Photo by Maksym Kaharlytskyi on Unsplash

Approximate Data Profiling

Prof. Dr. Felix Naumann
Tobias Bleifuß
Leon Bornemann
Youri Kaminsky

Summer 2022

Introduction: Data Profiling

"Data profiling is the process of examining the data available from an existing information source (e.g. a database or a file) and collecting statistics or informative summaries about that data" [0]

- Metadata is very helpful
- Metadata is often incomplete, outdated or completely unavailable
- → discover useful metadata, given just the (relational) dataset
- Our Focus: Unique Column Combinations (UCCs), Functional Dependencies (FDs) and Inclusion Dependencies (INDs)

Prof. Dr. Felix Naumann
Tobias Bleifuß
Leon Bornemann
Youri Kaminsky

[0] https://en.wikipedia.org/wiki/Data_profiling

Introduction: Unique Column Combinations (UCCs)

ID	Student ID	First Name	Last Name	Birthday
1	457825	John	Doe	14.02.1966
2	127894	John	Doe	13.11.1990
3	487992	Jim	Smith	17.04.1993
…	…	…	…	…

Unique Column Combination R.X:

For a set of attributes X in Relation R, X is a UCC, if for all t ∈ R there exists no t' ∈ R, so that t[X] = t'[X].

Example UCCs:

- {ID}
- {Student ID}
- {First Name, Last Name, Birthday}

Prof. Dr. Felix Naumann
Tobias Bleifuß
Leon Bornemann
Youri Kaminsky

Introduction: Functional Dependencies (FDs)

ID	Company	Street	Zip-Code	City
…	…	…	12307	Berlin
…	…	…	12307	Berlin
…	…	…	12359	Berlin
			23966	Wismar

ID	First Name	Last Name	Birthday
1	John	Doe	14.02.1966
2	John	Doe	13.11.1990
3	Jim	Smith	17.04.1993

Functional Dependency X → A:

For sets of attributes X and A the FD X → A holds if:
$t_1[X] = t_2[X] \Rightarrow t_1[A] = t_2[A]$.

Example FDs:

- Zip-Code → City
- First Name, Last Name, Birthday → ID

Prof. Dr. Felix Naumann
Tobias Bleifuß
Leon Bornemann
Youri Kaminsky

Introduction: Inclusion Dependencies (INDs)

Object	Density	Semi-major Axis
Mercury	…	…
Venus	…	…
Earth	…	…
Saturn	…	…

Name	Satellite of	Distance [km]
…	Saturn	…
…	Saturn	…
…	Earth	…

Inclusion Dependency X ⊆ Y:

For sets of attributes $R_1.X=\{X_1,...,X_n\}$ and $R_2.Y=\{Y_1,...,Y_n\}$ the IND $R_1.X \subseteq R_2.Y$ holds if $\Pi_X(R_1) \subseteq \Pi_Y(R_2)$

Example INDs:

- Satellite of ⊆ Object

Prof. Dr. Felix Naumann
Tobias Bleifuß
Leon Bornemann
Youri Kaminsky

Data Profiling Algorithms: Basic Strategies

Automatic discovery algorithms

- Typically discover minimal FDs/UCCs and maximal INDs
- Search space is still exponentially large
- → Employ pruning strategies

Examples for pruning:

- If {A,B} is no UCC, then neither {A} nor {B} can be UCCs
- If AB ↛ C then we also know that A ↛ C and B ↛ C

Data Profiling: Using Sampling

Motivation

- Runtime: even with state of the art algorithms, the discovery of dependencies can be infeasible for very large datasets
- Privacy: for data protection reasons, we may not access all data
- Cost: to collect or purchase all data might be too expensive
- Data quality: our database may be incomplete

Problem Statement

- Discover dependencies on a subset / sample

Research questions

- How close can we get to the true dependencies?
- What properties must our subset / sample have, so that we get reliable results?
- What is the tradeoff between sample size and correctness of results?

