Scalable Construction and Querying of Massive Knowledge Bases

Xiang Ren¹  Yu Su²  Pedro Szekely³  Xifeng Yan²

¹University of Southern California
²University of California, Santa Barbara
³Information Sciences Institute, University of Southern California
Tutorial website:
http://usc-is-i2.github.io/WWW18_1/

Slides, code, datasets, references
Growing Gap between Human and Data

What disease does the patient have?
• (EMR) Similar patients?
• (Literature) New findings?
• (Gene sequence) Suspicious mutations?
• ... ...

Ad-hoc information needs for on-demand decision making

Massive, heterogeneous data

86.9% adoption (NEHRS 2015)
27M+ papers, >1M new/year (PubMed)
$1000 gene sequencing
24x7 monitoring
How can AI Bridge the Gap?

Bottleneck #1: Knowledge

Bottleneck #2: Access

Bottleneck #3: Reasoning

Insights
Discoveries
Solutions
Broad Applications

“Which cement stocks go up the most when a Category 3 hurricane hits Florida?”

KENSFO
Constructing Domain Specific Knowledge Graphs

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?

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

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

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 \(t\)

(Barack Obama, wasBornOnDate, 1961-08-04),
(Barack Obama, hasGender, male),
...
(Hawaii, hasCapital, Honolulu),
...
(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/... escort, incall, ...

Shallow Ontology

Web → Domain Discovery → Corpus

url-1: HTML-1 url-2: HTML-2 url-3: HTML-3...

Knowledge Graph Construction

Knowledge Graph

GUI

Investigative Search Engine

SPARQL

Information Sciences Institute

USC Viterbi
Agenda

- Domain-Specific Search
- Why Knowledge Graphs?
- Knowledge Graph Construction
  - Short-Tail Extraction
  - Mapping Extractions To An Ontology
  - Long-Tail Extraction
  - Entity Resolution
- Knowledge Graph Search
- Knowledge Graph Completion
- 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
- Ambiguous
- Obfuscation
- How to adapt off-the-shelf tools?

Example:

<table>
<thead>
<tr>
<th>Italian 19 hello guys....My name is charlotte, New to town from kansas</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ GORGOUS BLONDE beauty] ? FROM Florida ? (Petite) ? [CURVy]?</td>
</tr>
<tr>
<td>NO DISAPPOINTMENTS. 34C..Brazilian,ITALIAN beauty....</td>
</tr>
<tr>
<td>Hey gentleman im Newyork and i’m looking for generous</td>
</tr>
<tr>
<td>Hi guy’s this is sexy newyork &amp; ready to party.</td>
</tr>
<tr>
<td>AVAILABLE NOW! ?? - (1 two 1) six 5 six - 0 9 one 2 - 21</td>
</tr>
</tbody>
</table>
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, traps, 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?
Schema-agnostic Knowledge Base Querying

Yu Su
University of California, Santa Barbara
**Structured Query: RDF + SPARQL**

### Triples in an RDF graph

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack_Obama</td>
<td>parentOf</td>
<td>Malia_Obama</td>
</tr>
<tr>
<td>Barack_Obama</td>
<td>parentOf</td>
<td>Natasha_Obama</td>
</tr>
<tr>
<td>Barack_Obama</td>
<td>spouse</td>
<td>Michelle_Obama</td>
</tr>
<tr>
<td>Barack_Obama_Sr.</td>
<td>parentOf</td>
<td>Barack_Obama</td>
</tr>
</tbody>
</table>

### SPARQL query

```sparql
SELECT ?x WHERE {
  Barack_Obama_Sr. parentOf ?y .
  ?y parentOf ?x .
}
```

### Answer

```sparql
<Malia_Obama>
<Natasha_Obama>
```
Why Structured Query Falls Short?

<table>
<thead>
<tr>
<th>Knowledge Base</th>
<th># Entities</th>
<th># Triples</th>
<th># Classes</th>
<th># Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freebase</td>
<td>45M</td>
<td>3B</td>
<td>53K</td>
<td>35K</td>
</tr>
<tr>
<td>DBpedia</td>
<td>6.6M</td>
<td>13B</td>
<td>760</td>
<td>2.8K</td>
</tr>
<tr>
<td>Google Knowledge Graph*</td>
<td>570M</td>
<td>18B</td>
<td>1.5K</td>
<td>35K</td>
</tr>
<tr>
<td>YAGO</td>
<td>10M</td>
<td>120M</td>
<td>350K</td>
<td>100</td>
</tr>
<tr>
<td>Knowledge Vault</td>
<td>45M</td>
<td>1.6B</td>
<td>1.1K</td>
<td>4.5K</td>
</tr>
</tbody>
</table>

* as of 2014

It’s more than large: High heterogeneity of KBs

If it’s hard to write SQL on simple relational tables, it’s only harder to write SPARQL on large knowledge bases

Even harder on automatically constructed KBs with a loosely-defined schema
Not Everyone Can Program...

“find all patients diagnosed with eye tumor”

WITH Traversed (cls, syn) AS (  
(SELECT R.cls, R.syn  
FROM XMLTABLE ('Document("Thesaurus.xml")'  
/terminology/conceptDef/properties  
[property/name/text()="Synonym" and  
property/value/text()="Eye Tumor"]  
/property[name/text()="Synonym"]/value'  
COLUMNS  
cls CHAR(64) PATH './parent::*/parent::*'  
/parent::*/name',  
tgt CHAR(64) PATH '} AS R)  
UNION ALL  
(SELECT CH.cls,CH.syn  
FROM Traversed PR,  
XMLTABLE ('Document("Thesaurus.xml")'  
/terminology/conceptDef/definingConcepts/  
concept[/text()=$parent]/parent::*/parent::*'  
properties/property[name/text()="Synonym"]/value'  
PASSING PR.cls AS "parent"  
COLUMNS  
cls CHAR(64) PATH './parent::*/'  
/parent::*/parent::*/name',  
syn CHAR(64) PATH '}) AS CH))  
SELECT DISTINCT V.*  
FROM Visit V  
WHERE V.disease IN  
(SELECT DISTINCT syn FROM Traversed)
In Pursue of Efficiency

find all patients diagnosed with eye tumor
In Pursue of Efficiency

find all patients diagnosed with eye tumor

Schema-agnostic Querying

```
WITH Patient (name) AS (
  SELECT name FROM Patients WHERE diagnosis = 'Eye Tumor')
FROM Patient
UNION ALL
SELECT name FROM Patient
UNION ALL
SELECT name FROM Treatment
WHERE diagnosis = 'Eye Tumor'
```

Database
Schema-agnostic KB Querying

"Barack Obama Sr. grandchildren"

Keyword query: query like search engine

Barack Obama Sr. grandchildren

Graph query: add a little structure

"Who are Barack Obama Sr.’s grandchildren?"

Natural language query: the holy grail
Tutorial Outline

Introduction

Part I: Domain-specific KB Construction

Lunch Break

Part II: Schema-agnostic KB Querying

Summary & Future Directions