Automated Mechanisms to Discover and Integrate Data from Web-based Tabular Collections

Edwin Yaqub, David Arnu, Ralf Klinkenberg – RapidMiner, Germany
Annalisa Gentile, Chris Bizer, Heiko Paulheim – Universität Mannheim, Germany

Relevance & Research Question

Keywords: *data search, integration, enrichment*

Growth in data is phenomenal in recent years. Data exists on public sources such as the web or corporate intranet sources. However, in order to effectively utilize the information contained in this data, supporting tools are needed to discover new data and integrate it to existing datasets. This raises several research questions:

- How to access heterogeneous data that resides on the web.
- How to discover data that strongly matches an entity-query (comprising of an existing table and attributes to be extended).
- How to incorporate integration and fusion algorithms to control the discovered data (tabular collections) within data mining processes.

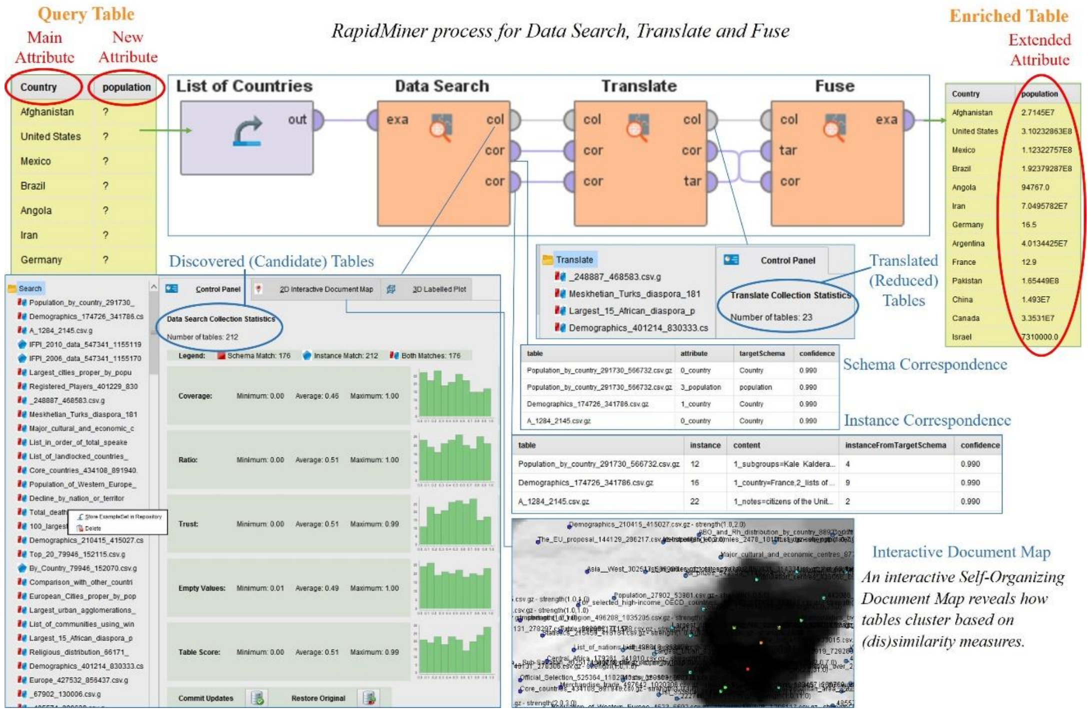
Methods & Data

Keywords: schema & instance matching, text corpus analysis, clustering

In project DS4DM [1], we developed Data Search extension for RapidMiner (an open source and free data mining tool). The extension provides Search-Join [2] as a data mining process for automated discovery and integration of tabular data.

A collection of half a million Wikipedia data tables is extracted and indexed in Lucene search engine for efficient retrieval. On this, the discovery algorithms compute schema and instance level matches for given query [3]. The result set is a space of candidate tables. The extension enables analyst to:

- Translate and fuse new data to existing tables using provided algorithms, or
- Perform manual refinements through i) exploratory visualizations (expose patterns among the vast corpus of tabular collections) and ii) graphical controls to manipulate intermediate outcomes of Search-Join process in real time (e.g. remove noisy tables) to prevent loss of valuable information.



