Inferencing in the Large
Characterizing Semantic Integration of Open Tabular Data

Asha Subramanian

International Institute of Information Techmology,
26/C, Electronics City, Hosur Road,, Electronic City, Bengaluru, Karnataka 560100, India
asha.subramanian@iiitb.org
www.iiitb.ac.in

Abstract. *Tables are a natural and ubiquitous way of representing related information. Actionable insight is usually gleaned from tabular datasets through data mining techniques assisted by domain experts. These techniques however, do not harness the semantics or the contextual reference underlying the datasets. Tabular datasets, especially the ones created as part of open data initiatives often contain information about entities fragmented across several datasets implicitly connected through some semantics thus giving them a contextual reference. Our work deals with harnessing this context (Thematic Framework) in which they can be reasoned further. This thesis aims at creating algorithmic support for a human to semantically integrate a collection of tabular data using ontologies from publicly available knowledge bases in Linked Open Data. The overall objectives of our work called "Inferencing in the Large" aims to go further than this, to enrich the mapped ontology with inferencing rules and generate enriched RDF (Schematic Framework), to enable the use of semantic reasoners.*

Keywords: Semantic Web · Linked Open Data · Context Abduction · Ontology · Graph Models

1 Introduction

Recent initiatives like Open Data by various governments, have resulted in a number of freely available tabular datasets containing actionable knowledge that could be relevant to several stakeholders. However, these published datasets are often created independently, with no overarching purpose or schematic structure. Indeed, there may be no overarching *thematic* structure – the datasets need not be about any one particular topic or theme. As a result, valuable knowledge remains fragmented across the datasets. While typical analytics efforts use data mining or machine learning algorithms to exploit data patterns and generate inferences, they fail to harness the implicit meaning of the data to make meaningful inferences in a semantic context.

There is a pressing need for *semantic integration* of such arbitrarily structured data.

Given a collection of tables, it is a daunting task even to determine what the set of tables are collectively about (that is, a topic or theme for the collection), let alone establish an overarching schematic framework.

We call this problem, “Inferencing in the Large” (in the wilderness), where in order to extract meaning from a collection of data, we first need to establish a framework within which meaning can be interpreted.

In a previous project called Sandesh (expanding to Semantic Data Mesh), we had proposed a knowledge representation framework based on Kripke Semantics called Many Worlds on a Frame (MWF) to integrate disparate datasets [1]. In this model, aggregated knowledge was represented in the form of several semantic “worlds” – each of which represented a schematic framework within which data was organized. Given a set of tabular data, and a hand-crafted set of seed worlds and their (type and location) relationships, the Sandesh toolkit reorganized the tabular data into data elements within the schematic frameworks of one or more worlds. MWF was meant to address the absence of a schematic and thematic framework in open datasets. However, the Sandesh framework still requires significant human effort in organizing the seed set of worlds and their schematic structures.

In this work, we aim to automate this process further, by using ontologies from the Linked Open Data cloud to explain a collection of tables. Firstly determine the Thematic Framework or the dominant concept(s) that the tables are collectively about and secondly, determine the Schematic Framework or entities/properties that each of the row values and column headers relate to in the context determined by the Theme. Such matched ontologies can be enriched in two ways: (a). they can be augmented with inference rules and new assertions to enable semantic reasoning within them, and (b). they can be interrelated to one another to form a global frame, that can support reasoning *across* them.

2 Related Work

Determining a meaningful context, extracting relevant ontologies and generating enriched RDF tuples from structured, unstructured and semi-structured data using appropriate ontologies from LOD cloud are all active research areas [6], [5], [7], [8].

