2N-2

OLAPを利用したLinked Dataの分析処理

井上寛之† 天笠俊之‡§ 北川博之‡

†筑波大学情報学群情報科学類 ‡筑波大学システム情報系情報工学域
§宇宙航空研究開発機構宇宙科学研究所宇宙科学情報解析研究系

1 はじめに

「Web of Data」として，Linked Data[1] の取り組みがW3C[1]によって推進されている．コンピュータが人間と同じように実時間で理解，共有，判断することを可能にする試みで，データの生成と相互運用を目指す概念・技術の集合である．

一方，アメリカやヨーロッパ諸国が様々な数値，統計データをWebに公開している．これらはOpen Dataと呼ばれ，Webから誰でも利用することが可能である．近年これらのデータがRDF（Resource Desciption Framework）形式で公開され，互いにリンクすることで，Linked Dataとして公開され始めた．

数値，統計データの分析には以前からOLAP（On-Line Analytical Processing）システムが広く利用されてきた．本研究ではLinked Dataから取得した数値，統計データを変換してOLAPシステムに格納し，OLAP分析を行うための手法を提案する．

2 前提知識

2.1 Linked Data

Linked Dataは構造化されたデータの公開方法の一つであり，データ同士が接続関係を持つことで，データの利用をより促進させる取り組みである．Linked Dataは標準化されたWeb技術（HTTP, URI）で構成され，Webページが人間向けの情報であるのに対し，Linked Dataはコンピュータが読み込み・理解可能なように情報を共有するものである．

Linked Dataでは情報の記述にRDFを用いる．RDFは主語（Subject），述語（Predicate），目的語（Object）の三つの要素から構成されるグラフ構造であり，リソースの識別子としてURIを利用する．主語は情報が格納されたWeb上のリソースで，目的語は主語に関する情報のプロパティもしくは特徴を定義し，目的語は述語の対象となる値を格納する．これらは三つ組でトリプルと呼ばれる．

2.2 OLAP (Online Analytical Processing)

OLAPとは，データウェアハウスに格納されたデータ対して多次元的な分析を行う手法である．分析対象となる数値データが格納された事実表と複数の次元表を対象に，ロールアップ，ドリルダウン，スライシングなどの操作を適用することによって，多様な集約処理を行なうことができる．このため，主にビジネスインテリジェンスの分野で利用されている．

3 関連研究

RDFやLinked DataをOLAPにより分析しようとする試みはこれまでにいくつか行なわれている．Benedikt ら [3] の取り組みは，QB Vocabulary [2]を利用し，RDFから多次元モデルへの変換，および一般的なOLAP操作の実現可能性を調査したものである．

これに対して本研究では，数値データを含む任意のLinked Dataで記述された数値，統計情報を対象に，既存のOLAPシステムによる分析処理を可能にする一般的な変換手法について議論している点が異なる．

4 提案手法

4.1 データセットの取得とRDBへの格納

分析対象のデータが含まれるデータセットのURIを元にデータを取得する．このとき，同一ドメイン内に限りリンクを辿り，リソースを再帰的に取得する．得られたデータは逐次関係データベース (RDB) に格納する．

データセットは，述語表アプローチ [4] を用いて格納する．これは，あるリソースについて，1) リソースのタイプ（`rdf:type`[3]）毎に独立した関係表を作成し，2) リテラルを値として持つ述語を関係表の属性とする，3) リテラル以外を参照する述語は，外部キー参照とし

Analytical Processing of Linked Data using OLAP
Hiroyuki INOUE†(inohiro@kde.cs.tsukuba.ac.jp),
Toshiyuki AMAGASA‡§(amagasa@cs.tsukuba.ac.jp) and
Hiroyuki KITAGAWA‡(kitagawa@cs.tsukuba.ac.jp)
†College of Information Sciences, University of Tsukuba
‡Faculty of Engineering, Information and Systems, University of Tsukuba
§Institute of Space and Astronautical Science, Japan Aerospace Exploration Agency

