Visual Analysis of Scientific Content

Belgin Mutlu and Vedran Sabol, Know-Center GmbH

Abstract—The steadily increasing amount of scientific publications demands for more powerful, user-oriented technologies supporting querying and analyzing scientific facts therein. Current digital libraries that provide services to access scientific content are rather closed in a way that they deploy their own meta-models and technologies to query and analyse the knowledge contained in scientific publications. The goal of the research project CODE is to realize a framework based on Linked Data principles which aims to provide methods for federated querying within scientific data, and interfaces enabling user to easily perform exploration and analysis tasks on received content. The main focus in this paper lies on the one hand on extraction and organization of scientific facts embedded in publications and on the other hand on an intelligent framework facilitating search and visual analysis of scientific facts through suggesting visualizations appropriate for the underlying data.

Keywords—Research Data, Linked Data, RDF Data Cube, Visualization, Visual mapping

1 Introduction

It is an accepted fact that managing, querying and analysing research data contained in scientific publications using traditional Web technologies is a difficult and unintuitive task. Digital libraries which control the life-cycle of the scientific publications, i.e., publishing and making them accessible for certain community, mainly expose the research knowledge using domain-specific meta-models and technologies, such as the widely used Dublin Core meta-model. This domain-specificity limits the ability to effectively find the desired information. In addition, scientific data or facts included in publications are unstructured or, at best, in tabular format, so that once the information is found, there is a hardly way to reuse it.

To address this issue we provide methods to automatically (i) extract facts (i.e., statistical data such as evaluation results) from scientific publications, (ii) represent them in an unified format by using the RDF Data Cube Vocabulary (RDF-DCV) [1], and (iii) organise them within the Linked Open Data (LOD) global repository using semantic web technologies. The strength of LOD lies in its interlinking structured data in a format that can be read and processed by computers which in turn makes the knowledge more accessible and transparent. We utilise the semantic information to support our central contributions: (iv) easy to use discovery of information in the LOD, and (v) interactive visual analysis of the retrieved Linked Data.

The semantic web is getting more and more popular with the growing need for automating the information retrieval process. However, the problem here is the tedious process of accessing this data without having in-depth knowledge of semantic technologies. There are browsers/interfaces which support users to access Linked Data, but they are either not user-friendly or have been provided only for special tasks. Hence, there is a need for a tool which enables non-expert users to easily perform exploration and analysis tasks on Linked Data. Besides supporting functions like filtering and navigation, a Linked Data collection can also be analysed using visual means, for instance for analysing statistical facts helping users to gain a deeper insight into the data.

For extracting statistical facts in the form of tables from publications in PDF format we apply a **PDF Extractor** [6]. Subsequently, a **Data Extractor** [2] converts the table in an RDF Data Cube which is saved in a triple store. To allow users to query, analyse and organise the extracted statistical data, we provide

- *B. Mutlu and V. Sabol are with Know-Center GmbH, Inffeldgasse 13, 8010 Graz, Austria.*
E-mail: bmutlu, vsabol@know-center.at

Manuscript received April 30, 2015; revised XX, 2015.

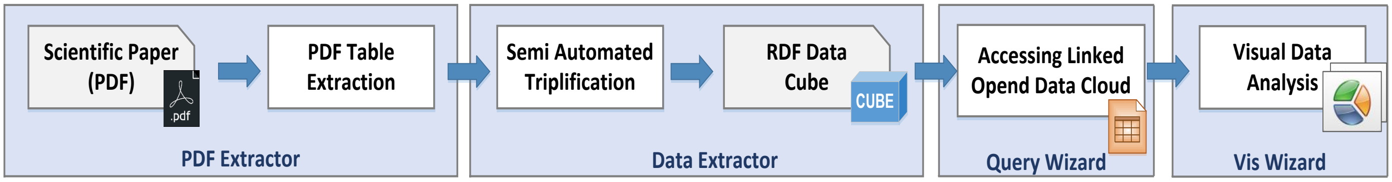


Fig. 1: Schematic representation of the CODE pipeline to extract and analyse scientific facts embedded in publications.