We divide this broad literature into the following groups :

- **Identifying and Relating Concepts and Entities from Content**
  Tools such as Open Calais[1], FRED [2], Apache Stanbol[2], Fox[3] work on unstructured content, extract concepts and entities such as places, events, people, organisations etc and relate them to universally known entities from knowledge bases such as DBPedia, Freebase, Geonames etc. While Open Calais and FRED amongst these are the most advanced tools with capabilities to extract context and related entities, the ontology/metadata they use internally are proprietary, in the sense that the disambiguated entities refer to an internal Calais or a FRED URI/id. Our objective is to extract concepts for the identified context that can be related to an openly available knowledge base from the Linked Open Data Cloud without using any proprietary vocabulary. In the context of datasets in LOD cloud, Lalithsena et

[1] Open Calais - http://viewer.opencalais.com/

[2] Apache Stanbol - https://stanbol.apache.org/overview.html

[3] FOX: Federated knOwledge eXtraction Framework - http://aksw.org/Projects/FOX.html

al. in [9] use an interesting technique to identify domains for such datasets with an aim to annotate/categorize the datasets appropriately. They rely on the Freebase knowledge base to identify topic domains for LOD.

- **Extraction of RDF tuples from CSVs**
  Several research efforts have addressed extraction of RDF tuples from CSV files. Some prominent tools in this area include RDF Converter[4], Virtuoso Sponger[5], Open Refine[6] and RDF123 [3]. However, the generated RDF tuples are mostly still raw data without any contextual reference. These RDF tuples need to be semantically linked to knowledge sources such as the ones constituting the Linked Open Data Cloud[7] or other formal ontologies to extract meaningful inferences from the data.
- **State of Art : Understanding Semantics of Tables and generating enriched RDF**
  Some of the most recent and relevant work that compares to our research includes work by Mulwad [4]. Mulwad's work is quite comprehensive in determining the meaning of a table and uses parameterized graphical model to represent a table. Their core module performs joint inferencing over row values, column headers and relations between columns to infer the meaning of the table by using a semantic message passing scheme that incorporates semantics into the messages. The graph is parameterised on three variables 1) one to determine the classes the column should map to 2) second to determine the appropriate relations between the data values in a row 3) third to determine relation between the column headers. The joint inferencing module is an iterative algorithm that converges when the model variables agree on the alloted LOD classes/entities for the column headers, relations between columns and row values. They also generate enriched RDF encapsulating the meaning of the table.

While these efforts are attractive and generate quality linked data keeping the intended meaning of the data in mind, they still work on a single table and are largely data values driven. In our challenge, we are looking for the ontology/collection of classes from related ontologies that fit best a *set* of tables, each table contributing a set of its columns to the identified ontology(ies). We expect a set of tables to have different utilitarian views depending upon the desired context. Our research aims to provide this semantic framework wherein a set of tables are mapped to domain from LOD. The columns from various tables are linked to relevant properties of the domain classes that the tables are collectively about, the data values and relations between columns in the tables are instantiations of the domain classes and their properties. Our overall objectives from this research is also, given a set of tables, generate enriched RDF data that can be further exploited by semantic reasoners.

[4] RDFConverter: http://www.w3.org/wiki/ConverterToRdf

[5] Virtuoso Sponger: http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VirtSponger

[6] Open Refine: https://github.com/OpenRefine

[7] Linked Data Design Issues: http://www.w3.org/DesignIssues/LinkedData.html

3 Research Questions and Hypothesis

My thesis attempts to answer the overall research objective to semantically integrate a collection of tabular data sets by inferring a Thematic and Schematic Framework. The following research questions detail the research objective.

1. Given a Collection of arbitrary tabular datasets, is it possible to determine what the collections of tables is about? Can we relate this inferred theme in terms of known concept(s) from Linked Open Data (LOD) or a custom knowledge base? We call this the Thematic Framework. We envisage the Thematic Framework to contain *Concept/Domain Classes* from LOD that best describe the collection of datasets.
2. For the identified Thematic Framework, can we relate the column headers in the various tables as properties of the identified concept classes from LOD and data values as instances or entities of the concept classes? We call this the Schematic Framework consisting of the T-Box (class and property definitions) and A-Box (class and property assertions). We envisage the Schematic Framework to contain a) properties of the dominant classes that are most relevant for the columns in the datasets in line with the identified Thematic Framework b) A-Box instantiations for all the data values in the datasets using the dominant classes and their properties c) T-Box definitions for the hierarchy of dominant classes and their properties derived from the respective vocabularies in LOD
3. Finally, can the enriched RDF generated using the Thematic and Schematic Framework discussed above, be processed by a reasoner such as Apache Jena to perform semantic inferences?

