Data mining in unusual domains with information-rich knowledge graph construction, inference and search

Mayank Kejriwal, Pedro Szekely

Information Sciences Institute,
USC Viterbi School of Engineering
Agenda

Unusual domains
Knowledge graphs
KG construction
Case study

Knowledge graph completion
Entity resolution
Probabilistic soft logic
KGs in latent space
Searching knowledge graphs
Unusual domains
Disaster Assistance In Regions Using Low Resource Languages

Data
Social media and news in low resource languages
Akan, Amharic, Hausa, Tagalog, Uyghur, Wolof, Yoruba, …

Example Questions
Identify areas where people are in greatest need
Identify threats to relief personnel
Characterize the evolution of the disaster

Technical Challenges
Very noisy translation
Clustering documents according to need, location and entities
Streaming data
Identify Illegal Firearm Sales

Data
Classified ads and forums
Open and dark web

Example Questions
Identify buyers who purchase on behalf of others
Identify people who buy and sell without proper licenses
Identify vendors who illegally sell across state lines
Identify stolen firearms for sale

Technical Challenges
Most firearm sales are legal
Huge volume of pages and web sites
Unusual language model
Identify Counterfeit Electronics Vendors

Data

Online catalogs and forums
Internal sources containing suspicious vendors, shipping addresses

Example Questions

Identify clusters of companies under control of one organization
Identify fraudulent ads across vendor sites

Technical Challenges

Many pages in Chinese
Fake images on catalogs
Frequent creation of new shell vendor companies
Identify Narcotics Vendors In The Dark Web

Data
Dark web marketplace and forum pages
Open web social media sites

Example Questions
Identify dark web personas in the open web

Technical Challenges
Pages in multiple languages
Unusual language model
Deception, fake vendors, fake reviews
Careful concealment of identifying information
Combat Fraud In The Penny Stock Market

Data
Web pages and social media about companies in the Over The Counter (OTC) market

Example Questions
Identify in-progress pump-and-dump scams
Identify chain of shell companies and individuals involved
Identify promoters, attorneys, etc.
Identify evidence of unlawful behavior (e.g., false statements in promotional materials)

Technical Challenges
Pump-and-dump scams are carefully planned and controlled to look legal
Wide variety of page genres: company profiles, press releases, promotions, financial data
Identify Human-Trafficking Victims &Prosecute Traffickers

Data
100,000,000 escort ads published on the web
50,000 images

Example Questions
Identify all ads for an escort given soft identifies
Tag ads as high-risk for human trafficking
Identify stables of escorts controlled by one trafficker
Identify fake images
Identify networks of phone numbers

Technical Challenges
Noise, obfuscation, deception and large size
Unique language model
Long-tailed set of sources
Lack of identifiers and reference data
Work on unusual domains has significant social impact
Find Locations Where An Escort Was Advertised

25 of 46,190,422 Results

2.06

100hh Special ?H1t C??a ??? P?a???? ????? R?? 100% W??t? T??

V?h?? - Chicago escorts - backpage.com

Oct 6, 2015

Locations: Chicago

Telephone Numbers: 323-8111

Services Provided: Fetish friendly

Provider Ages: 20

No Provider Eye Colors

No Provider Hair Colors

Provider Heights: 5'10"

Price: 100 per hh

Provider Ethnicities: Hispanic

No Social Media IDs

No Review IDs

Provider Names: raia

Website: backpage.com

Cached Ad Webpage: Open

What is a cached webpage?

Phone: 323-8111

Location Drops Timeline

25 of 435 Ads

100 special ? B7?at??? C77a ?? B7? B7?? T????

C7??? ? 2Girls Avail 100% W??t? T??? - Indianapolis escorts - backpage.com

Oct 10, 2015

Locations:

 Prices:

Backpage use: 325

No Telephone Numbers:

Yes

No Social Media IDs:

No Provider Names:

raia

100 special ? B7?at??? C77a ?? B7? B7?? T????

