Managing Distributed Statistical Data in the Semantic Web

Sebastian Bayerl[1]

Abstract: The RDF Data Cube Vocabulary is the W3C recommendation for publishing multi-dimensional statistical data in a semantic web format. A large and growing number of such data cubes is already available in the linked data cloud, but the data is distributed and isolated in many remote repositories. Having access to the data, applications can now be developed to explore and use it. To handle and correlate the distributed cubes, a process is necessary to merge them. This research aims at developing a homogeneous management process for such cubes reaching from creation to visualization. The focus is put on the merging of RDF data cubes. In order to access and integrate the distributed cubes, a suitable ranking and discovery mechanism is needed. Therefore, a similarity measure for cubes must be developed.

Keywords: Knowledge Discovery, Data Integration, RDF Data Cube Vocabulary, Linked Open Data

1 Motivation

The linked data cloud[2] offers open access to a vast amount of information. Here, companies, governments and research institutions publish various datasets, like financial statistics or historical data. The core idea of this semantic web is to publish and link different datasets using specialized vocabularies to later gain new insights from the data. The resource description framework (RDF) [WLC14] should be used for the datasets and to connect the data, but there are also lots of datasets available as comma separated values (CSV). The RDF Data Cube Vocabulary [CR14] is the W3C recommended data format for statistical linked data. It is mainly used to publish statistical datasets to the linked data cloud. The cube vocabulary defines a multi-dimensional data cube, which reproduces a data structure - known from traditional Online Analytical Processing (OLAP), namely the OLAP cube [CD97].

This research project targets the management of RDF data cubes in the semantic web to enable an integrated view on the currently distributed and heterogeneous cubes. Hereby, the schema mapping and the data fusion problem must be tackled, which is stated as one of the fundamental research questions for linked data [BHBL09].

Therefore, a homogeneous management process for RDF data cubes will be developed. This process focuses on the discovery and the merging of distributed data cubes, while taking the semantic nature of the datasets into account.

[1] University of Passau, Faculty of Computer Science and Mathematics, Innstraße 43, 94032 Passau, sebastian.bayerl@uni-passau.de

[2] `http://linkeddata.org`

2 Problem Setting

The process of managing data cubes can be divided into a sequence of tasks, which are shown in Figure 1. This process is based on the RDF Data Cube Vocabulary to employ a homogeneous data structure for statistical data that fosters semantic properties like interlinked and disambiguated concepts. These properties are an additional value that is mostly not available in traditional data warehousing. This enables new functionality in all steps of the managing process, like the semantic comparison of data.

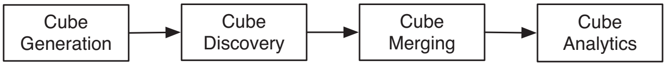


Fig. 1: RDF Data Cube Management Process

To open up this process for other data formats a *cube generation* step is necessary. Hereby, e.g. HTML tables or CSV documents are converted to RDF data cubes. Consequently, the following process can be used with statistical data that is initially not in the RDF data cube format. In order to find and access relevant cubes for the current use case, a suitable *cube discovery* mechanism is needed. Therefore, decentralized and remote repositories must be searched. Also a suitable similarity measures for cubes must be developed to be able to order cubes based on their content or structure.

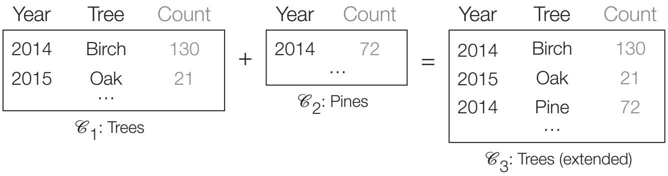


Fig. 2: Cube merging example

Cube merging is the central problem setting while managing data cubes. The main goal of cube merging is to integrate multiple isolated and potentially distributed cubes into a single dataset. Hereby, the validity of the structure and the data must be traced and maintained by finding suitable transformations. The merging process therefore produces an integrated view on a set of cubes. Figure 2 shows the simplified and abstract concept of the cube merging problem. Cube $\mathscr{C}_3$ contains all information from $\mathscr{C}_1$ and $\mathscr{C}_2$. To do so, it is necessary to introduce the column *Tree* with the default value *Pine* in cube $\mathscr{C}_2$. Similar to traditional data warehousing processes, OLAP operations like *roll-up* or *slicing* can now be performed on the integrated data cubes. This *cube analytics* process must efficiently enable the application of these operations. Visualizing the cubes can help to understand the integrated data and to draw new conclusions from it.