Query Wizard [3] and Visualization Wizard (*VisWizard*) [4]. The former aims as an interface to the Linked Data repositories that simplifies the process of accessing, displaying, filtering, exploring and navigating the Linked Data. The later enables visual analysis of the Linked Data with the emphasis on automation of the visualisation process. For that purpose, the *VisWizard* uses a generic mapping model that maps the extracted data to the integrated visualisations and generates in turn the appropriate visualisations for the underlying data.

Thus, the contribution of this work can be summarised as a publication processing pipeline consisting of (1) a strategy to extract and organise scientific facts embedded in publications, (2) an approach for easily accessing, exploring and navigating Linked Data, and (3) an interactive visual interface which automatically generates the appropriate visualisations for the current data.

2 Related Work

The closest approach to our work for analysing scientific facts contained in research publications is a system described by Attwood et al. [5] which uses a PDF reader to simplify the search for relevant information. This intelligent tool provides additional data (links to suitable web resources and metadata) and methods to interactively analyse the content of the evaluation results embedded in papers (in tables to be exact). For the visualisation, the table content and figures are enriched with provenance information. Visualisations are created following a very common principle: columns are statically mapped to axes of a scatter plot. There is no vocabulary defining the structure of this data and, therefore, no intermediate model which would allow to make more sophisticated visualisation mappings.

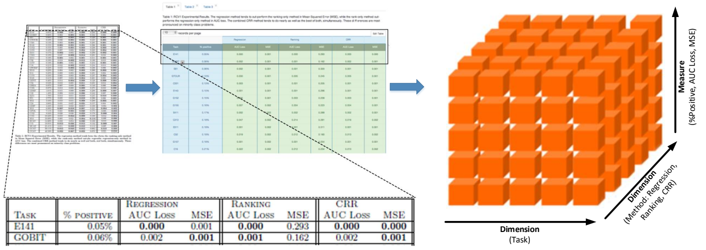
Salas et al. [1] developed a framework which supports the visualisation and visual querying for statistical data sets, stored in form of the RDF Data Cube. However, in contrast to our approach, there is no support for automated suggestions of possible visualisations and also no support for large cubes (i.e. cubes having varying number of dimensions and multiple measures). Furthermore, the framework does not rely on semantic description of charts and it offers just a small number of simple chart types.

3 Analysing Content of Scientific Publications

As shown in Figure 1, in order to access and analyse scientific facts embedded in scientific publications, they first have to be extracted. To achieve this, we use an embedded PDF extractor that takes a PDF file as input and returns extracted tables as output. Once the tables are extracted the content have to be represented in a uniform format. To do so, a triplifier transforms the extracted data into RDF Data Cube(s) which we then publish in a Linked Data endpoint. Using the Query Wizard user can access and explore the content of the generated cubes. To visualise the Cube content, the *VisWizard* is activated, which automatically creates the appropriate visualisations. In the following subsections we briefly describe each of the these units.

3.1 PDF Extractor and Data Extractor

The PDF extractor is responsible for extracting and managing tabular information embedded in scientific publication. This is achieved



Task	% positive	Regression		Ranking		CRR	
		AUC Loss	MSE	AUC Loss	MSE	AUC Loss	MSE
E141	0.05%	**0.000**	0.001	**0.000**	0.293	**0.000**	**0.000**
GOBIT	0.06%	0.002	**0.001**	**0.001**	0.162	0.002	**0.001**

Measure
(%Positive, AUC Loss, MSE)
Dimension
(Method: Regression, Ranking, CRR)
Dimension
(Task)

Fig. 2: Pipeline to extract and structure scientific facts embedded in publications.

through the analysis of publication structure using unsupervised machine learning techniques and heuristics. Contiguous text blocks and geometrical relations between them are extracted from the character stream. The blocks are categorised into different classes resulting in a logical structure of the document. Table extraction starts from a ”table” caption, and then labels neighbouring sparse blocks recursively as table blocks, if their vertical distance is within a specific threshold.

