Constructing Domain Specific Knowledge Graphs

Mayank Kejriwal, Craig Knoblock and Pedro Szekely
Information Sciences Institute,
University of Southern California
Domain-specific search (DSS)
Emerging opportunities for DSS

Fighting human trafficking

Predicting cyberattacks

Stopping Penny Stock Fraud

Accurate geopolitical forecasting
DARPA/IARPA programs

DARPA Memex

IARPA Hybrid Forecasting Competition

DARPA AIDA

DARPA Causal Exploration

DARPA LORELEI

IARPA CAUSE
DSS is more than keyword search

Lead Investigation
What is the ad with the earliest post date containing number 7075610282?

Aggregations/Lists
List all ads in Seattle, WA that include an ethnicity in the ad text. In the answer field, concatenate and list ethnicities

Indicator Mining
List all ads that have high probability of movement
List all ads in the Chicago area advertising multiple people at once

Dossier Generation
Collect and show me all information on the phone number 7075610282
Google Knowledge Graph
What is a Knowledge Graph?

set of triples, where each triple \((h, r, t)\) represents a relationship \(r\) between head entity \(h\) and tail entity \(t\)

\[(\text{Barack Obama, wasBornOnDate, 1961-08-04}),\]
\[(\text{Barack Obama, hasGender, male}),\]
...
\[(\text{Hawaii, hasCapital, Honolulu}),\]
...
\[(\text{Michelle Obama, livesIn, United States})\]
How do we construct domain specific knowledge graphs over web data for powerful DSS applications
Knowledge Graphs for DSS

Relevance Model
http://site1/...
http://site2/...
estort, incall, ...

Web

Domain Discovery

Corpus
url-1: HTML-1
url-2: HTML-2
url-3: HTML-3

Knowledge Graph Construction

Knowledge Graph

Guider

Investigative Search Engine

Schema

- eye color
- hair color
- ethnicity
- content
- vendor
- phone
- email

- Ad

- Vendor

GUI

SPARQL

scored entities

forest of SBT queries

ranked entities, attributes, aggregations
Agenda

- Domain-Specific Search
- Why Knowledge Graphs?
- Knowledge Graph Construction
  - Knowledge Graph Search
  - Knowledge Graph Completion
- Short-Tail Extraction
- Mapping Extractions To An Ontology
- Long-Tail Extraction
- Entity Resolution

Domains and Data
What is (or even isn’t) a domain?

Some dictionary definitions

(Merriam Webster) A sphere of knowledge, influence or activity

(Oxford) A specified sphere of activity or knowledge

Specifying the sphere

Rules
Scope (e.g., the legal system)
Syllabi (for classrooms)
Examples

How do domain experts specify the sphere?

Examples
Ontology
Domain-Specific Challenges

- Subject matter
- Complex nature
- Obfuscation
- How to adapt off-the-shelf tools?
- Ambiguous

Example text:

Italian 19 hello guys.... My name is charlotte, New to town from kansas

[ GORGEOUS BLONDE beauty] ? FROM Florida ? (Petite) ? [ CURVv ]?

NO DISAPPOINTMENTS. 34C...Brazilian,ITALIAN beauty....

Hey gentleman im Newyork and i'm looking for generous
Hi guy's this is sexy newyork , & ready to party.
AVAILABLE NOW! ?? - (1 two 1) six 5 six - 0 9 one 2 - 21
Specifying investigative domains

Functional
I have some questions I’d like answers to
Domain is the scope of the answers
Presents interesting cognitive dilemma!
I know what I want but can’t define it precisely

Two major functional steps

Data Acquisition
• Find me the data from a universe aka the Web that can help me answer my questions

Ontological Specification
• Let me define fields and field properties that will help me unambiguously represent questions and interpret answers
Specifying investigative domains

**Functional**

I have some questions I’d like answers to
Domain is the scope of the answers
Presents interesting cognitive dilemma!
I know what I want but can’t define it precisely

**Two major functional steps**

Data Acquisition
• The data from a universe aka the Web that can help me answer my questions

Ontological Specification
• The classes and fields that will help me unambiguously represent questions and interpret answers
In practice...

...investigators think of a domain as a tri-faceted combination of:

1. Questions
2. Entity types (a shallow ontology)
   - Ad, Posting Date, Title, Content, Phone, Email, Review
   - ID, Social Media ID, Price, Location, Service, Hair
   - Color, Eye Color, Ethnicity, Weight, Height
3. Examples/Annotations
Crawling Challenges

Scale, cost, speed
DNS, fetching, parsing/extracting, memory/disk

Errors, redirects, localization
Need sophisticated software

Deep web, forms, dynamic pages, infinite scrolling
Identify and fill in forms, render pages while crawling (headless browser)

Counter-crawling measures
Login, captchas, trap, fake errors, banning

Freshness and deduplication
Identify and re-crawl new content
Domains have a long tail

The human-trafficking domain: 140 million pages

Many interesting things to be found, but how do we automate it at scale?