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

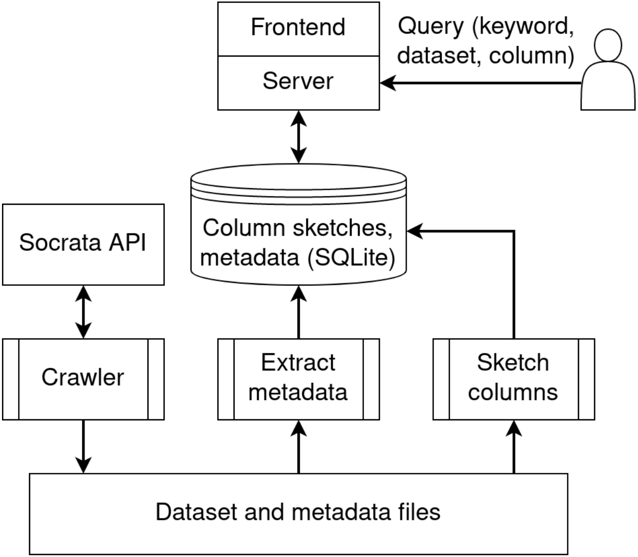
Joinable table search

Unionable table search

Semantic Keyword Search

Future work

System overview


Frontend
Server
Query (keyword, dataset, column)
Socrata API
Column sketches, metadata (SQLite)
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

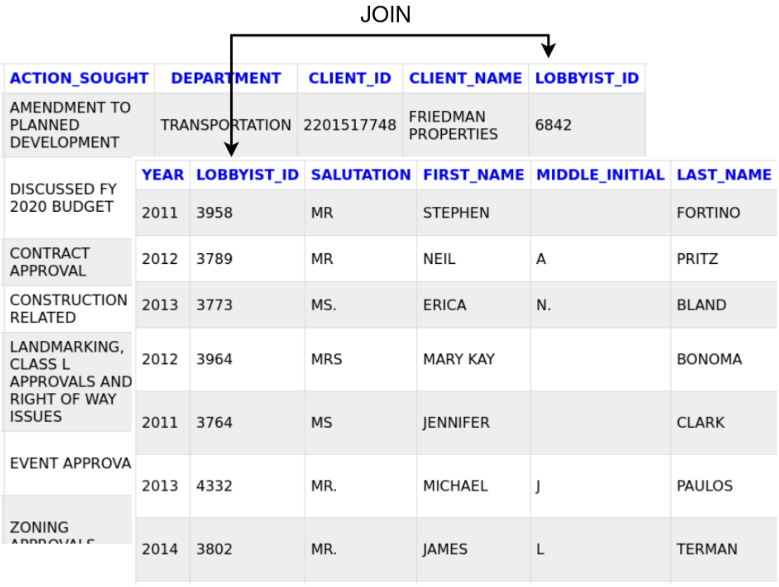
Joinable table search

Unionable table search

Semantic Keyword Search

Future work

Joinable table search



ACTION_SOUGHT	DEPARTMENT	CLIENT_ID	CLIENT_NAME	LOBBYIST_ID
AMENDMENT TO PLANNED DEVELOPMENT	TRANSPORTATION	2201517748	FRIEDMAN PROPERTIES	6842
DISCUSSED FY 2020 BUDGET				
CONTRACT APPROVAL				
CONSTRUCTION RELATED				
LANDMARKING, CLASS L APPROVALS AND RIGHT OF WAY ISSUES				
EVENT APPROVA				
ZONING APPROVALS				

YEAR	LOBBYIST_ID	SALUTATION	FIRST_NAME	MIDDLE_INITIAL	LAST_NAME
2011	3958	MR	STEPHEN		FORTINO
2012	3789	MR	NEIL	A	PRITZ
2013	3773	MS.	ERICA	N.	BLAND
2012	3964	MRS	MARY KAY		BONOMA
2011	3764	MS	JENNIFER		CLARK
2013	4332	MR.	MICHAEL	J	PAULOS
2014	3802	MR.	JAMES	L	TERMAN

Joinable table search

- Attributes are treated as sets.
- Sets are encoded with minhash data sketches.
- A table T is joinable with the query U if $Containment(X \in T, Q \in U) \geq t$.
- We use an LSH index for fast querying.