The Data Extractor uses a (semantically enriched) HTML table to define and semantically describe the components of the cube(s). The columns of the table are automatically classified as either nominal/categorical data (if the cell content is non-numeric), numerical data (for numeric cell content), or multi-value data (if there are multiple values in at least one of the cells of the given column). This process is illustrated in Fig. 2. The cube is subsequently saved in a triple store and made accessible through a SPARQL endpoint.

3.2 Query Wizard

The Query Wizard aims to help users to select relevant data from Linked Data repositories. It can be used to explore RDF Data Cubes available through a SPARQL endpoint and also to explore raw RDF. The user just performs a keyword search in the desired endpoints and obtains the resulting data presented in an easy-to-use web-based interface very similar to the spreadsheet applications (e.g. Microsoft Excel Table). A row corresponds to a single subject and a column represents a predicate. Cells contain objects, i.e. any number of literals and/or entities, depending on the row and column. Users can select columns to be shown, load any number of rows, and perform operations such as filtering.

3.3 Visualization Wizard

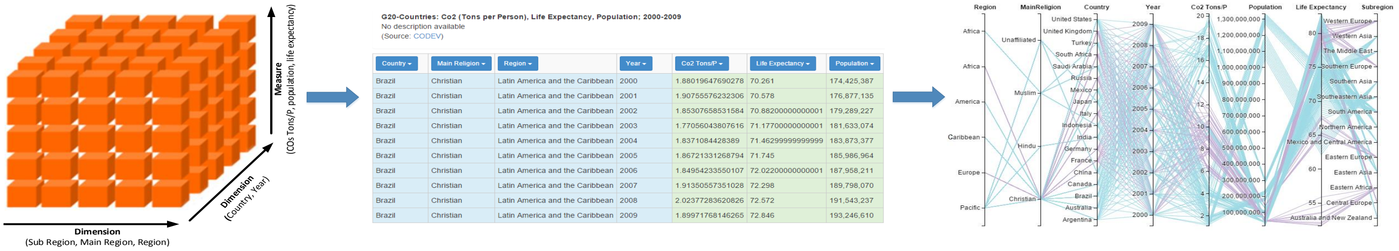
In contrast to other available solutions for visualising Linked data [7] the *VisWizard* automatically suggests the appropriate visualisations for the current data. The following parts contribute to this feature:

Vocabularies. The RDF-DCV is used to represent statistical data as a collection of so called observations, each consisting of a set of dimensions and measures. The dimensions identify the observations and measures are related to concrete values. The format of RDF-DCV guarantees a uniform representation for all unstructured statistics, thereby enabling the *VisWizard* to access the data in a standard way defined by the RDF-DC specification.

The Visual Analytics (VA) Vocabulary is used to represent the semantic of visualisations. Our semantic description strictly focuses on describing the visual encoding process, hence we represent visualisations in terms of their visual channels (e.g., axes of a visualisation, colours etc.) together with data types they support.

Mapping Vocabularies. The mapping between both mentioned vocabularies is a relation from dimensions and measures (cube components) of the Data Cube to the corresponding visual channels of the visualisation. This relation is valid if the data types of the cube components and visual channels of a visualisation are compatible. Beyond the data type compatibility, a valid mapping needs to account the structural compatibility. Thus, only the combinations with the same instantiation patterns that match the observations of the RDF Data Cube are candidates for valid mappings.

Visualisation Process. Based on provided RDF Data Cube model, *VisWizard* proposes the mapping candidates, i.e., (1) the visualisations and (2) possible variants of the visualisations. The former is done by mapping the dimensions and measures with the provided visual channels. However, when multiple dimensions are presented in a RDF Data Cube, more than one visualisation can be suggested. In this case



G20-Countries: Co2 (Tons per Person), Life Expectancy, Population; 2000-2009
No description available
(Source: CODEV)