We hypothesise that for tabular datasets with data values that can be linked to LOD or some available custom knowledge base, it is possible to infer dominant concepts that relate to concept classes from known ontologies and further map the column headers and data values of the tables to properties and entities of those concept classes respectively. The definitions of properties and classes described explicitly in an ontology (DBPedia, Yago and others from Linked Open Data Cloud) and those implicitly derived from the instance assertions together with additional evidence from column headers can be combined with graphical modelling techniques to achieve the research objective. Table 1 shows a sample Thematic and Schematic Framework output for a set of two input data files (Table a , Table b) that have information on some Indian states and their capitals and rivers and their state of origin.

4 Proposal

Our first goal is to find dominant concept classes that relate maximally to a given set of tables. This paper showcases preliminary results towards this first goal. We propose two approaches for the concept class(es) identification and combine the two to obtain an overall scoring. In the *bottom-up* approach, entities are searched from LOD to obtain classes that maximally subsume the data values in a column. We also use the *bottom-up* technique to mine properties for the columns that best relate to the relation between the data values in a pair of columns. In the *top-down* approach, we rely completely on the

Table 1: Sample Thematic and Schematic Framework

Table a: StatesandCapitals.csv

State	**Capital**
Andhra Pradesh	Hyderabad
Maharashtra	Mumbai
Karnataka	Bangalore
Tamil Nadu	Chennai
Uttarakhand	Dehradun

Table b: RiversandSourceState.csv

River	**Source**
Ganges	Uttarakhand
Yamuna	Uttarakhand
Godavari	Maharashtra
Krishna	Maharashtra
Kaveri	Karnataka

Thematic Framework : Dominant Concept Classes
http://dbpedia.org/ontology/PopulatedPlace
http://dbpedia.org/ontology/River
Schematic Framework
StateandCapitals/State a dbpedia-owl:PopulatedPlace
StateandCapitals/Capital a dbpedia-owl:PopulatedPlace
RiversandSourceState/River a dbpedia-owl:River
RiversandSourceState/Source a dbpedia-owl:PopulatedPlace
StateandCapitals/Capital a owl:ObjectProperty
StateandCapitals/Capital rdfs:domain dbpedia-owl:PopulatedPlace
dbpedia.org/resource/Karnataka a dbpedia-owl:PopulatedPlace
dbpedia.org/resource/Bangalore a dbpedia-owl:PopulatedPlace
dbpedia.org/resource/Karnataka dbpedia-owl:Capital dbpedia.org/resource/Bangalore

column header literals or other information in the table description to arrive at candidate properties and their respective domain classes. For columns containing arbitrary literals, only top-down technique is applicable. We assume that the literals used to label the column headers are relevant to the data contained in the respective columns as otherwise, it will be practically impossible to ascertain what the data is about even by humans. We combine results from the top-down and bottom-up techniques and create a consolidated graph linking columns from tables to their respective candidate classes (derived from *bottom-up technique*) using *cc* edge label (*cc* used to denote candiate class link), and candidate properties for the columns (derived from *top-down technique* and *bottom-up technique*) to their respective *domain classes* using *d* edge label (*d* used to denote link to a domain class). We use DBPedia to generate the preliminary list of domain classes for the columns and call it the *Hypothesis Set* and expand the search for candidate properties to all the equivalent classes from LOD (determined by the *owl:equivalentClass* property for each domain class of the candidate property). This way we can identify dominant concept classes for a given set of tables across LOD. We use two Abduction Reasoning Heuristics namely a) Consistency and b) Minimality to arrive at the dominant class(es) [10].

