7a.003.UL – Ontology-based Fast Semantic Indexing for Structured and Unstructured Data in Health Care

Satya Katragadda[1], Raju Gottumukkala[1], Vijay Raghavan[1], Adeola Siwoku[1], Michael Lucito[2], Chris Cotteleer[2]
University of Louisiana at Lafayette[1], Schumacher Clinical Partners[2]

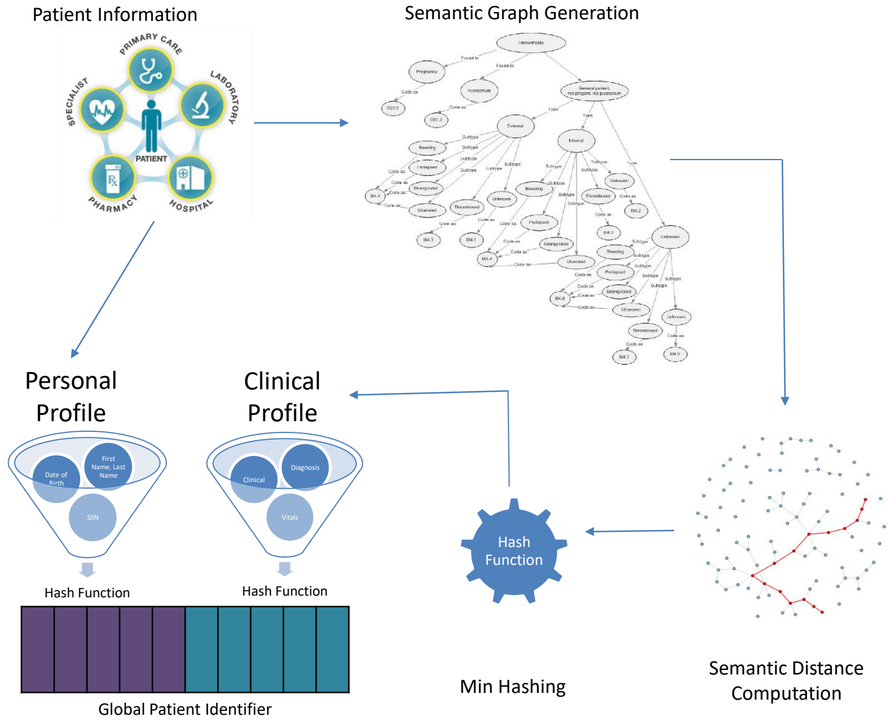
Project Start: 10/01/2018	End Date: 07/31/2019	Project Budget: 40K	Spent: ~12k

Project Summary:

In the current big data environment, most of the data is gathered from multiple sources. Entity resolution or duplication of data is a major problem in this scenario. This duplicate data is more pronounced in patient data from health care. Recent studies indicate that about 15% of the Master Patient Index of major hospitals are duplicate entries. Issues like heterogeneous data, incomplete information, constantly changing properties associated with entities, and temporal information pose major challenges to identifying duplicate entities in the data. To solve this problem, we propose a indexing technique that identifies duplicate information from databases using ontology based semantic measures. The proposed approach generates a global identifier for each entity based on the distances of the properties associated with the entity to core nodes within the semantic graph extracted from the ontology. Partial and complete match algorithms will be applied on the global identifier to identify duplicate records. The identifier can be updated based on changes to the properties associated with the entity. Our project proposes a proof of concept to identify duplicate records in a Master Patient Index that indexes the data using a global patient identifier that is based on the demographic and clinical profile of the patient. We aim to significantly improve the performance of the deduplication algorithm over the traditional baseline algorithms.

Details of Progress/Achievements:

- Surveyed various deduplication, instance matching, entity co-reference, and entity matching techniques using ontologies
- Built an semantic graph using ICD10 Ontology and BBC Core Concepts Ontology
- Identified and implemented various structural similarity measures to measure similarity between different entities
- Currently, working on generating semantic data using LANCE developing approaches to identify sub-components of the ontology graph



PROJECT DELIVERABLES

Deliverable	Achievements	Remaining To Do
1. Investigate various instance matching based on entity recognition, record linkage, and entity co-reference approaches in current literature	80% complete.	The team will continue reviewing the literature
2. Develop a global identifier for each instance based on the properties or features associated with that instance	50% complete. Implemented semantic distance measures that can be used to build a global profile of a user	Develop techniques to identify core nodes in the graph to represent the instance
3. Design a blocking technique that identifies the matching between two instances based on the global identifier	30% complete. Implemented the max and min techniques to map the global profile and match instances	Extend the global profile to be more representative for temporal data
4. Build a prototypical system for healthcare data to identify duplicate entries in a Master Patient Index	0%. Planned for March, 2019	The team will implement the prototypical system after the approach is tested on synthetic data