Approximate Data Profiling: Related Work

State of the art dependency discovery algorithms on full data sets:

- Functional Dependencies: HyFD [1]
- Inclusion Dependencies: BINDER [2]
- Unique Column Combinations: HyUCC [3]

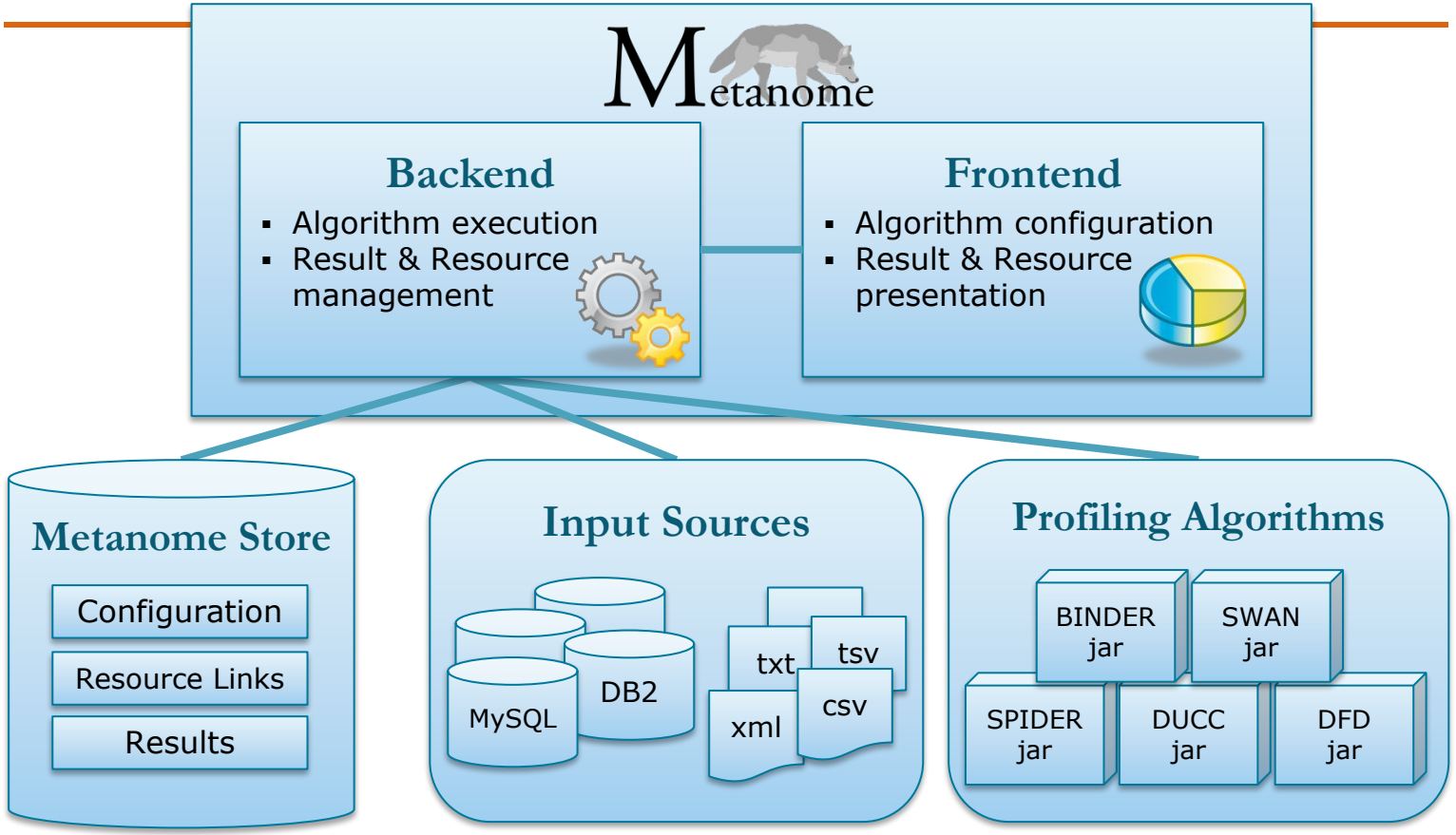
Sampling/Sketching has already been applied to reduce runtime:

