができる. リレーショナルデータベース, XML, CSV, JSON, TSV に対応している.

4 提案手法

提案手法のフレームワークを, 図 1 に示す.

本システムは, 非 RDF データを SPARQL エンドポイント中の RDF データと透過的に連携できるシステムを提案する. 非 RDF データから RDF へのマッピングに RML を利用する. システム管理者は事前に, 統合対象の非 RDF データ (CSV など) を用意し，それを統合先の RDF スキーマにマップするための RML ファイルを作成しておく. また統合したいデータ自身のデータ (タイトル, 作成者, 作成日など) も RDF メタデータとして用意する.

エンドユーザは, 統合先の RDF スキーマに沿った RDF データがすでに SPARQL エンドポイント中に存在するものとして問合せを記述する. システムはそれを受けとり, 問合せ解析の結果から必要なデータ特定する. 次に, 特定されたデータを RML によって RDF

に変換し, SPARQL エンドポイントに投入する. このとき, 必ずしもデータ全体を変換する必要はなく, ユーザが与えた問合せの結果に必要な部分に限って RML による変換を行なうことで, ストレージおよび変換コストの削減を可能にする.

4.1 SPARQL 問合せの解析

ユーザから与えられる SPARQL 問合せには, 統合先の RDF スキーマを想定した問合せ条件が含まれている. 具体的には,

1. 情報源そのものに関する条件 (作者や作成日など)
2. データの内容に関する条件
3. 変換対象以外の RDF データに関する条件

など, 複数の条件が含まれる. システムはこれらを適切に分類し変換に必要な情報源 (非 RDF データ) を特定するとともに, 問合せを処理するのに貢献するデータに絞り込んで RML による変換を行なう. 情報源内の探索条件の抽出時のシステムの処理をそれぞれ説明する.

SPARQL 問合せから情報源の抽出をする際は, ユーザが記述した SPARQL 文中にある, メタデータの述語と, 情報源の内容を示す目的語を利用する. メタデータの述語をもとに情報源の内容を特定したあと, 該当する情報源はどれか, 各情報源の情報を格納したファイルシステム中のメタデータに問合せを行うことで情報源を特定する.

SPARQL 問合せから情報源内の探索条件の抽出をする際は, ユーザが記述した SPARQL 文中にある, 述語と FILTER を利用する.

問合せ条件中に出現する述語に着目すると, 情報源中のどの情報 (CSV の場合はカラム名など) が参照されるかが分かる. これには, RML 中の述語と情報源との対応関係を調べることで判別することができる. 同様に, FILTER 句にフィルタリングの条件が与えられている場合は, それを情報源に対するフィルタリング条件に変換することで, 情報源から変換すべきデータを特定することができる. FILTER 句に関しては, 条件式に出現する変数から対応する情報源を特定するとともに, 条件式を情報源のデータに対するフィルタリングの条件に変換する. さらに, その条件にマッチするレコードのみを抽出した上で RML による変換をすることで, 結果に寄与しないデータの変換を避けることができる.

4.2 ファイルシステム内のメタデータの構成

図 1 におけるファイルシステムのメタデータについて述べる. メタデータには, 情報源と RML でのマッピングルールの対応関係, 情報源の内容など, 4.1 節に利用する情報が格納されている. 本手法では, メタデータは RDF のフォーマットで保存されており, 1 つの情報源に対して, 1 つのメタデータを持っている. メタデータを RDF のフォーマットにすることで, 情報源が追加された際のメタデータの追加が容易でありまた, SPARQL でメタデータ内の検索をできるため, 目的の情報源を容易に発見することができる.

5 本システムの応用例

本研究では, 実データを使って提案手法が妥当であるか検証する. 使用する実データの分野は, 観光・レジャー分野と, 宇宙物理学分野の 2 つである. 観光・レジャー分野については, 異なる CSV ファイルに格納された統計情報などの関連情報を, SPARQL エンドポイント上の RDF データを連携・統合した利用を目指す. また, 宇宙物理学分野については, 複数の情報源から得られる異なる天体の観測情報を RDF データ上のオントロジーで統合的に利用することを可能にする.

6 結論

本稿では SPARQL 及び, RML を用いて, 多様なフォーマットのデータを RDF として透過的に利用できるフレームワークを提案した. 今後は, システムの実装に加えて, 複数情報源をまたがった問合せの最適化について検討する予定である.

参考文献

[1] Antonino Lo Bue, Alberto Machì. Open Data Integration Using SPARQL and SPIN: A Case Study for the Tourism Domain. AI*IA 2015, LNAI 9336, pp. 316–326, 2015.

[2] Chengbin Wang, Xiaogang Mab, Jianguo Chena. Ontology-driven data integration and visualization for exploring regionalgeologic time and paleontological information. Computers and Geosciences 115 (2018) 12–19.

[3] Weiming Huang, Khashayar Kazemzadeh, Ali Mansourian, Lars Harrie. Towards Knowledge-Based Geospatial Data Integration and Visualization: A Case of Visualizing Urban Bicycling Suitability. IEEE Volume: 8, May 2020.

[4] Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, Rik Van de Walle. RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. LDOW2014, April 8, 2014, Seoul, Korea