3 Approach

Figure 3 shows, how the different steps of the cube management process can be approached. In the following, these steps will be introduced in more detail.

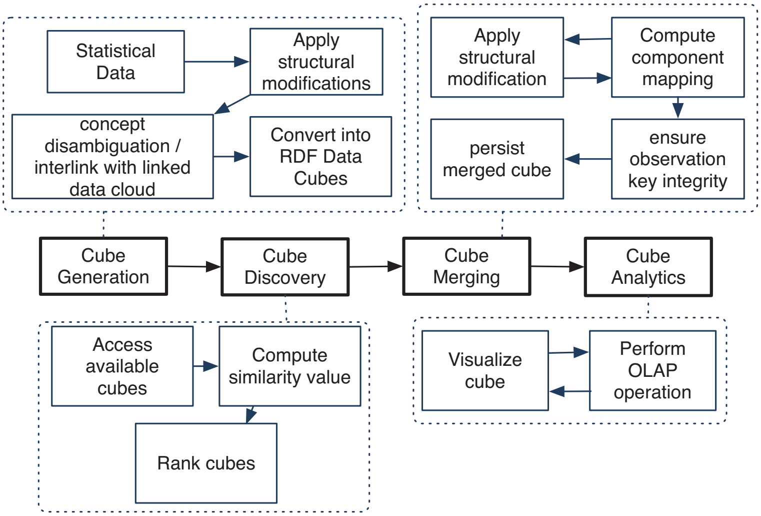

Statistical Data
Apply structural modifications
concept disambiguation / interlink with linked data cloud
Convert into RDF Data Cubes
Apply structural modification
Compute component mapping
persist merged cube
ensure observation key integrity
Cube Generation
Cube Discovery
Cube Merging
Cube Analytics
Access available cubes
Compute similarity value
Rank cubes
Visualize cube
Perform OLAP operation

Fig. 3: Detailed Cube Management Process

The main problem of converting statistical data into cubes lies within the normalization of the potentially complex structured data to a homogeneous structure. HTML tables for example are designed to be human-readable but do not reassemble the standardized multidimensional data cubes. A simple row-by-row based conversion approach is not sufficient and therefore, the complex structured tables must be validated, modified and transformed, while preserving the meaning of the data. Also, the data values must be disambiguated and interlinked with the linked data cloud to enable the semantic features of the following process. This approach is already implemented and published for a subset of the German Reich Statistics dataset [BG15b].

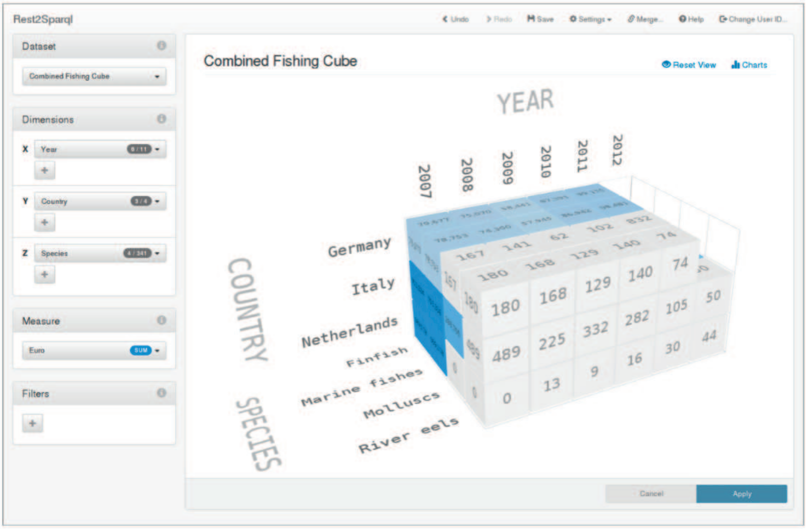
Cube discovery can be approached as an information retrieval problem, because selecting an appropriately ordered list from all available cubes satisfies an information need. Hereby, syntactic and semantic properties of the cubes can be considered to develop similarity measures for cubes. Currently, measures based on string similarity, concept equality, hierarchical relatedness and semantic similarity are considered good candidates to compare the structure of the cubes. Therefore, a graph-based shortest-path algorithm that utilizes the DBpedia category dataset[3] and a Word2Vec [Mi13] model will be implemented. This research will show if these measures or a combination of them are suitable to carry out the discovery process.

