The Information Profiling Approach for Data Lakes

Ayman Alserafi; Alberto Abelló; Oscar Romero; Toon Calders

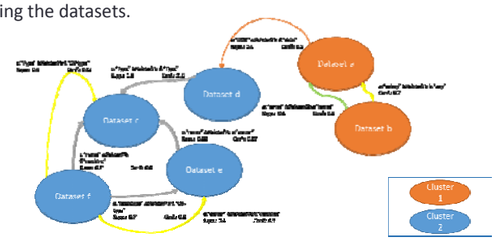
KEY GOAL & DEFINITIONS

Our goals is to be able to profile **multi-structured textual datasets** in **the Data Lake** (DL) for identifying relationships and information overlaps between them using **information profiling** techniques. This supports in effective *data governance* of DLs.

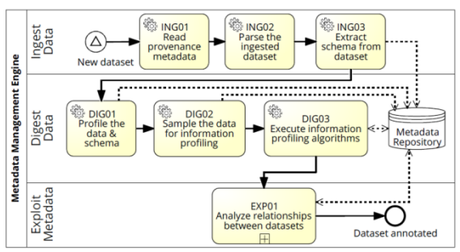
Data Lake: A repository for storing raw data in their original formats for Business Intelligence purposes.

Multi-structured Datasets: A collection of differently structured datasets of instances stored in a common structured / semi-structured schema (e.g. CSV, JSON, XML, etc.).

Information Profiling: Finding relationships between data items stored in different multi-structured datasets using schema matching techniques and metadata describing the datasets.

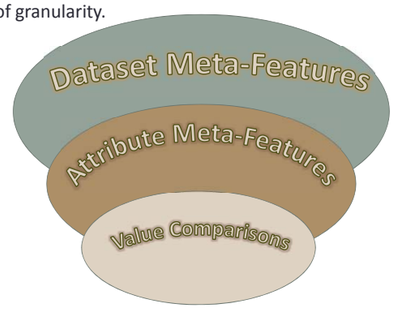


OUR FRAMEWORK



THE APPROACH

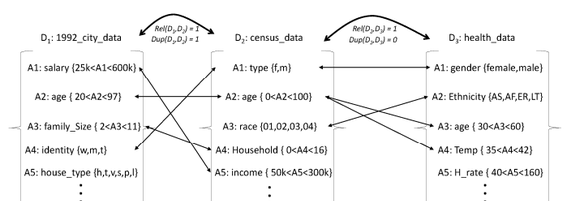
We aim to improve the efficiency of the information profiling task by applying a layered approach which encompasses increasing levels of computational complexity. Each layer filters out non-necessary comparisons using specific metadata collected at different levels of granularity.



DATASETS META-FEATURES MATCHING

Type	Meta-feature	Description
General	Number of Instances	The number of instances in the dataset
	Number of Attributes	The number of attributes in the dataset
	Dimensionality	The ratio of number of attributes to number of instances
Attributes by Type	Number per Type	The number of attributes per type (Nominal or Numerical)
	Percentage per Type	The percentage of attributes per type (Nominal or Numerical)
Nominal Attributes	Average Number of Values	The average number of distinct values per nominal attribute
	Standard Deviation of Number of Values	The standard deviation in the number of distinct values per nominal attribute
	Minimum/Maximum Number of Values	The minimum and maximum number of distinct values per nominal attribute
Numeric Attributes	Average Numeric Mean	The average of the means of all numeric attributes
	Standard Deviation of the Numeric Mean	The standard deviation of the means of the numeric attributes
	Minimum/Maximum Numeric Mean	The minimum and maximum mean of numeric attributes
Missing Values	Missing Attribute Count	The number of attributes with missing values
	Missing Attribute Percentage	The percentage of attributes with missing values
	Minimum/Maximum Number of Missing Values	The minimum and maximum number of instances with missing values per attribute
	Minimum/Maximum Missing Values Percentage	The minimum and maximum percentage of instances with missing values per attribute
	Mean Number of Missing Values	The mean number of missing values from each attribute
	Mean Percentage of Missing Values	The mean percentage of missing values from each attribute

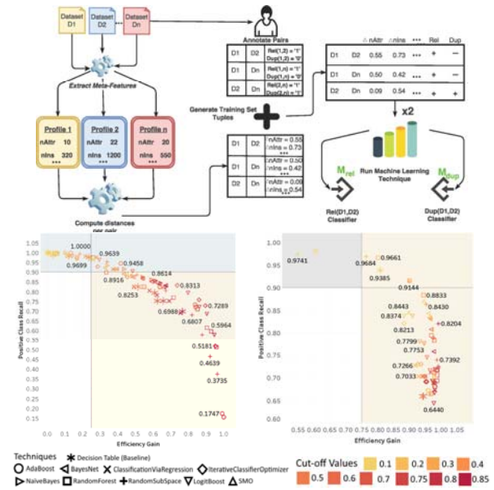
$$dist_{mi}(d_1,d_2) = \frac{\max(m_i(d_1),m_i(d_2)) - \min(m_i(d_1),m_i(d_2))}{\max(m_i(d_1),m_i(d_2))}$$



EARLY PRUNING BASED ON DATASETS META-FEATURES

We use supervised machine learning techniques to use meta-features describing the datasets for predicting whether a pair is a possible candidate for:

1. Schema matching: the pair has related data items and information
2. De-duplication: the pair has identical data items and information



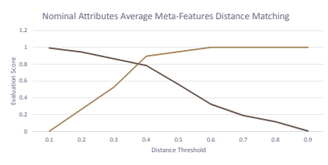
ATTRIBUTES META-FEATURES MATCHING

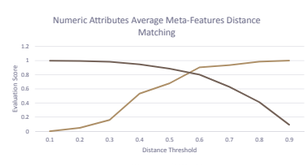
Nominal attributes

- Number of distinct values
- Percentage of missing values
- Value distribution percentage

Numeric attributes

- Number of distinct values
- Percentage of missing values
- Min,Max,Avg,STDEV
- Range, Coefficient of variance





CONCLUSIONS

- We propose an effective and efficient stratified approach for information profiling in the DL
- Efficiency gains using early-pruning techniques can support in large-scale environments

FUTURE WORK

- Implementing an efficient technique for value comparisons between attributes in the datasets
- Handling different varieties of multi-structured data within our framework

REFERENCES

1. S. Kruse, et al., Data Anamnesis : **Admitting Raw Data into an Organization**, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering (2016)
2. E. Rahm, **The Case for Holistic Data Integration**, in: East European Conference on Advances in Databases and Information System (ADBIS), Springer, 2016.
3. Alserafi, A., Abelló, A., Romero, O., & Calders, T. (2016). **Towards Information Profiling: Data Lake Content Metadata Management**. In *Data Mining Workshops (ICDMW), IEEE*
4. P. a. Bernstein, J. Madhavan, E. Rahm, **Generic Schema Matching , Ten Years Later**, Proceedings of the VLDB Endowment 4 (11) (2011) 695 - 701.
5. I. Terrizzano, P. Schwarz, M. Roth, J. E. Colino, **Data Wrangling: The Challenging Journey from the Wild to the Lake**, in: 7th Biennial Conference on Innovative Data Systems Research CIDR'15, 2015.