RapidMiner process for Data Search, Translate and Fuse

Query Table

Main Attribute · New Attribute

Country	population
Afghanistan	?
United States	?
Mexico	?
Brazil	?
Angola	?
Iran	?
Germany	?

Enriched Table

Extended Attribute

Country	population
Afghanistan	2.7145E7
United States	3.10232863E8
Mexico	1.12322757E8
Brazil	1.92379287E8
Angola	94767.0
Iran	7.0495782E7
Germany	16.5
Argentina	4.0134425E7
France	12.9
Pakistan	1.65449E8
China	1.493E7
Canada	3.3531E7
Israel	7310000.0

Discovered (Candidate) Tables

Search
Population_by_country_291730_
Demographics_174726_341786.cs
A_1284_2145.csv.g
IFPI_2010_data_547341_1155119
IFPI_2006_data_547341_1155170
Largest_cities_proper_by_popu
Registered_Players_401229_830
_248887_468583.csv.g
Meskhetian_Turks_diaspora_181
Major_cultural_and_economic_c
List_in_order_of_total_speake
List_of_landlocked_countries_
Core_countries_434108_891940
Population_of_Western_Europe_
Decline_by_nation_or_territor
Total_death
100_largest
Store ExampleSet in Repository
Delete
Demographics_210415_415027.cs
Top_20_79946_152115.csv.g
By_Country_79946_152070.csv.g
Comparison_with_other_countri
European_Cities_proper_by_pop
Largest_urban_agglomerations_
List_of_communities_using_win
Largest_15_African_diaspora_p
Religious_distribution_66171_
Demographics_401214_830333.cs
Europe_427532_856437.csv.g
_67902_130006.csv.g
Control Panel
2D Interactive Document Map
3D Labelled Plot

Data Search Collection Statistics

Number of tables: 212

Legend: ▨ Schema Match: 176 ● Instance Match: 212 ▨ Both Matches: 176

Coverage:	Minimum: 0.00	Average: 0.46	Maximum: 1.00
Ratio:	Minimum: 0.00	Average: 0.51	Maximum: 1.00
Trust:	Minimum: 0.00	Average: 0.51	Maximum: 0.99
Empty Values:	Minimum: 0.01	Average: 0.49	Maximum: 1.00
Table Score:	Minimum: 0.00	Average: 0.51	Maximum: 0.99

Commit Updates
Restore Original
Translate
_248887_468583.csv.g
Meskhetian_Turks_diaspora_181
Largest_15_African_diaspora_p
Demographics_401214_830333.cs
Control Panel

Translate Collection Statistics

Number of tables: 23

Translated (Reduced) Tables

Schema Correspondence

table	attribute	targetSchema	confidence
Population_by_country_291730_566732.csv.gz	0_country	Country	0.990
Population_by_country_291730_566732.csv.gz	3_population	population	0.990
Demographics_174726_341786.csv.gz	1_country	Country	0.990
A_1284_2145.csv.gz	0_country	Country	0.990

Instance Correspondence

table	instance	content	instanceFromTargetSchema	confidence
Population_by_country_291730_566732.csv.gz	12	1_subgroups=Kale Kaldera...	4	0.990
Demographics_174726_341786.csv.gz	16	1_country=France,2_lists of ...	9	0.990
A_1284_2145.csv.gz	22	1_notes=citizens of the Unit...	2	0.990

Demographics_210415_415027.csv.gz - strength(1.0,2.0)
The_EU_proposal_144129_286217.csv
Major_cultural_and_economic_centres_87

Interactive Document Map

An interactive Self-Organizing Document Map reveals how tables cluster based on (dis)similarity measures.

Results:

- Domain-independent data enrichment solution to harness web tables.
- Quality of search results is evaluated on the basis of various statistical metrics such as coverage, trust, ratio, empty values, table score.

Added Value:

- RapidMiner platform has been extended to incorporate data discovery and integration methods as first class citizens of data mining processes.

References
[1] http:// ds4dm.de
[2] Extending RapidMiner with data search and integration capabilities. Gentile, Anna Lisa, et al., International Semantic Web Conference, 2016
[3] Entity Matching on Web Tables: a Table Embeddings approach for Blocking. Gentile, Anna Lisa, et al., (accepted submission), International Conference on Extending Database Technology, 2017