C7??? ? 2Girls Avail 100% W??t? T??? - Indianapolis escorts - backpage.com

Oct 10, 2015

Locations:

Prices:

Backpage use: 325

No Telephone Numbers:

Yes

No Social Media IDs:

No Provider Names:

raia

Kejriwal, Szekely
Number Of HT Investigations Increased

FROM <1% TO 62%
Man sentenced to 97 years in human trafficking case

By Vivian Ho  Updated 7:07 am, Friday, April 22, 2016

97-TO-LIFE
4 ACCOMPLICES

- **Deshawn Birden**
  - Pimp
  - Present in victim “B” case
  - Las Vegas, Oakland, SF

- **Brazil Harris**
  - Pimp
  - Los Angeles, San Diego, SF
  - Arr. 4/2016 for possession for sale
  - Co-de said she was forced

- **Jermaine Fulgham**
  - Self-admitted pimp
  - Assisted Geeter in jail
  - Present for victim 1 "Z"  

- **Jhontay Wills**
  - Pimp
  - Prior arrest for 653.23 on Capp St.
  - Friend of Geeter
  - Was assisting in bringing in character witnesses (minors)
Victims

Cardise Burns
- Mother of Geeter’s first child
- Reached out post verdict to SFDA about victimization
- Sgt. Flores has been in contact

Yasmine Malone
- Missing 15 yo victimized in 12/2014
- MH issues; assaulted SFSD officer
- Birden/Walls in contact post-trial and assisting in exploitation (2015)

Shakarri Miller
- Missing 15 yo victimized
- Arrested in Broadmoor, CA with Geeter

Courtney Taylor
- Missing 17 yo victimized
- Arrested in Garden Grove, CA with Geeter and Birden

+20 additional, possible victims of ring

Zurline Hurst

Belinda Carson

>25 VICTIMS
Case Study 4: Recovery of Missing Juvenile

**Problem:** Missing juvenile via NCMEC who reached out to friends, saying she was being trafficked throughout CA

**Outcome:** A. Green was recovered in Atlanta, Georgia by the FBI after I provided NCMEC and local law enforcement with her most recent sex ad and location.
<table>
<thead>
<tr>
<th>Usual</th>
<th>Unusual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good English</td>
<td>Jargon, ungrammatical</td>
</tr>
<tr>
<td>Mostly correct information</td>
<td>Obfuscation &amp; deception</td>
</tr>
<tr>
<td>Large reference datasets</td>
<td>No reference datasets</td>
</tr>
<tr>
<td>Small number of sources</td>
<td>Long tail</td>
</tr>
<tr>
<td>Sources readily available</td>
<td>Ephemeral pages, dark web, anti-crawling measures</td>
</tr>
</tbody>
</table>
Two guiding questions

How do we model, and represent knowledge in, unusual domains?

Are there general lessons to be learned across domains?

How do we do search and inference in unusual domains?
how to represent KGs?
KG Definition

a directed, labeled multi-relational graph representing facts/assertions as triples

(h, r, t)   head entity, relation, tail entity
(s, p, o)   subject, predicate, object
Simplest Knowledge Graph

Entities

- LGYV
- Legacy Ventures International Inc
- Damn Good Penny Stocks

Easiest to build
Simple, But Useful KG

Entities + properties

stock-ticker

company

promoter

LGYV

Legacy Ventures International Inc

Damn Good Penny Stocks

“Easy” to build
Semantic Web KG (RDF/OWL)

Entities + properties + classes

Company

is-a

is-a

Legacy Ventures International Inc

stock-ticker

promoter

Damn Good Penny Stocks

LGYV

Very hard to build
“Ideal” KG

Entities + properties + classes + qualifiers

Company

is-a

Legacy Ventures International Inc

stock-ticker

LGYV

Damn Good Penny Stocks

is-a

promoter

source

stockreads.com

start-date

June 2017

Very very hard to build
How about Ontologies?
Ontologies?

Domain ontologies don’t exist for unusual domains

Build new one or extend existing one?
Deep or shallow?
Ontologies?

Goal is to help users solve problems

Let users tell you what entities matter to them

Build the shallowest, simplest ontology that captures the entities users care about
Ontologies?

Domain ontologies don’t exist for unusual domains

Build new one or extend existing one?
Deep or shallow?
Where to Store KGs?
Serializing Knowledge Graphs

Resource Description Framework (RDF)
Database (triple store): AllegroGraph, Virtuoso,
Query: SPARQL (SQL-like)

Key-Value, Document Stores
Data model: Node-centric
Databases: Hbase, MongoDB, Elastic Search, ...
Query: filters, keywords, aggregation (no joins)

Graph Databases
Data model: graph
Databases: Neo4J, Cayley, MarkLogic, GraphDB, Titan, OrientDB, Oracle, ...
Query: GraphQL, Gremlin, Cypher
Popularity Ranking Of Graph Databases

