Theme : Data Profiling and Curation

- Sub Theme : Data Profiling and Curation

Currently, silo and heterogeneous data is integrated manually based on different understanding of each person based on traditional ETL(Extract-Transform-Load) method. The method is not able to handle large-sized data and does not ensure scalability.

Development a platform (set of methods) for semi-automated data curation that efficiently integrates Samsung's siloed data

- Data curation requires identifying of potential relationships among schema/entities in heterogeneous data sources that need to be integrated in a data-driven approach
- A data-driven approach is for substantially eliminating the necessity of human's understanding of the target data and intervention into the processes.

Specification of the data curation

- Module for Data ingestion
  - Hundreds of heterogeneous data sources in different forms (i.e. Hive,
  - Oracle, MySQL, PostGres, Excel, CSV, TXT…etc.)
  - A configuration file (IP, DBSpace, User Account) that indicates data sources is input
- Module for Data cleansing
  - Some errors in data types, NaN, Inf and/or missing values can exist
  - Description of the data is limitedly available
- Module for Analysis I - Identification of potential schema mapping
  - Finding the same columns with different column name and data type
    e.g. [Address] column in Table A in DB 1 may be the same with [ADDR] column in Table B in DB2 with score (confidence) level of 0.8

e.g. [Family name] column + [Given name] column in Table A in DB1 may be equivalent to [USRNAME] in Table B in DB2 with score (confidence) level of 0.78

- Machine learning or statistical methods could be devised
- Proper threshold level should be provided for automatically accepting the finding

- Module for Analysis II - Identification of potential entity mapping
  - Finding the same records (entities)
    e.g. ‘Barack Obama’ in Table A in DB1 may be equivalent to ‘Mr. Obama, Barack’ in Table B in DB2 according to the field contents associated with those entities with score (confidence) level of 0.91
  - Finding the subsumption records
    e.g. ‘Korea’ in Table A in DB1 may be subset of ‘Asia’ in Table B in DB2 with score (confidence) level of 0.99
  - Machine learning or statistical methods could be devised
  - Proper threshold level should be provided for automatically accepting the finding
- Module for Human confirmation on the identified mappings (w/ UI)
  - Proper visual representation on the identified mappings (e.g. Graph-representation)
  - Human should be able to manually check the identified mappings and accept/reject the finding
  - Output: Graph-structured map of identified schema/entity mappings
- Module for Updating process (w/ UI)
  - The steps above should re-run upon any necessity of update is sensed

Test dataset / Testbed

- Samsung’s internal data (TBD) will be provided for the development and performance test
- Cloud instances (e.g. Amazon) can be used for developmental period

Evaluation criteria

- Accuracy of the machine suggested mappings (F1-score >= 0.85)
  - Comparison with the ground truth mappings
- Scalability

- Independency of number, type and location of target data sources
- Independency of size of data in each data sources

Assumptions

- Not like other conventional tools for data integration, users of the platform does not have to know the primary, foreign and other keys for join operations
- The platform should be scalable enough to handle hundreds data sources
- Open source tools can be used

※ The topics are not limited to the above examples and the participants are encouraged to propose original idea.

※ Funding : Up to USD $200,000 per year