Open Data Link

A dataset search engine for open data

Paul Ouellette and Justin Fargnoli

Open Data Link

- Dataset search engine for open data.
- Search methods:
  - Semantic keyword search
  - Joinable table search
  - Unionable table search

Motivation

- Governments and other organizations publish a lot of open data, but discovery is still difficult.
- Data scientists can identify ways to integrate datasets.
- Data publishers can see the wider context of their data.

Demo

Outline

System overview

Joinable table search

Unionable table search

Semantic keyword search

Future work

Outline

System overview

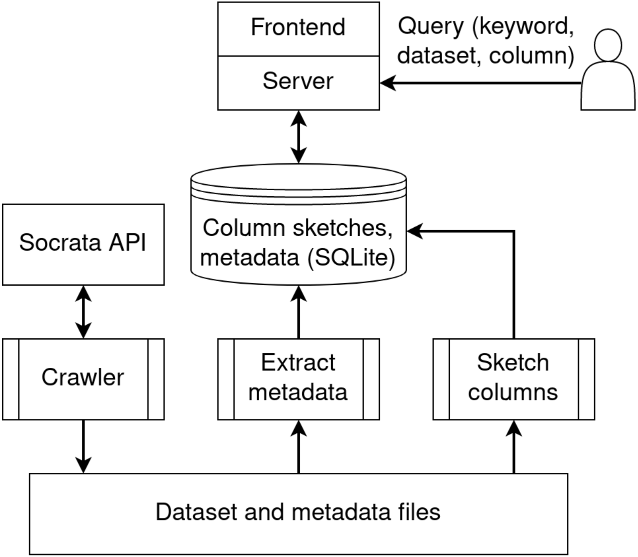
Joinable table search

Unionable table search

Semantic keyword search

Future work

System overview


Frontend
Server
Query (keyword, dataset, column)
Column sketches, metadata (SQLite)
Socrata API
Crawler
Extract metadata
Sketch columns
Dataset and metadata files

Dataset crawl

- 10k of 42k datasets on Socrata.
- 172k columns.
- Most datasets are small.
- Largest datasets have over 100 million rows.

Outline

System overview

Joinable table search

Unionable table search

Semantic keyword search

Future work

Minhash[2]

- Data sketch for estimating Jaccard similarity of sets.

$$J(S,T) = \frac{|S \cap T|}{|S \cup T|}$$

- A minhash signature is composed of the results of a number of minhashes.
- The probability that the minhashes for two sets are the same equals the Jaccard similarity of the sets[1].
- Minhash LSH hashes similar signatures to the same bucket.

[1] Mining of Massive Datasets, Chapter 3.

[2] A. Broder, "On the Resemblance and Containment of Documents", Compression and Complexity of Sequences 1997.

LSH Ensemble[3]

- Set containment is a better measure for computing joinability.

$$C(Q, X) = \frac{|Q \cap X|}{|Q|}$$

- We can convert Jaccard similarity to containment, given the sizes of the domains.
- The size of the indexed domain is not constant, so domains are partitioned by cardinality.
- A minhash LSH index is constructed for each partition.

[3] Erkang Zhu, Fatemeh Nargesian, Ken Q. Pu, Renée J. Miller, "LSH Ensemble: Internet-Scale Domain Search", VLDB 2016.

Outline

System overview

Joinable table search

Unionable table search

Semantic keyword search

Future work

Unionable table search

- The LSH Ensemble index is queried for each column of the query table.
- Candidate tables are those that appear in $\geq 40\%$ of the joinability queries.
- Candidates are ranked by alignment: the fraction of candidate columns that are unionable with a query column.

Outline

System overview

Joinable table search

Unionable table search

Semantic keyword search

Future work

Overview

- FastText: words -> vectors
- SimHash: vectors -> bit vectors
- LSH: similarity search on bit vectors

FastText

- Vectors represent the semantics of words
- Closer a pair of vectors, closer the semantics of the two words
- closeness or similarity of vectors := Cosine-Similarity

Simhash

- Vector of floats -> Vector of bits

hash := an array of length H For vector with dimension d: Compute wether it is above or below d hyperplanes H times

SimHash LSH

- L hash tables of bit vectors
- Query each L hash table for M candidates
- Compute cosnine similarity of unhashed vectors to return top-M results

LSH Forest

- Prefix Tree of bit vectors
- Variable length hash in tree solves tunability probelm
- Query L Prefix Trees (the LSH Forest) for M candidates
- Compute cosnine similarity of unhashed vectors to return top-M results

Outline

System overview

Joinable table search

Unionable table search

Semantic keyword search

Future work

Future work

- Organizing datasets into a directory structure for navigation.
- Use semantic similarity of attribute names in unionable table search.
- Similar dataset search based on metadata similarity.
- Keyword search over data values.