Our Scoring Model to determine the dominant classes is as follows:

1. Candidate Class Support (CCS), defines how well a class $\gamma \in \mathit{\Gamma}$ fits as a candidate class for columns across all the CSV files:

$$ccs(\gamma) = \frac{\sum f_k}{cscols(\gamma)} \tag{1}$$

2. Domain Class Support (DCS) defines how well a class $\gamma \in \mathit{\Gamma}$ corresponds to candidate properties for columns across all the CSV files:

$$dcs(\gamma) = \frac{|dscols(\gamma)|}{|cols|} \tag{2}$$

Here, Γ represents the Hypothesis Set. Each member of this class γ will have incoming edges representing one of the following: a) candidate class for some column c_k with its corresponding support f_k represented by an incoming cc link, and/or b) domain class for some property p represented by an incoming d link. $\sum f_k$ is the sum of the support from each column connected to class γ with a cc link. $cscols(\gamma)$ is the set of nodes of type column that have a path leading to γ with a cc link. Similarly $dscols(\gamma)$ is the set of nodes of type column having a path to γ with a d link. $cols$ is the set of all nodes of type column.

ccs and dcs calibrate the prolific nature of the class across all the CSV files. In addition to the above scores, a "universality score" is associated with a class that describes how prolific is this class across different tables. This score called Tabular Support (TS) is defined as:

$$ts(\gamma) = \frac{|tabs(\gamma)|}{|tabs|} \tag{3}$$

Here, $tabs(\gamma)$ is the set of nodes labeled "table" in the graph that have a path to γ via any of the labeled edges and $tabs$ is the set of nodes labeled "table" in the graph.

The class score vector for class γ is a vector representing ccs and dcs scores:

$$csv(\gamma) = [ccs(\gamma), dcs(\gamma)]$$

The overall score representing the suitability of a class γ as a domain class is defined as:

$$Score(\gamma) = \|csv(\gamma)\|_2 \cdot H[csv(\gamma)] \cdot ts(\gamma) \tag{4}$$

Here $\|csv(\gamma)\|_2$ represents the L_2 norm of the class score vector and $H[csv(\gamma)]$ represents the entropy of the class score vector, given by:

$$H[csv(\gamma)] = - \sum_{i \in ccs(\gamma), dcs(\gamma)} p_i(\gamma) \log p_i(\gamma) \tag{5}$$

From the Overall Scores for each entry in the Hypothesis set, we use a *user defined threshold* to select the dominant concept(s) for the collection of tables. Our approach uses data-driven techniques and additional evidence from column headers and looks for convergence to domain classes in the context determined by the data. This is one of the differences from the State of the Art techniques discussed in *section 2*

5 Preliminary Results

As of this writing, we have considered a variety of tabular datasets ranging from hand crafted to publicly available csv datasets including those from data.gov.in.

Table 2 shows the preliminary results from our scoring model to identify dominant concept classes from a collection of tables using a cutoff threshold at 0.75.

Table 2: Dominant Concept Classes for the various collection of tabular datasets

Tabular Datasets	Description	Dominant Concept Classes
1) StatesCapitals.csv, 2) RiversSources.csv	Arbitrary Indian States and their capitals and Prominent Indian Rivers and their Source States	http://dbpedia.org/ontology/Place http://dbpedia.org/ontology/PopulatedPlace http://schema.org/Place http://schema.org/BodyOfWater http://dbpedia.org/ontology/BodyOfWater http://dbpedia.org/ontology/River http://schema.org/RiverBodyOfWater
1) PM.csv, 2) Prez.csv	Indian Prime Ministers and Presidents	http://xmlns.com/foaf/0.1/Person http://dbpedia.org/ontology/Person http://schema.org/Person
1) TechStartupUS.csv, 2) USStateCities.csv	Details of type and location of Technology Start-up companies in the US and arbitrary US State and Cities	http://dbpedia.org/ontology/Place http://schema.org/Place http://dbpedia.org/ontology/PopulatedPlace http://dbpedia.org/ontology/Location http://dbpedia.org/ontology/Organisation http://dbpedia.org/ontology/Settlement