To be able to apply the merging process to cubes using the RDF Data Cube Vocabulary, it is necessary to analyze their structure and content to determine the properties of the resulting cube. An iterative process is used to adapt the structure of the cubes until a bijective mapping of the components can be found. This process tackles the following essential problems:

[3] http://wiki.dbpedia.org/services-resources/datasets/dbpedia-datasets

1. Modify the structure of the input cubes to fit the structure of the resulting cube (Integration on schema level).
2. Detect and handle duplicate observations and therefore maintain the global key integrity. This makes it necessary to develop an update strategy (Integration on instance level).

Beside other metadata about the process, the provenance information is stored with the resulting cube to track modification of the data and to ensure data quality. See [BG15a] for more information on this topic.


Dataset: Combined Fishing Cube

Dimensions:
X Year 6/11
Y Country 3/4
Z Species 4/341

Measure: Euro SUM

Fig. 4: Cube Visualization GUI

Cube analytics is the last step in the cube management process. A graphical user interface was implemented to visualize the (integrated) data cubes, shown in figure 4. It enables basic OLAP operations to manually browse and aggregate the data, using a 3D-rendered model.

4 Related Work

The presented process is similar to traditional Data Warehousing in various aspects. On the one hand, both deal with the problems of cleaning, transforming, merging and aggregating data [Le03]. On the other hand, handling RDF-based and interlinked open data from the linked data cloud yields advantages and new challenges, like semantic hierarchies and similarities.

There is work describing the process of converting statistical data into RDF data cubes. The *CSV2DataCube* tool converts CSV files [Sa12] and the *Data Extractor* is able to handle HTML and tables, extracted from PDF documents [St12]. These approaches allow basic restructuring, but complex structural modifications cannot be applied to the data.

Multidimensional hierarchical spaces are an essential part in OLAP processes. They are employed for data aggregation, but can also be used to define similarity measures for multidimensional data [BRV11]. Here, the similarity of cubes is computed by utilizing the distances of the cube facts according to the appropriate hierarchies. This approach handles OLAP cubes, but does not consider the specific properties of the RDF Data Cube Vocabulary. Several approaches to determine the similarity of words or linked data concepts have been presented in literature. An extensive description of semantic measures can be found in [Ha13]. Here, general definitions, the basics and a comprehensive classification can be found. The author of [Re99] describes the semantic similarity in a taxonomy in great detail. The survey [EAM14] compares state-of-the-art semantic similarity measures.

The authors of [KSH14] propose an approach to integrate RDF Data Cubes into a global cube based on the Drill-Across operation [ASS02]. If correct mappings for the contained cubes are known, merging and querying operations are possible. The reuse phase of the OpenCube lifecycle [Ka14] proposes approaches, how RDF data cubes can be handled.

5 Conclusion and Future Work

This paper presents a management process for RDF data cubes. This process can be divided into the independent sub-steps creation, discovery, merging and analytics. First results regarding cube creation and cube merging are already published [BG15b, BG15a]. Therefore, research prototypes have been implemented and evaluated. Currently, sophisticated similarity measures for cubes are developed to replace a naive approach, connecting the pool of available cubes with the merging process. This will show, which syntactic or semantic properties of RDF Data Cubes can be reasonably used for cube discovery. This will enable the comparison and ranking of data cubes.

6 Acknowledgement

This research is supervised by Prof. Dr. Michael Granitzer, University of Passau, Faculty of Computer Science and Mathematics / MiCS - Media Computer Science.

References

[ASS02] Abelló, Alberto; Samos, José; Saltor, Felix: On Relationships Offering New Drill-across Possibilities. In: Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP. ACM, pp. 7–13, 2002.