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

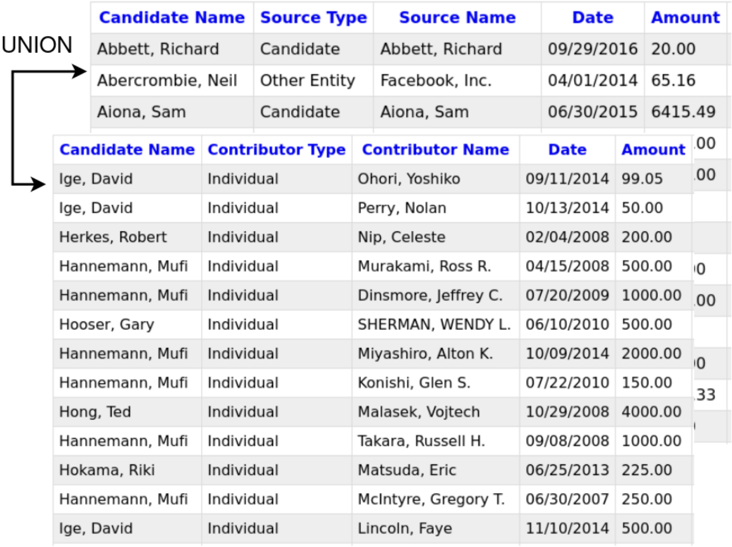
Joinable table search

Unionable table search

Semantic Keyword Search

Future work

Unionable table search



UNION

Candidate Name	Source Type	Source Name	Date	Amount
Abbett, Richard	Candidate	Abbett, Richard	09/29/2016	20.00
Abercrombie, Neil	Other Entity	Facebook, Inc.	04/01/2014	65.16
Aiona, Sam	Candidate	Aiona, Sam	06/30/2015	6415.49

Candidate Name	Contributor Type	Contributor Name	Date	Amount
Ige, David	Individual	Ohori, Yoshiko	09/11/2014	99.05
Ige, David	Individual	Perry, Nolan	10/13/2014	50.00
Herkes, Robert	Individual	Nip, Celeste	02/04/2008	200.00
Hannemann, Mufi	Individual	Murakami, Ross R.	04/15/2008	500.00
Hannemann, Mufi	Individual	Dinsmore, Jeffrey C.	07/20/2009	1000.00
Hooser, Gary	Individual	SHERMAN, WENDY L.	06/10/2010	500.00
Hannemann, Mufi	Individual	Miyashiro, Alton K.	10/09/2014	2000.00
Hannemann, Mufi	Individual	Konishi, Glen S.	07/22/2010	150.00
Hong, Ted	Individual	Malasek, Vojtech	10/29/2008	4000.00
Hannemann, Mufi	Individual	Takara, Russell H.	09/08/2008	1000.00
Hokama, Riki	Individual	Matsuda, Eric	06/25/2013	225.00
Hannemann, Mufi	Individual	McIntyre, Gregory T.	06/30/2007	250.00
Ige, David	Individual	Lincoln, Faye	11/10/2014	500.00

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.
  - $0 \leq t \leq 1$
- Motivation: Data scientists want a simple way to find new and insightful datasets

Our Approach

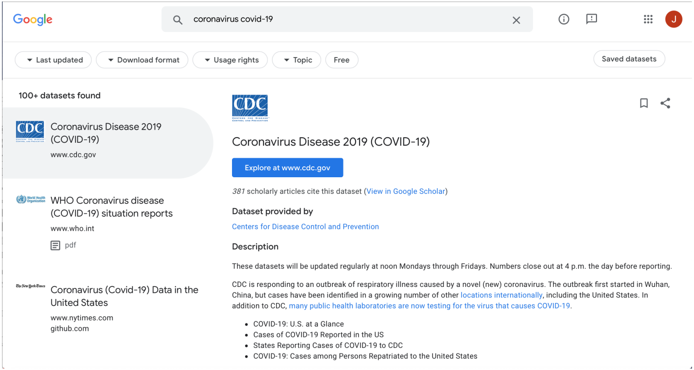
- Search on the metadata, not on the data in the dataset
  - Data in dataset is too noisy
- Metadata that we have:
  - Dataset description
  - Column description
  - Datasets tags

Our Approach (Cont.)

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

Others Approach

- Google Dataset Search



100+ datasets found

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

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

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

Coronavirus Disease 2019 (COVID-19)

381 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: 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.