DB-Engines Ranking of Graph DBMS
ElasticSearch, MongoDB & Neo4J Have Wide Adoption

DB-Engines Ranking

Score (logarithmic scale)

© August 2017, DB-Engines.com

Kejriwal, Szekely

https://db-engines.com/en/ranking_trend/graph+dbms
KGs I can Reuse
DBpedia

RDF graph derived from Wikipedia
http://wiki.dbpedia.org/

4.58 million things
4.22 million are classified in a consistent ontology

1,445,000 persons

735,000 places
478,000 populated places

411,000 creative works
123,000 music albums, 87,000 films and 19,000 video games

241,000 organizations
58,000 companies and 49,000 educational institutions

251,000 species

6,000 diseases
YAGO Knowledge Base


Derived from Wikipedia, WordNet and GeoNames

10 million entities
120 million assertions
persons, organizations, cities, etc.

350,000 classes

many fine grained classes, inferred from the data
Collaborative, multilingual
collecting structured data to provide support for Wikipedia

31,419,072 items
534,615,360 edits since the project launch
Google Knowledge Graph
https://developers.google.com/knowledge-graph/how-tos/search-widget-example

derived from many sources, including the CIA World Factbook, Wikidata, and Wikipedia

powers a "knowledge panel"

the Knowledge Graph now holds 70 billion facts
Other Knowledge Graphs

Internet Movie Firearms Database
Firearms used or featured in movies, television shows, video games, and anime
22,159 articles, extensive coverage and ontology
http://www.imfdb.org/wiki/Category:Gun

Microsoft Satori
Large knowledge graph similar to Google KG, e.g., 1.8 million bottles of wine
Many streaming channels of real-time data, e.g., bitcoin, transportation, ...
https://www.satori.com/

LinkedIn Knowledge Graph
450M members, 190M historical job listings, 9M companies, 28K schools,
1.5K fields of study, 600+ degrees, 24K titles and 35K skills in 19 languages
https://engineering.linkedin.com/blog/2016/10/building-the-linkedin-knowledge-graph
Knowledge ‘base’ vs ‘graph’

Terms often used interchangeably in the literature

Knowledge base is more of a catch-all e.g., Wikipedia vs. DBpedia

We use the term ‘knowledge graph’ where possible
Why Knowledge Graphs?

Combine advantages of databases and unstructured text

Machine and human understandable

Useful in search and data mining

Many interesting extensions

Interest in multiple communities incl. NLP, data mining...

Many more!
KGs in Unusual domains
Knowledge Graphs In Unusual Domains

Many challenges (and frontier research questions)!

Data skew
Lack of training data
Importance of data exploration
Keeping domain experts in the loop
Multi-faceted prediction and inference
Scalability

...
KG For Unusual Domains

Entities + properties + provenance + confidence + qualifiers

Simple, shallow ontology

customized to domain, if not needed, leave it out

Rich provenance and confidences

essential for end-users, useful for knowledge graph improvement

Hybrid text/structured representation

keep the text, essential for machine learning and search

“Easy” to build
Knowledge Graph Construction
How can I **build** a KG?
KG Construction Problem

Web
> $10^{13}$ pages

Databases

Knowledge Graph
Steps To Build a KG

1. Web: > $10^{13}$ pages
2. Crawling
3. Databases
4. Information Extraction
5. Knowledge Snippets
6. Knowledge Graph Construction
7. Knowledge Graph

Kejriwal, Szekely
Crawling the web
The web is very BIG

Google Index

Deep Web

Dark Web

Web
We can only get a tiny part.
Crawling Research Challenges

Domain discovery
Identifying relevant sites, datasets and pages

Crawling
Building models of relevant content
Identifying new content
Downloading dynamic content
Overcoming anti-crawling measures
Exhibiting human-like behavior
Crawling Tools

**Scrapy (targeted crawling)**
Open source and collaborative framework for extracting data from websites
https://scrapy.org/ACHE

**Apache Nutch (massive crawling)**
highly extensible, highly scalable Web crawler
http://nutch.apache.org

**Deep Deep (Adaptive crawler)**
reinforcement learning to learn which links to follow
https://github.com/TeamHG-Memex/deep-deep

**ACHE (focused crawler)**
https://github.com/ViDA-NYU/ache

**ScrapingHub (service)**
https://scrapinghub.com/data-on-demand
Unusual Domains Have Long Tails

Web sites (second level domains), ordered by number of pages (log scale)
principles and challenges in Information Extraction
Information Extraction (IE)
Apple to Open Its First Retail Store in New York City

MACWORLD EXPO, NEW YORK--July 17, 2002--Apple's first retail store in New York City will open in Manhattan's SoHo district on Thursday, July 18 at 8:00 a.m. EDT. The SoHo store will be Apple's largest retail store to date and is a stunning example of Apple's commitment to offering customers the world's best computer shopping experience.