- AIDFD [4]
- FAIDA [5]

[1] Papenbrock, Thorsten, and Felix Naumann. "A hybrid approach to functional dependency discovery." Proceedings of the 2016 International Conference on Management of Data. 2016.
[2] Papenbrock, Thorsten, et al. "Divide & conquer-based inclusion dependency discovery." Proceedings of the VLDB Endowment 8.7 (2015): 774-785.
[3] Papenbrock, Thorsten, and Felix Naumann. "A hybrid approach for efficient unique column combination discovery." Datenbanksysteme für Business, Technologie und Web (BTW 2017) (2017).
[4] Bleifuß, Tobias, et al. "Approximate discovery of functional dependencies for large datasets." Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016.
[5] Kruse, Sebastian, et al. "Fast approximate discovery of inclusion dependencies." Datenbanksysteme für Business, Technologie und Web (BTW 2017) (2017).

Prof. Dr. Felix Naumann
Tobias Bleifuß
Leon Bornemann
Youri Kaminsky

Metanome



Prof. Dr. Felix Naumann
Tobias Bleifuß
Leon Bornemann
Youri Kaminsky

Metanome

http://www.metanome.de

https://owncloud.hpi.de/s/WSmgz18Z6yceG7h

Prof. Dr. Felix Naumann
Tobias Bleifuß
Leon Bornemann
Youri Kaminsky

Metanome Resources

Repositories:

- Metanome https://github.com/HPI-Information-Systems/metanome
- Algorithms https://github.com/HPI-Information-Systems/metanome-algorithms
- Metanome-CLI https://github.com/sekruse/metanome-cli

Datasets:

- UCC/FD discovery: https://hpi.de/naumann/projects/repeatability/data-profiling/fds.html#c168191
- IND discovery: https://hpi.de/naumann/projects/repeatability/data-profiling/metanome-ind-algorithms.html

Goals and grading

Goals

- Learn about the research area data profiling
- Read papers and understand them
- Craft a novel solution to the problem of sample-based profiling
- Run experiments and evaluate results
- Present results in written and oral form

Grading

- Approach (35%)
- Written report (35%)
- Presentations and discussions in the seminar (30%)

Requirements

You **bring**:

- Java (at least basic skills or willing to learn)

You want to **learn** about:

- Data Profiling
- Algorithmic problems and how to solve them
- Experimental evaluation on large datasets

You do **not** need to bring:

- Prior knowledge about Data Profiling

You should **agree** to:

- Publish the results as open-source

Seminar Roadmap (preliminary)

03.05.2022
- Technical introduction
- Introduce Metanome and algorithms repository

10.05.2022
- Present related work (1 previously assigned paper)
- Report initial execution of related work (modify input data)
- Potentially: Initial ideas for a sampling-based variant

17.05.2022
- Discuss metrics for evaluation
- Discuss initial ideas / general approaches

24.05.2022
- Implement evaluation metrics and evaluate related work on sample

14.06.2022
- Milestone: Technical Evaluation
- Present preliminary results

19.07.2022
- Final presentation (chair is invited)

29.07.2022
- Deadline paper-style submission (max 12 pages)
- If preferred, we can postpone the deadline until after the exam periods

Every week we offer to meet for status updates, questions and feedback

Register for the seminar

Structure:

- Work in teams of **two** students

Register:

- Email to youri.kaminsky@hpi.de
- Use subject: **Approximate Data Profiling seminar registration**
- Optionally, include topics that are interesting for you
- Optionally, include team partner
- Optionally, include another time slot that fits your schedule better