

Open Data Link

A dataset search engine for open data

Paul Ouellette and Justin Fagnoli

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

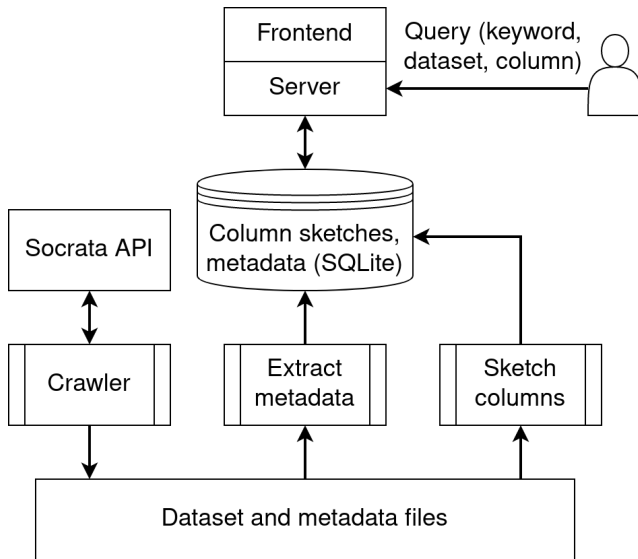
Joinable table search

Unionable table search

Semantic Keyword Search

Future work

System overview



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

Minhash²

- ▶ 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 sets¹.
- ▶ Minhash LSH hashes similar signatures to the same bucket.

¹Mining of Massive Datasets, Chapter 3.

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

LSH Ensemble³

- ▶ 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.

³Erkang 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

Semantic Keyword Search

- ▶ Problem: Given a list of keywords, return datasets which are more similar than threshold t .
- ▶ Motivation: Data scientists want a simple way to find new and insightful datasets

Our Approach

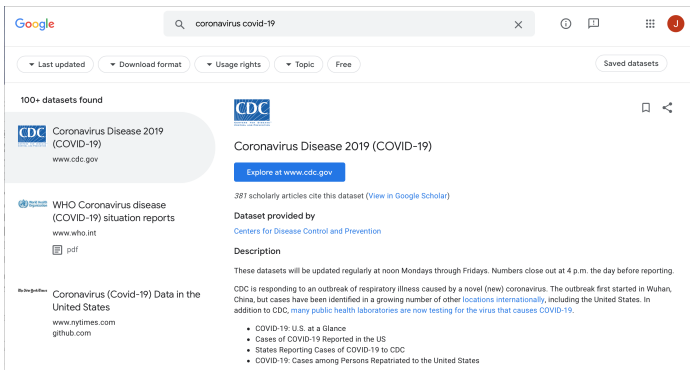
- ▶ Search on the metadata, not on the data in the dataset
 - ▶ Data in dataset is too noisy
- ▶ Metadata that we have:
 - ▶ Dataset description
 - ▶ Column description
 - ▶ Datasets tags

Our Approach (Cont.)

- ▶ Use semantic NOT syntactic similarity
 - ▶ Example: Fish & Seafood
 - ▶ Example: Coronavirus & Respiratory System

Others Approach

► Google Dataset Search



The screenshot shows the Google Dataset Search interface. At the top, the search bar contains 'coronavirus covid-19'. Below the search bar, there are filters for 'Last updated', 'Download format', 'Usage rights', 'Topic', and 'Free'. A 'Saved datasets' button is on the right. The results section shows '100+ datasets found'. The first result is 'Coronavirus Disease 2019 (COVID-19)' from the CDC, with a link to 'www.cdc.gov'. The second result is 'WHO Coronavirus disease (COVID-19) situation reports' from 'www.who.int', with a PDF icon. The third result is 'Coronavirus (Covid-19) Data in the United States' from 'www.nytimes.com' and 'github.com'. The CDC result is expanded, showing a description: '387 scholarly articles cite this dataset (View in Google Scholar)'. It also lists 'Centers for Disease Control and Prevention' as the provider and includes a 'Description' section stating that the datasets are updated regularly and that CDC is responding to the outbreak. A list of four specific datasets is provided at the bottom of the CDC result.