"Fourteen months after opening our first retail store, our 31 stores are attracting over 100,000 visitors each week," said Steve Jobs, Apple's CEO. "We hope our SoHo store will surprise and delight both Mac and PC users who want to see everything the Mac can do to enhance their digital lifestyles."
Dimensions of IE

Document features

Scope

Pattern complexity

Relevance
Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.
Scope

Web site specific

Genre specific (e.g., forums)

Wide, non-specific

Kejriwal, Szekely
# Pattern Complexity

**E.g., word patterns**

<table>
<thead>
<tr>
<th>Closed set</th>
<th>Regular set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U.S. states</strong></td>
<td><strong>U.S. phone numbers</strong></td>
</tr>
<tr>
<td>He was born in <em>Alabama</em>…</td>
<td>Phone: <em>(413) 545-1323</em></td>
</tr>
<tr>
<td>The big <em>Wyoming</em> sky…</td>
<td>The CALD main office can be reached at <em>412-268-1299</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Complex pattern</strong></th>
<th><strong>Ambiguous patterns, needing context and many sources of evidence</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U.S. postal addresses</strong></td>
<td>Person names</td>
</tr>
<tr>
<td>University of Arkansas P.O. Box 140 Hope, AR 71802</td>
<td>Pawel Opalinski, Software Engineer at WhizBang Labs.</td>
</tr>
<tr>
<td>Headquarters: 1128 Main Street, 4th Floor Cincinnati, Ohio 45210</td>
<td>…was among the six houses sold by Hope Feldman that year.</td>
</tr>
</tbody>
</table>

“YOU don’t wanna miss out on ME :) Perfect lil booty Green eyes Long curly black hair Im a Irish, Armenian and Filipino mixed princess :) 💚 Kim 💚 7o7~7two7~7four77 💚 HH 80 roses 💚 Hour 120 roses 💚 15 mins 60 roses”

---

*Courtesy of Andrew McCallum*
small amount of relevant content

irrelevant content very similar to relevant content
IE In Unusual Domains
Full spectrum of in all dimensions

<table>
<thead>
<tr>
<th>Document Features</th>
<th>Scope</th>
<th>Pattern Complexity</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>• text paragraphs</td>
<td>• website specific</td>
<td>• closed set</td>
<td>• all relevant</td>
</tr>
<tr>
<td>• grammatical, some formatting</td>
<td>• genre-specific</td>
<td>• regular set</td>
<td>• significant irrelevant content</td>
</tr>
<tr>
<td>• ungrammatical, rich formatting</td>
<td>• wide</td>
<td>• complex pattern</td>
<td>• active deception</td>
</tr>
<tr>
<td>• tables</td>
<td>•</td>
<td>• ambiguous pattern</td>
<td>• purposeful obfuscation</td>
</tr>
<tr>
<td>• charts</td>
<td>•</td>
<td>• unusual language model</td>
<td></td>
</tr>
</tbody>
</table>

How to adapt existing techniques without much supervision?
currently, only one universal extractor
Practical Considerations

How good (precision/recall) is necessary?
High precision when showing extractions to users
High recall when used for ranking results

How long does it take to construct?
Minutes, hours, days, months

What expertise do I need?
None (domain expertise), patience (annotation), simple scripting, machine learning guru