6 Conclusion and Future Directions

The first goal of our research objective namely *Thematic Framework* extraction, as of now has been verifed with satisfactory results on tables, where the dominant concepts are about persons, places, organisations or some identifiable concept defined in LOD. The main challenge is the ability to identify LOD entities/resources from the data accurately especially when the Information Content/Entropy in the data from columns is low. Additionally the column header may capture the essence of the properties for a domain class using words that have a similar word-sense. We would like to test and refine the algorithm on variety of tables where the data values are a combination of known LOD entities and arbitrary values. Additionally our proposal faces challenges to converge to any dominant theme when the dataset is about a complex concept such as *Rice Prices on a particular date in various districts of India* or a dataset about *Real Estate Sales Transations in a particular locality*. Such instances occur when we do not have appropriate classes/ontologies in LOD that relate to the dataset in hand or the data values in the tables do not map to any entity in the LOD. In such cases, we would like to explore the use of SKOS(Simple Knowledge Organization System) categories and Yago concept classes together with Wordnet (to address the problem of similar words that capture the essence of the column header literals) as they seem to closely relate to the purpose/context of the data. Additionally, we would like to incorporate human input/custom ontology to validate the suggestions on concepts/properties returned by the algorithms. The next step from here is to expand the *Thematic Framework* to the corresponding *Schematic Framework* (T-Box and A-Box assertions) and device an appropriate scoring algorithm for abduced properties in line with the Thematic Framework and finally generate enriched RDF.

7 Evaluation Plan

We intend to use evaluation methods measuring a) Coverage b) Accuracy c) Applicability. **Coverage** will measure the percentage of tables/columns mapped to LOD classes.

The goal is to cover as many columns in all the datasets to relevant properties from LOD in line with the Thematic Framework abduced for the datasets. **Accuracy** will compare the scores of the Dominant Concept(s) in the Thematic Framework and scores of the properties suggested by the Schematic Framework with the actual concepts/properties suggested by human evaluators. **Applicability** will measure the relevance of the newly abduced A-Box instantiations and newly inferred LOD properties for relations between column headers and the Dominant Concepts for a collection of datasets.

Acknowledgments. I am deeply grateful to my thesis advisor Prof. Srinath Srinivasa for his support.

References

1. Srinivasa, Srinath and Agrawal, Sweety V. and Jog, Chinmay and Deshmukh, Jayati: Characterizing Utilitarian Aggregation of Open Knowledge In: Proceedings of the 1st IKDD Conference on Data Sciences, pp. 6:1–6:11. ACM, New York 2014
2. Presutti, Valentina, Francesco Draicchio, and Aldo Gangemi. Knowledge extraction based on discourse representation theory and linguistic frames. In Knowledge Engineering and Knowledge Management, pp. 114-129. Springer Berlin Heidelberg, 2012.
3. Han, Lushan, Tim Finin, Cynthia Parr, Joel Sachs, and Anupam Joshi. RDF123:from Spreadsheets to RDF. Springer Berlin Heidelberg, 2008.
4. Mulwad, Varish Vyankatesh. TABEL - A Domain Independent and Extensible Framework for Inferring the Semantics of Tables. PhD diss., University of Maryland, 2015.
5. Buitelaar, Paul, Philipp Cimiano, and Bernardo Magnini. Ontology learning from text: An overview. Vol. 123. 2005.
6. Lau, Raymond YK, Jin Xing Hao, Maolin Tang, and Xujuan Zhou. Towards context-sensitive domain ontology extraction. In System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on, pp. 60-60. IEEE, 2007.
7. Gerber, Daniel, Sebastian Hellmann, Lorenz Buhmann, Tommaso Soru, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. Real-time RDF extraction from unstructured data streams. In The Semantic Web - ISWC 2013, pp. 135-150. Springer Berlin Heidelberg, 2013.
8. Augenstein, Isabelle, Sebastian Pado, and Sebastian Rudolph. Lodifier: Generating linked data from unstructured text. In The Semantic Web: Research and Applications, pp. 210-224. Springer Berlin Heidelberg, 2012.
9. Lalithsena, Sarasi, Pascal Hitzler, Amit Sheth, and Paril Jain. Automatic domain identification for linked open data. In Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on, vol. 1, pp. 205-212. IEEE, 2013.
10. Asha Subramanian, Srinath Srinivasa, Pavan Kumar, and S. Vignesh. 2015. Semantic Integration of Structured Data Powered by Linked Open Data. In Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics (WIMS '15) ACM, New York, NY, USA.