Towards Heterogeneous Keyword Search

Chunbin Lin, Jianguo Wang, Chuitian Rong

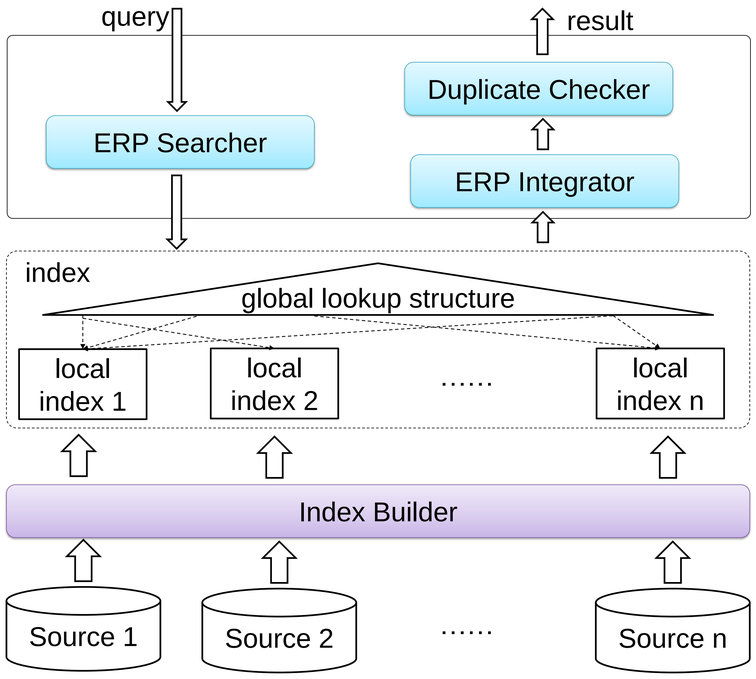
Motivation

- Data is usually resident in heterogeneous data sources including unstructured data, semi-structured data and structured data.
- Existing keyword search systems are designed and tuned for one specific data model. They cannot answer heterogeneous keyword queries.

Contribution

Build a **heterogeneous keyword search system** that performs keyword queries upon diverse data sources rather than just one type of data source.

System Architecture



Challenge 2 : Ranking Function

- final answers are integrated from different data sources
- each data source has its own features

→ New ranking functions

A **local ranking function** for each kind of data source

A **global ranking function** to compute the scores for final ERPs.

Challenge 3 : Index Structure

To support efficient heterogeneous keyword search over diverse data sources

→ New index structure

Local lookup structure. Each data source has a local index. It is an inverted index with keywords as keys and the (ERP, score) pairs as values.

Global lookup structure. A hash table with keywords as keys and points to local indexes as values.

Challenge 4 : Top-k Query Processing

Existing top-k algorithms, e.g., TA and NRA, cannot be applied for the heterogeneous keyword search problem

→ Design a new top-k algorithm

Challenge 5 : Fuzzy Mapping

Answers from different data sources may contain duplicate attributes and entities

→ String similarity measure

Syntactic similarities
- Token-based similarity
- Character-based similarity

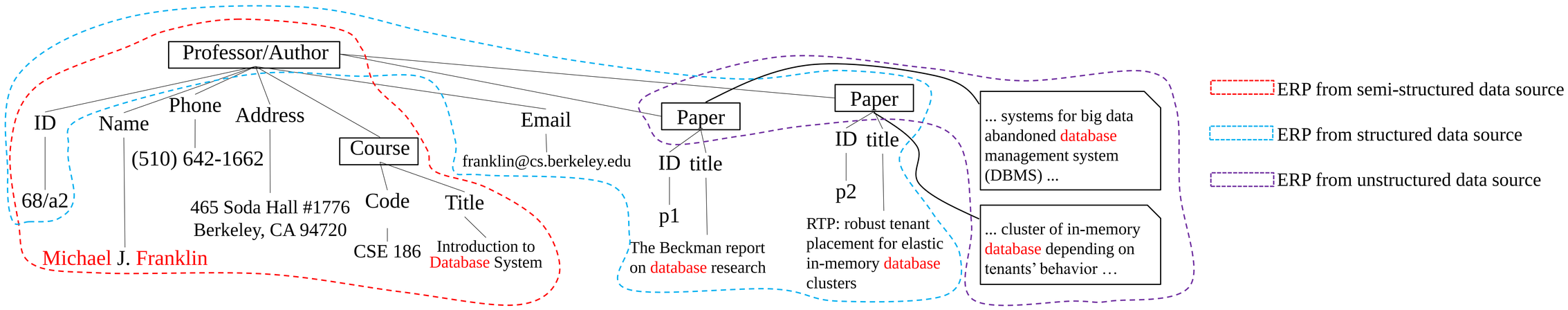
Semantic similarities
Apply synonym rules to evaluate the maximal similarities

Challenge 1 : Unified Result Format

Heterogeneous keyword search system integrates partial answers from different data sources

→ Need to define an unified result format. (1) It should be powerful enough to express answers in unstructured data, semi-structured data and structured data; (2) It should capture entitles and the relationships among entities

→ Entity-relationship pattern (ERP)



An ERP of query "michael, franklin, database", which is constructed by three ERPs from unstructured data (highlighted in purple), semi-structured data (highlighted in red), and structured data (highlighted in blue)