What tools can I use?
Many ...
# Practical IE Technologies

<table>
<thead>
<tr>
<th>Glossary</th>
<th>Regex</th>
<th>NLP Rules</th>
<th>Semi-Structured</th>
<th>CRF</th>
<th>NER</th>
<th>Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effort</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>assemble glossary</td>
<td>hours</td>
<td>hours</td>
<td>minutes</td>
<td>O(1000) annotations</td>
<td>zero</td>
<td>O(10) annotations</td>
</tr>
<tr>
<td>Expertise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>minimal</td>
<td>high, programmer</td>
<td>low</td>
<td>minimal</td>
<td>low-medium</td>
<td>zero</td>
<td>minimal</td>
</tr>
<tr>
<td>Precision</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>medium (ambiguity)</td>
<td>high</td>
<td>high</td>
<td>high</td>
<td>medium-high</td>
<td>medium-high</td>
<td>high</td>
</tr>
<tr>
<td>Recall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>medium (formatting)</td>
<td>low</td>
<td>medium</td>
<td>high</td>
<td>medium</td>
<td>medium</td>
<td>high</td>
</tr>
<tr>
<td>Coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wide</td>
<td>wide</td>
<td>wide</td>
<td>single site</td>
<td>genre</td>
<td>news wire</td>
<td>narrow</td>
</tr>
</tbody>
</table>
Case Study

Combating fraud in the penny stock market
Microcap Stock Fraud

**Microcap stock fraud** is a form of securities fraud involving stocks of "microcap" companies, generally defined in the United States as those with a market capitalization of under $250 million. Its prevalence has been estimated to run into the billions of dollars a year.

**Pump and dump** schemes, involving use of false or misleading statements to hype stocks, which are "dumped" on the public at inflated prices. Such schemes involve telemarketing and Internet fraud.

https://en.wikipedia.org/wiki/Microcap_stock_fraud
Most Relevant Websites

4-traders.com
advfn.com
analystratings.net
barchart.com
bitcointalk.org
blogspot.com
blogspot.in
businessinsider.com
businessprofiles.com
dividend.com
dynamoo.com
etf.com
facebook.com
fifighter.com
financialcontent.com
finanzen.nl
finanzen100.de
hotstocked.com
index.co
investorshangout.com
marketnewscall.com
marketwatch.com
minyanville.com
moneyhub.net
nasdaq.com
openpr.com
otecmarkets.com
pennystock101.org
pennystocktweets.com
pinkinvesting.com
prnewswire.com
rumas.de
sify.com
siliconinvestor.com
stockguru.com
stockopedia.com
stockreads.com
superstockscreener.com
tdameritrade.com
thehotpennystocks.com
thelion.com
theotc.tod
traders.com
trading-treff.de
twitter.com
uservoice.com
wikinvest.com
yahoo.com
## Microcap Fraud Ontology

Defined by two SEC users in 45 minutes

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>address</td>
<td>US postal service address</td>
</tr>
<tr>
<td>city</td>
<td>Specific cities mentioned in a page</td>
</tr>
<tr>
<td>compensation_amount</td>
<td>Amount a promoter was compensated</td>
</tr>
<tr>
<td>counsel</td>
<td>Counsel</td>
</tr>
<tr>
<td>country</td>
<td>Countries mentioned in a page</td>
</tr>
<tr>
<td>date_of_post</td>
<td>Post date of the blog or article</td>
</tr>
<tr>
<td>disclaimer</td>
<td>Disclaimer in promotion</td>
</tr>
<tr>
<td>email</td>
<td>Email addresses mentioned in the page</td>
</tr>
<tr>
<td>industry</td>
<td>Industry of organization</td>
</tr>
<tr>
<td>market</td>
<td>Stock market of the issuer</td>
</tr>
<tr>
<td>message_board</td>
<td>Name of message board</td>
</tr>
<tr>
<td>message_board_category</td>
<td>Category of the message board</td>
</tr>
<tr>
<td>organization_name</td>
<td>Name of the organization</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>org_registration_date</td>
<td>Date organization was created</td>
</tr>
<tr>
<td>org_registration_number</td>
<td>State registration number for the organization</td>
</tr>
<tr>
<td>organization_status</td>
<td>Status of the organization's state registration</td>
</tr>
<tr>
<td>paying_party</td>
<td>Entity that is paying for the promotion</td>
</tr>
<tr>
<td>phone</td>
<td>Phone numbers mentioned in the page</td>
</tr>
<tr>
<td>posted_date</td>
<td>Any date mentioned in a page</td>
</tr>
<tr>
<td>promoter_name</td>
<td>Name of the promoter</td>
</tr>
<tr>
<td>state</td>
<td>Specific states mentioned in a page</td>
</tr>
<tr>
<td>stock_ticker</td>
<td>Stock tickers mentioned in the page</td>
</tr>
<tr>
<td>twitter_tag</td>
<td>Tags in a tweet</td>
</tr>
<tr>
<td>user_registration_date</td>
<td>Date the user registered for the site</td>
</tr>
<tr>
<td>username</td>
<td>User name in a website</td>
</tr>
<tr>
<td>website</td>
<td>Website where page was published</td>
</tr>
</tbody>
</table>
myDIG Demo
setting up a new domain (after crawling)

