Applying Multidimensional Navigation and Explanation in Semantic Dataset Summarization

James R. Michaelis, Deborah L. McGuinness, Cynthia Chang, Joanne S. Luciano, and James Hendler

Tetherless World Constellation, Rensselaer Polytechnic Institute,
110 8th Street, Troy, NY 12180
{michaelis,dlm,csc,jluciano,hendler}@cs.rpi.edu

Abstract. A key objective of multidimensional dataset analysis is to reveal patterns of interest to users, but can be difficult to conduct due to the challenges of both presenting and navigating large datasets. This work explores how initial summarizations of multidimensional datasets can be generated (designed to reduce the number of data points which would need to be displayed), using *summarization policies* based on provided dataset values. Additionally, functionality for explaining the derivation of summarizations is being designed in line with prior work on aiding analyst interactions with data processing systems. To help drive development of this work, as well as provide illustrative use cases, we are presently designing a dataset summarization generator as part of greater work being done on an infrastructure for managing evidence of technical emergence in varying research disciplines via automated review of published materials.

Keywords: OLAP, Explanation, Provenance

1 Introduction

A key objective of multidimensional dataset analysis is to reveal patterns of interest to analysts. In many cases, these analyses will involve navigation over a dataset to expose content likely to have interesting patterns. However, multidimensional analysis has been observed to be challenging to analysts for the following reasons [1]:

1. They may be overwhelmed by a data space evidence set if it is too large.
2. They may not have time or expertise to perform extensive navigation.

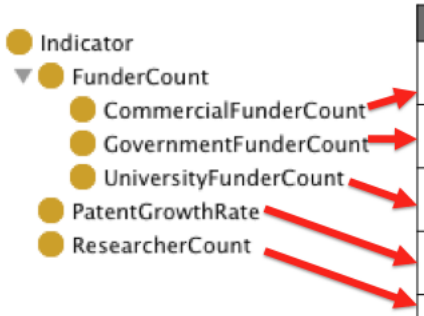
This work explores how initial summarizations of multidimensional datasets can be generated for consuming parties (designed to reduce the number of data points which would need to be displayed) driven by *summarization policies* based on provided dataset values. Focus has been given to RDF-based dataset encodings, due largely to RDFs flexibility in linking to outside data sources (e.g., ontologies for expressing possible data values). Finally, functionality for explaining the derivation of summarizations is being developed - in line with prior work for aiding analyst interactions with data processing systems [2].

2 Evidence Summarization in the ARBITER System

To help drive development of this work, as well as provide illustrative use cases, we are presently developing a dataset summarization generator for the Abductive Reasoning Based on Indicators and Topics of EmeRgence (ARBITER) system - being jointly developed by Rensselaer, BAE Systems, NYU, Brandeis and 1790 Analytics as part of IARPA's Foresight and Understanding from Scientific Exposition (FUSE) program. ARBITER's design objective is to scan for signs of technical emergence in published literature - where technical emergence is defined in the FUSE program as [3]: *the process by which research domains appear, mature, and if conditions are favorable, make a significant impact.*

In ARBITER, sets of one or more evidence entries are evaluated to make hypotheses about emergence-related questions for a given topic and time period. For example: *Has a practical application for DNA Microarrays been established in the time period of 2006-2010, based on the document collection PubMed-42?*

In this setting, evidence entries are defined as *emergence indicators*, calculated based on analysis over document collections. Indicators are classified according to an OWL ontology of indicator types, where each indicator is defined to have at least one RDF type, as well as a set of numerical scoring metrics to define relationship of evidence to hypothesis. For brevity, an example is provided with five indicators, each with a single RDF type and two numerical properties (*value* and *relevance to the question answer*, where a higher value is better).



Indicator Type	Value	Relevance
Count of Commercial Funding Agencies	19	0.70
Count of Government-based Funding Agencies	82	0.58
Count of University-based Funding Agencies	5	0.42
Growth Rate for Assigned Patents	1.09	0.68
Count of Participating Researchers	518	0.53

Fig. 1. *Left:* Part of ARBITER's ontology, which defines indicators. *Right:* A sample of indicator data. Here, indicators are aligned with their corresponding ontology classes.

Currently, these evidence entries are presented as a 2-dimensional spreadsheet. To reduce the number of rows directly presented, policy-based summarization techniques are being explored - deriving from established navigation techniques in OLAP [1]: grouping rows into *collection-based* entries, as well as filtering table entries - each based on specified criteria. For this submission, the following two summarization policies are provided for illustrative purposes:

1. **Grouping:** Group entries together that are SKOS[1] subconcepts of the "FunderCount" class.
2. **Filtering:** Remove entries with relevance scores below 0.55.