[1] World Wide Web Consortium, http://www.w3.org/

[2] http://publishing-statistical-data.googlecode.com/

[3] http://www.w3.org/1999/02/22-rdf-syntax-ns#type

表 1: ID-URI の対応表

ID	URI
1	http://www.example.com/inohiro
2	http://www.w3.org/2001/sw/

表 2: 垂直表現による RDF トリプルの格納

ID	URI_ID	Attribute	Value	リテラル
1	1	foaf:name	“井上寛之”	true
2	1	foaf:interest	“2”	false

て表現する．ただし，データを辿る段階では，全ての属性を列挙することができないため，スキーマを決定することができない．そこでタプルを ID，属性，値の組に分解して固定したスキーマに格納できる垂直表現 [2] を利用し，一度 RDB へ格納する．具体的には，主語（Subject）から得られる `rdf:type` 毎に作成する．また，RDB の中での ID とリソース（URI）の組み合わせを保存するテーブルも作成する（表 2）．

全てのリソースを参照し RDB に格納し終えると，各表の属性を把握することができるため，`rdf:type` 毎のテーブルを水平表現として作りなおす（表 3）．

全ての関係表を作成し終わった後，関係表の簡単化を行う．1 対 1，1 対 N（N 対 1）といった関係で関連付けられている関係表を一つに統合することで，データの構造を簡潔にすることができる．参照制約は，実際に格納されているデータの間の対応関係を調べることで推定する．

4.2 次元表の導出

OLAP 分析を行うためには，分析の軸に対応する次元表が必要である．Linked Data では，以下のような形で次元表を導出することが考えられる．

データに直接記述されたリテラルを利用する リテラルで記述された時刻や地理空間情報（緯度経度高度）などは，データそれ自身が概念階層と関連付けられていると考えられる．時刻であれば Time Ontology[4]，地理空間情報であれば WGS84 Geo Position Ontology[5]などの利用が考えられる．

データに直接記述された階層構造を利用する RDF はそもそも概念体系を記述ことができるため，データセット自身が概念階層を記述している場合がある．このような階層関係は，多くの場合リソースの間の参照関係によって表現されているため，簡単化によって生成された表自身が次元表をなしている場合があり，その場合はそれをそのままを利用する．

データセット外部にある階層構造を利用する Linked Data の特徴である，外部データセットへのリンクを利用して，階層構造を得る手法である．例えばあるデータが，地理情報の記述に GeoNames[6]へのリンクを持っている場合，GeoNames から得られるリソースを解析することで概念階層を得ることが可能である．特に GeoNames の場合，ある地点のリソースは，それを包含する親となる地点のリソースをリンクしているため，比較的簡単に階層構造を得ることが可能である．

表 3: 水平表現による RDF トリプルの格納

ID	URI_ID	foaf:name	foaf:interest
1	1	“井上寛之”	2
2	...	...	...

[4]http://www.w3.org/2006/time
[5]http://www.w3.org/2003/01/geo/

4.3 OLAP 分析

データセットの変換および次元表の導出が終わった段階で，スキーマを利用者に提示し，1) 集計対象となるメジャー属性，2) 分析の軸となる次元表の選択を依頼する．あとは，通常の OLAP システムを利用することで，Linked Data の OLAP 分析が可能となる．

5 まとめ

本論文では，Linked Data に対して通常の OLAP システムを使った多次元分析を可能にする手法を提案した．今後はシステムの実装と実データによる評価を行う予定である．

謝辞

本研究の一部は科学研究費補助金若手研究 (B)（#23700102）による．

参考文献

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. International Journal on Semantic Web and Information Systems (IJSWIS), 2009.

[2] J. Beckmann, A. Halverson, R. Krishnamurthy, and J. Naughton. Extending RDBMSs To Support Sparse Datasets Using An Interpreted Attribute Storage Format. In ICDE, 2006.

[3] B. Kampgen, and A. Harth. Transforming Statistical Linked Data for Use in OLAP Systems, I-SEMANTICS 2011, 7th Int. Conf. on Semantic Systems, 2011.

[4] K. Wilkinson. Jena Property Table Implementation, Development HPL-2006-1, 12 (2006).

[6]http://www.geonames.org/