Country	Main Religion	Region	Year	Co2 Tons/P	Life Expectancy	Population
Brazil	Christian	Latin America and the Caribbean	2000	1.88019647690278	70.261	174,425,387
Brazil	Christian	Latin America and the Caribbean	2001	1.90755576232306	70.578	176,877,135
Brazil	Christian	Latin America and the Caribbean	2002	1.85307658531584	70.88200000000001	179,289,227
Brazil	Christian	Latin America and the Caribbean	2003	1.77056043807616	71.17700000000001	181,633,074
Brazil	Christian	Latin America and the Caribbean	2004	1.8371084428389	71.46299999999999	183,873,377
Brazil	Christian	Latin America and the Caribbean	2005	1.86721331268794	71.745	185,986,964
Brazil	Christian	Latin America and the Caribbean	2006	1.84954233550107	72.02200000000001	187,958,211
Brazil	Christian	Latin America and the Caribbean	2007	1.91350557351028	72.298	189,798,070
Brazil	Christian	Latin America and the Caribbean	2008	2.02377283620826	72.572	191,543,237
Brazil	Christian	Latin America and the Caribbean	2009	1.89971768146265	72.846	193,246,610

Fig. 3: Pipeline to explore, navigate and analyse scientific facts extracted from publications.

(the option (2)), the framework creates a candidate table including all possible combinations between both models. For each combination variant, a specific visualisation is created.

The querying and visualisation process of the Cubes are illustrated in Fig. 3.

4 Usage Scenario

Consider following scenario: Jane studies the paper *Combined Regression and Ranking* and feels overwhelmed with the numbers spread across the tables through the pages. She wishes to explore the data in the Table 1 displaying the experimental results (see Figure 2, on the left). To do so, she uploads the PDF file into the Data Extractor which internally uses the PDF Extractor to extract the tables. Next, Jane selects the table (experiment results) she would like to analyse and activates the Data Extractor to generate a RDF Data Cube (see Figure 2, on the right) from the content of the table. Once the Cube is generated, Jane can either explore the Cube content using Query Wizard or visualise it using *VisWizard*. If a visual analysis required, the *VisWizard* suggests parallel coordinates as one of the possible visualizations which shows Jane the trends of the evaluated algorithms.

5 Conclusion

This paper introduces a framework that helps user to organise and analyse publications. The framework uses (1) a PDF Extractor to extract tables from publications, (2) Data Extractor to organise and structure the extracted data in a standard way using RDF Data Cube Vocabulary, (3) and interactive tools (Query- and Visualization Wizard), to explore and analyse the content of the extracted data. The research data has to traverse several stages, starting from the textual, over tabular to its final structured form and the visualisations. The essential aspect of the approach is the automated support for this workflow which simplifies the process to explore, analyse and manage scientific data without having in depth knowledge about PDF extraction, semantic web technologies or visualisations.

The CODE framework[1] has been released online since over two years and has been actively used for accessing and analysing Linked Open Data. The components of the framework have been evaluated separately whereby the results of each evaluation can be taken from the corresponding publication.

Acknowledgements. This work is funded by the EC 7th Framework project EEXCESS (grant 600601). The Know-Center GmbH is funded within the Austrian COMET Program - managed by the Austrian Research Promotion Agency (FFG).

References

[1] P.E.R. Salas et al., *Publishing Statistical Data on the Web*: Semantic Computing (ICSC), 2012.

[2] C. Seifert et al.,*Crowdsourcing fact extraction from scientific literature*: SouthCHI, 2013.

[3] P. Hoefler et al., *Linked Data Query Wizard: A novel Interface for Accessing SPARQL Endpoints* : LDOW at WWW, 2014.

[4] B. Mutlu et al., *Suggesting Visualizations for Published Data*: VISIGRAPP, 2014.

[5] T.K. Attwood et al., *Utopia documents: linking scholarly literature with research data*: Bioinformatics, Oxford Journals, 2010.

[6] S. Klampfl et al., *An unsupervised machine learning approach to body text and table contents extraction from digital scientific articles*: TPDL, 2013

[7] D. Aba-Sah et al., *Approaches to visualizing linked data: a survey*: Semant. Web, IOS Press, 2011.

[1] http://code-research.eu/results/