Ultimately, the following system conditions are assumed: (i) A maximum number of *summary rows* will be specified, which will appear in the presented summary; (ii) A pre-defined collection of policies will be accessible by ARBITER, along with a pre-defined ordering for their execution; (iii) Policies will be sequentially applied to the evidence set until the summary row count is reached, or all policies have been applied. Initially, an evidence dataset D_0 will represent content directly generated by evidence gathering routines in ARBITER. Each policy execution will yield a transformed dataset view $D_{1...n}$, up until condition (iii) is satisfied.

While initial summarization can be a powerful aid for analyst users, care has to be taken in their usage, since one summarization strategy may not be appropriate for all users and information-seeking tasks. To help analysts keep track of applied strategies, summaries will be accompanied by explanations of their derivation - accessible for individual entries. In Figure 2, an example summary view - along with a supporting explanation - is provided.

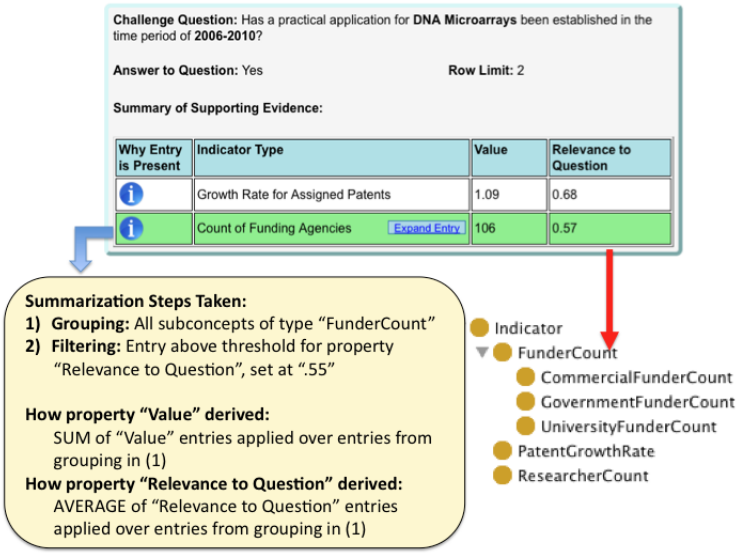


Fig. 2. Here, a summary view is displayed, with a collection-based entry highlighted in green (which aggregates the 3 funding entries from Figure 1). An explanation of why this entry is present - and how column values were derived - can be accessed by clicking the "i" icon.

[1] RDF SKOS Vocabulary: http://www.w3.org/TR/skos-reference/

System Development: ARBITERs summary generator is being designed to take three inputs: (i) A set of fine-grained evidence; (ii) A set of SPARQL-encoded preference policies, along with an accompanying execution order; and (iii) Corresponding ontologies for encoding the preference and evidence data. For encoding evidence, we are now exploring use of the RDF Datacube[2] vocabulary - given its support for representing multidimensional data.

Upcoming Directions: In upcoming work, focus will be given to the following three issues: (i) selection of summarization policies which align with an analysts perceived preferences, (ii) based on the summarization explanations provided, enabling analysts to tweak applied strategies to generate new summarizations, and (iii) enabling analysts to identify source documents used to create evidence entries (similar to efforts discussed in [2]). For situations where significant numbers of evidence entries are presented (e.g., over 100), all three issues are expected to need addressing.

3 Acknowledgements

We would like to thank our collaborators at BAE Systems, Sean Stromsten, Dan Hunter and Olga Babko-Malaya for their assistance in this work. Support has been provided by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20154. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

1. Giacometti, A. and Marcel, P. and Negre, E. A framework for recommending OLAP queries. 11th International Workshop on Data Warehousing and OLAP (DOLAP08), 73-80, 2008.
2. Murdock, J., McGuinness, D., Pinheiro da Silva, P., Welty, C., and Ferrucci, D. Explaining conclusions from diverse knowledge sources. Proceedings of ISWC 2006, 861-872, 2006.
3. Foresight and Understanding from Scientific Exposition (FUSE) Program - Broad Agency Announcement (BAA) [IARPA-BAA-10-06]. Retrieved from: http://www.iarpa.gov/solicitations_fuse.html. Date Last Accessed: 07/28/2012.

[2] RDF Datacube Vocabulary: http://www.w3.org/TR/vocab-data-cube/