Specifying websites
Defining the domain ontology
Defining extractors
Building the knowledge graph
Customizing the search engine
# Featured Extractors

<table>
<thead>
<tr>
<th></th>
<th>Glossary</th>
<th>Regex</th>
<th>NLP Rules</th>
<th>Semi-Structured</th>
<th>CRF</th>
<th>NER</th>
<th>Table</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Effort</strong></td>
<td>assemble glossary</td>
<td>hours</td>
<td>hours</td>
<td>minutes</td>
<td>0(1000) annotations</td>
<td>zero</td>
<td>0(10) annotations</td>
</tr>
<tr>
<td><strong>Expertise</strong></td>
<td>minimal</td>
<td>high, programmer</td>
<td>low</td>
<td>minimal</td>
<td>low-medium</td>
<td>zero</td>
<td>minimal</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>medium (ambiguity)</td>
<td>high</td>
<td>high</td>
<td>high</td>
<td>medium-high</td>
<td>medium-high</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td>medium (formatting)</td>
<td>low f(# regex)</td>
<td>medium f(# rules)</td>
<td>high</td>
<td>medium</td>
<td>medium</td>
<td>high</td>
</tr>
<tr>
<td><strong>Coverage</strong></td>
<td>wide</td>
<td>wide</td>
<td>wide</td>
<td>single site</td>
<td>genre</td>
<td>news wire</td>
<td>narrow</td>
</tr>
</tbody>
</table>

Kejriwal, Szekely
Get Sample Pages

Retrieve sample pages from the CDR for testing and visualization. This tool will cache a range of your own TLDs as well as a random selection of pages from other TLDs.

Number of Pages per TLD: 200

Publish Project Files

Upload all the project files to the myDIG protected GitHub repository, backing up your files to allow going to previous versions or preserving files in case of disk failure.

Run Extractions and Load Index For Sample Pages

Run all extractors currently defined and load the extracted data into a new Elasticsearch index. You need to perform the Update To New Index command to replace your current index with the new index.
Inferlink Extractor

Automatic extraction from semi-structured pages

https://github.com/inferlink
Inferlink Extractor

input: a pile of pages

Classify by Templates

pages clustered by template

Infer Extractor

extractor
# Inferlink Extraction Accuracy

**firearms domain, 10 websites, 5 pages each**

<table>
<thead>
<tr>
<th>fields</th>
<th>Title</th>
<th>Desc</th>
<th>Seller</th>
<th>Date</th>
<th>Price</th>
<th>Loc</th>
<th>Cat</th>
<th>Member Since</th>
<th>Expires</th>
<th>Views</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect</td>
<td>1.0</td>
<td>.76</td>
<td>.95</td>
<td>.83</td>
<td>.87</td>
<td>.51</td>
<td>.68</td>
<td>1.0</td>
<td>.52</td>
<td>.76</td>
<td>.97</td>
</tr>
<tr>
<td></td>
<td>(50/50)</td>
<td>(37/49)</td>
<td>(40/42)</td>
<td>(40/48)</td>
<td>(39/45)</td>
<td>(34/50)</td>
<td>(35/35)</td>
<td>(15/29)</td>
<td>(19/25)</td>
<td>(35/36)</td>
<td></td>
</tr>
<tr>
<td>Pretty Good</td>
<td>1.0</td>
<td>.98</td>
<td>.95</td>
<td>.83</td>
<td>.98</td>
<td>.84</td>
<td>.88</td>
<td>1.0</td>
<td>.55</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>(50/50)</td>
<td>(48/49)</td>
<td>(40/42)</td>
<td>(40/48)</td>
<td>(44/45)</td>
<td>(44/50)</td>
<td>(35/35)</td>
<td>(16/29)</td>
<td>(25/25)</td>
<td>(36/36)</td>
<td></td>
</tr>
</tbody>
</table>

Pretty Good: useful to user
- extra tokens present
- non-essential tokens missing
Glossary Extraction

Simple in principle
list of words or phrases to extract

Challenges
Ambiguity: Charlotte is a name of a person and a city
Colloquial expressions: “Asia Broadband, Inc.” vs “Asia Broadband”