Google coronavirus covid-19

Last updated Download format Usage rights Topic Free Saved datasets

100+ datasets found

CDC Coronavirus Disease 2019 (COVID-19)
www.cdc.gov

WHO Coronavirus disease (COVID-19) situation reports
www.who.int
pdf

nytimes.com github.com Coronavirus (Covid-19) Data in the United States
www.nytimes.com
github.com

CDC Coronavirus Disease 2019 (COVID-19)
Explore at www.cdc.gov

387 scholarly articles cite this dataset (View in Google Scholar)

Dataset provided by
Centers for Disease Control and Prevention

Description
These datasets will be updated regularly at noon Mondays through Fridays. Numbers close out at 4 p.m. the day before reporting.

CDC is responding to an outbreak of respiratory illness caused by a novel (new) coronavirus. The outbreak first started in Wuhan, China, but cases have been identified in a growing number of other locations internationally, including the United States. In addition to CDC, many public health laboratories are now testing for the virus that causes COVID-19.

- COVID-19: U.S. at a Glance
- Cases of COVID-19 Reported in the US
- States Reporting Cases of COVID-19 to CDC
- COVID-19: Cases among Persons Repatriated to the United States

System Overview

- ▶ FastText: word in dataset's metadata \rightarrow embedding vector
- ▶ SimHash: embedding vector \rightarrow bit vector
- ▶ Locality Sensitive Hashing (LSH): build index on the bit vector of each word

- ▶ Word in dataset's metadata -> embedding vector
- ▶ Embedding vector represent the semantics of words
- ▶ Embedding vectors are learned from wikipedia articles

⁴A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, *Bag of Tricks for Efficient Text Classification*

FastText (Cont.)

- ▶ closeness or similarity of vectors $:=$ Cosine-Similarity
- ▶ Closer a pair of vectors, closer the semantics of the two words
 - ▶ PICTURE

Simhash

- ▶ Embedding Vector \rightarrow Bit Vector
 - ▶ PICTURE

Locality Sensitive Hashing (LSH)

- ▶ Underlying data structure: Hash Table
 - ▶ Predefined # of buckets
- ▶ Insert SimHashed embedding vectors into hash table
- ▶ Collisions in hash table buckets are candidate pairs.

SimHash LSH (Cont.)

- ▶ Perdefined # of hash tables
- ▶ Query each L hash table for M candidates
 - ▶ $M \geq k$
- ▶ Order M candidates into a top-k list by the cosine similairty of embedding vectors

Problem with SimHash LSH

- ▶ The # of hash tables and # of buckets in each hash table must be **hand tuned**
- ▶ Must be retuned when data size significantly changes
- ▶ PICTURE

LSH Forest

- ▶ Underlying data structure: Prefix Tree or Trie
- ▶ Similar to LSH
 - ▶ Predefined # of prefix trees
 - ▶ Query each L hash table for M candidates
 - ▶ $M \geq k$

LSH Forest (Cont.)

- ▶ Variable length hashing in prefix tree solves LSH's problems
- ▶ PICTURE
- ▶ Prefix Tree expands and contracts to account for # of embedding vectors
 - ▶ Thus, no hand tuning

Answering Queries

- ▶ Query the index with each keyword in the keyword list
- ▶ Add the results to a list
- ▶ Rank datasets by how often they appear in the list

Problems

- ▶ No semantic relationships **between** words
 - ▶ Example: Keyword List := “traffic violations”
 - ▶ Produces good results for “traffic” and “violations”, but not “traffic violations”

Outline

System overview

Joinable table search

Unionable table search

Semantic Keyword Search

Future work

Future work

- ▶ Improve ability to see semantic relationships between words
- ▶ Organize datasets into a directory structure
- ▶ Use semantic similarity of column names in unionable table search.
- ▶ Similar dataset search based on metadata similarity.
- ▶ Keyword search over data values.