Contribution ID : **28** Type : **Demonstration**

Realizing smart data by automating tabular search, integration and extraction methods

Thursday, 12 October 2017 12:15 (30)

There is an increased focus on deriving business value from data. To exploit data, it is usually necessary to fetch it out of various silos to get a clear and holistic picture for the domain of discourse. The challenges lie in merging diverse data sets so data mining tasks can be performed on enriched data sets, which better represent the specific use case at hand. In current art, there is a lack of software tools and methods that assist in curating data from heterogenous sources to realize smarter transformations compared to their archaic forms. The problem is complicated because data exists in various formats and lies on public or private sources. These inadequacies block progress in extracting value out of data. To resolve these shortcomings, we are researching and developing user interface based tools and scientific methods in the ongoing research project "Data Search for Data Mining (DS4DM)" [1]. More concretely, we use structured data tables as basis to address *data extraction* and *data enrichment* by developing extensions for the open-source data science platform "RapidMiner". Thus, the presented work leads to smart exploitations of data in graphically designed and highly reusable data mining processes.

Data Extraction: Our work towards data extraction enables data scientists to conveniently retrieve data tables from popular sources directly into RapidMiner. These include Wikipedia articles, websites, PDF documents, online (Google) spreadsheets, etc. This adds to the out-of-box features already available in RapidMiner, which allow for reading spreadsheet documents, CSV and XML files, RDBMS databases, Cloud storage (Amazon S3) and NoSQL data-stores such as Cassandra and MongoDB, etc.

Data Enrichment: Our work towards data enrichment implements the Search-Join [2] method within the graphical user interface of RapidMiner. Search-Join is a two-folds structured (tabular) search method. First, relevant data tables are searched from potentially large tabular corpus for a provided query. The tabular corpus serves as a reliable data store. Our current prototype consists of half a million data tables extracted from Wikipedia, but data can be ingested from organizational data-stores as well. The search query is comprised of an existing data table and an additional attribute, which needs to be discovered. This way, new tabular columns can be discovered and appended with imprecise or vague knowledge such as text keywords.

The query is resolved by discovery algorithms that compute schema (table's header) level and instance (table's row) level matches for the query. This returns a space of candidate tables, which have strong contextual resemblances with the query and hence may add value to the original data table. As data search is susceptible to noise, the potentially large number of discovered tables need to be refined. This introduces data integration challenges, namely i) manual integration and ii) automatic integration of data, without which the practicality of discovered results remains of lesser value.

- *Manual Data Integration:* Ideally, the data scientist needs to manually examine results and remove noisy tables, so that only value-contributing tables are considered. To guide the human in removing noise but preventing loss of informative tables, exploratory visualization techniques are developed, e.g., i) A Self-Organizing Document Map reveals how tables cluster based on similarity measures and ii) Graphical controls to manipulate intermediate outcomes of Search-Join process in real time i.e. removing noisy tables or observing distributions of certain statistical metrics among discovered candidate tables, which help to understand resemblances.
- *Automatic Data Integration:* This option allows to execute the Search-Join as a fully automated process using default options, so extraction and enrichment can be operationalized.

Results

The presented work implements a domain-independent solution to realizing smart data (through extraction and enrichment). The quality of search results considers statistical metrics such as coverage, trust, ratio and empty values, which are useful for data integration. As a result, the RapidMiner platform has been extended to incorporate data discovery and integration methods in data mining processes. Four extensions are developed and made publicly available at [3].

Acknowledgements

This work is sponsored by the German ministry of education and research (BMBF) under grant agreement number 01IS15027A-B.

[1] DS4DM project website, weblink: http://ds4dm.de
[2] The Mannheim Search Join Engine, C. Bizer et al., Web Semantics: Science, Services and Agents on the W
[3] The RapidMiner Marketplace, weblink: https://marketplace.rapidminer.com/UpdateServer/faces/index.xhtml

Track

SDIC

Primary author(s) : Mr. ARNU, David (Rapidminer GmbH); Dr. YAQUB, Edwin (RapidMiner GmbH); Mr. SCHLUNDER, Phillipp (RapidMiner GmbH); Mr. KLINKENBERG, Ralf (RapidMiner GmbH)

Presenter(s) : Dr. YAQUB, Edwin (RapidMiner GmbH)