[BG15a] Bayerl, Sebastian; Granitzer, Michael: Bacon: Linked Data Integration Based on the RDF Data Cube Vocabulary. In: Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics. WIMS '15, ACM, New York, NY, USA, pp. 14:1–14:6, 2015.

[BG15b] Bayerl, Sebastian; Granitzer, Michael: Data-transformation on Historical Data Using the RDF Data Cube Vocabulary. In: Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business. i-KNOW '15, ACM, New York, NY, USA, pp. 15:1–15:8, 2015.

[BHBL09] Bizer, Christian; Heath, Tom; Berners-Lee, Tim: Linked Data – The Story So Far. International Journal on Semantic Web and Information Systems, 5(3):1–22, 2009.

[BRV11] Baikousi, Eftychia; Rogkakos, Georgios; Vassiliadis, Panos: Similarity measures for multidimensional data. In: Data Engineering (ICDE), 2011 IEEE 27th International Conference on. IEEE, pp. 171–182, 2011.

[CD97] Chaudhuri, Surajit; Dayal, Umeshwar: An Overview of Data Warehousing and OLAP Technology. ACM SIGMOD Record, 26(1):65–74, March 1997.

[CR14] Cyganiak, Richard; Reynolds, Dave: The RDF Data Cube Vocabulary. W3C recommendation, W3C, January 2014. http://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/.

[EAM14] Elavarasi, S Anitha; Akilandeswari, J; Menaga, K: A Survey on Semantic Similarity Measure. International Journal of Research in Advent Technology, 2(3):389–398, March 2014.

[Ha13] Harispe, Sébastien; Ranwez, Sylvie; Janaqi, Stefan; Montmain, Jacky: Semantic Measures for the Comparison of Units of Language, Concepts or Entities from Text and Knowledge Base Analysis. CoRR, 2013.

[Ka14] Kalampokis, Evangelos; Karamanou, Areti; Nikolov, Andriy; Haase, Peter; Cyganiak, Richard; Roberts, Bill; Hermans, Paul; Tambouris, Efthimios; Tarabanis, Konstantinos: Creating and Utilizing Linked Open Statistical Data for the Development of Advanced Analytics Services. In: Second International Workshop for Semantic Statistics SemStats. 2014.

[KSH14] Kämpgen, Benedikt; Stadtmüller, Steffen; Harth, Andreas: Querying the Global Cube: Integration of Multidimensional Datasets from the Web. In: Knowledge Engineering and Knowledge Management, pp. 250–265. Springer, 2014.

[Le03] Lehner, Wolfgang: Datenbanktechnologie für Data-Warehouse-Systeme: Konzepte und Methoden. dpunkt-lehrbuch. dpunkt-Verlag, 2003.

[Mi13] Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey: Efficient Estimation of Word Representations in Vector Space. CoRR, 2013.

[Re99] Resnik, Philip et al.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res.(JAIR), 11:95–130, 1999.

[Sa12] Salas, Percy E. Rivera; Martin, Michael; Mota, Fernando Maia Da; Auer, Sören; Breitman, Karin; Casanova, Marco A.: Publishing Statistical Data on the Web. In: Sixth IEEE International Conference on Semantic Computing, ICSC 2012, Palermo, Italy, September 19-21, 2012. pp. 285–292, 2012.

[St12] Stegmaier, Florian; Seifert, Christin; Kern, Roman; Höfler, Patrick; Bayerl, Sebastian; Granitzer, Michael; Kosch, Harald; Lindstaedt, Stefanie N.; Mutlu, Belgin; Sabol, Vedran; Schlegel, Kai; Zwicklbauer, Stefan: Unleashing Semantics of Research Data. In: Specifying Big Data Benchmarks - First Workshop, WBDB 2012, San Jose, CA, USA, May 8-9, 2012, and Second Workshop, WBDB 2012, Pune, India, December 17-18, 2012, Revised Selected Papers. pp. 103–112, 2012.

[WLC14] Wood, David; Lanthaler, Markus; Cyganiak, Richard: RDF 1.1 Concepts and Abstract Syntax. W3C recommendation, W3C, February 2014. http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.