Research
Improving precision of glossary extractions
Extending glossaries automatically
Extraction Using Regular Expressions

Too difficult for non-programmers

regex for North American phone numbers:

```regex
^\(?:\d{3}\)?\s*\(?:\d{3}\)?\s*\(?:\d{4}\)\s*\?\?\?\?
```

Brittle and difficult to adapt to unusual domains

unusual nomenclature and short-hands

obfuscation
NLP Rule-Based Extraction

Tokenization for unusual domains

tokenize on white-space, punctuation and emojis

Token properties

literal, part of speech tag, lemma, in/out of dictionary
dependency parsing relationships (advanced)
type (alphanumeric, alphabetic, numeric)
shape (pattern of digits and characters), capitalization, prefix and suffix
number of characters, range (numbers)

Pattern

Sequence of required/optimal tokens
positive and negative patterns
Named Entity Recognizers

Machine learning models (Conditional Random Field)
people, places, organizations and a few others

SpaCy
complete NLP toolkit, Python (Cython), MIT license
code: https://github.com/explosion/spaCy
demo: http://textanalysisonline.com/spacy-named-entity-recognition-ner

Stanford NER
part of Stanford’s NLP software library, Java, GNU license
code: https://nlp.stanford.edu/software/CRF-NER.shtml
demo: http://nlp.stanford.edu:8080/ner/process
myDIG: A KG Construction Toolkit

Python, MIT license, https://github.com/usc-isii2

Enable end-users to construct domain-specific KGs
end users from 5 government orgs constructed KGs in less than one day

Suite of extraction techniques
semi-structured HTML pages, glossaries, NLP rules, NER, tables (coming soon)

KG includes provenance and confidences
enable research to improve extractions and KG quality

Scalable
runs on laptop (~100K docs), cluster (> 100M docs)

Robust
Deployed to many law enforcement agencies

Easy to install
Aug 31 2017: Docker deployment with single “docker compose up” installation
Knowledge Graph Completion

Our thanks to Lise Getoor for some slides on Entity Resolution and PSL
Problem

Extractions are noisy

Noise is not random

Postal code got extracted as phone
Email ID, social network ID got interchanged

Entity Disambiguation: Charlotte, NC vs. Charlotte the person

Complete the knowledge graph by inferring wrong links, new links
Some solutions we’ll cover today

Entity Resolution (ER)

Probabilistic Soft Logic (PSL)

Knowledge Graphs in Latent Space aka knowledge graph embeddings
Entity Resolution (ER)
Entity Resolution (ER)

The problem of clustering mentions that refer to the same underlying entity
Example: Linking Dbpedia to Freebase
Knowledge graphs contain duplicate entities
Knowledge graph nodes are multi-type
How to do ER?

Popular methods use some form of machine learning; see surveys by Kopcke and Rahm (2010), Elmagarmid et al. (2007), Christophides et al. (2015)

<table>
<thead>
<tr>
<th>Probabilistic Matching Methods</th>
<th>Supervised, Semi-supervised</th>
<th>Active Learning</th>
<th>Rule Based</th>
<th>Distance Based</th>
<th>Unsupervised</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hierarchical Graphical Models</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SVM Christen (2008)</td>
</tr>
</tbody>
</table>
With graph representation

Can propagate similarity decisions Melnik, Garcia-Molina and Rahm (2002)

More expensive but better performance

Can be generic or use domain knowledge e.g., citation/bibliography domain Bhattacharya and Getoor (2006, 2007)
Example (co-authorship)

Bhattacharya and Getoor (2006, 2007)
Feature functions - I

First line of attack is \textit{string matching}

Available Packages: \textit{SecondString, FEBRL, Whirl\ldots}
Feature functions - II

Unusual domains have many non-text fields Feature functions may have to be hand-crafted or learned from data:

- Tracking numbers for mail fraud/narcotics
- Review IDs for human trafficking
- Stock tickers for SEC
Learnable string similarity

Example: adaptive edit distance

Sets of equivalent string pairs (e.g., <Suite 1001, Ste. 1001>)

Learned parameters

Bilenko and Mooney (2003)
After training...

Apply classifier i.e. similarity function to every pair of nodes? **Quadratic complexity!**

\[ O(|V|^2) \]

Linked mentions
Blocking trick

Like a configurable inverted index function
What is a good blocking key?

Achieves high recall

Achieves high reduction

Good survey on blocking: Christen (2012)
How do we learn a good blocking key?

Key idea in existing work is to learn a DNF rule with indexing functions as atoms

CharTriGrams(Last_Name) U (Numbers(Address) X Last4Chars(SSN))

Putting it together: two-step ER

Entity Resolution

Similarity:
- machine learning, rules, heuristics, graph propagation

Blocking:
- Blocking scheme learner+blocking method

Training set of duplicates/non-duplicates

Graph inputs

Candidate pairs

Resolved entities

Kejriwal, Szekely
ER packages

Not many, still tend to be inefficient or for specific domain

**FEBRL** was designed for **biomedical** record linkage *(Christen, 2008)*

**Dedupe** crashes on graphs with fewer than a million nodes
[https://github.com/dedupeio/dedupe](https://github.com/dedupeio/dedupe)

**LIMES, Silk** mostly designed for RDF data, often require **pre-specified** similarity functions *(Ngonga Ngomo and Auer, 2008; Isele et al. 2010)*
Not all attributes are equal

Phones/emails important in human trafficking
(names are unreliable)

Names can be important in SEC
(nothing special about phones)

How do we use this knowledge?
Domain knowledge

Especially important for unusual domains but how do we express and use it?

Use rules? Too brittle, don’t always work!

Use machine learning? Training data hard to come by, how to encode rule-based intuitions?
Probabilistic Soft Logic (PSL)
Practical methods to resolve logic/probability dilemma

Statistical Relational Learning (SRL)

- Stochastic Logic Programs (SLPs) Muggleton (1996)
- Probabilistic Relational Models (PRMs) Koller (1999)
- Bayesian Logic Programs (BLPs) Kersting and De Raedt (2001)
- Markov Logic Networks (MLNs) Richardson and Domingos (2006)
Intuitive example (PSL)

\[ \text{vote}(A,P) \land \text{friend}(B,A) \Rightarrow \text{vote}(B,P) : 0.3 \]

\[ \text{vote}(A,P) \land \text{spouse}(B,A) \Rightarrow \text{vote}(B,P) : 0.8 \]
Why PSL?

- Continuous Random Variables
- Mathematical Foundation
- Logic Foundation
- Inference & Learning
- Sets and Aggregators
- Extensible
- High Performance
Probabilistic Soft Logic (PSL)

Good framework for expressing domain constraints

Code available, many tutorials and videos on how to get started

http://psl.lings.org/index.html
Rules need to be pre-specified

But weights can be learned from training data

\[ \text{vote}(A, P) \land \text{friend}(B, A) \rightarrow \text{vote}(B, P):? \]

\[ \text{vote}(A, P) \land \text{spouse}(B, A) \rightarrow \text{vote}(B, P):? \]

Some recent work has also tried to learn rules (but may lose interpretability)
Case Study: Toponym Resolution

Toponym resolution: resolving extracted location mention to a canonical geolocation in a KB like GeoNames

Applied to human trafficking domain

...I’m from beautiful downtown London in Ontario, Canada...
Rules for toponym resolution

Can easily encode rules (in PSL) such as a city can only be in one state and one country, a city must actually exist in a state and country...

Use elements like population to assign weights
Knowledge Graphs in Latent Space
Low-dimensional vector spaces

Very popular for documents, graphs, words...
Knowledge graph embeddings

Many ways to model the problem: entities are usually vectors, relations could be vectors or matrices
Objective/loss/energy functions

What is an ‘optimal’ vector/matrix for an entity or relation?

<table>
<thead>
<tr>
<th>Model</th>
<th>Score function $f_r(h, t)$</th>
<th># Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>TransE (Bordes et al. 2013b)</td>
<td>$|h + r - t|<em>{\ell</em>{1/2}}, \ r \in \mathbb{R}^k$</td>
<td>$O(n_e k + n_r k)$</td>
</tr>
<tr>
<td>Unstructured (Bordes et al. 2012)</td>
<td>$|h - t|_2^2$</td>
<td>$O(n_e k)$</td>
</tr>
<tr>
<td>Distant (Bordes et al. 2011)</td>
<td>$|W_{rh} h - W_{rt} t|<em>1, \ W</em>{rh}, W_{rt} \in \mathbb{R}^{k \times k}$</td>
<td>$O(n_e k + 2n_r k^2)$</td>
</tr>
<tr>
<td>Bilinear (Jenatton et al. 2012)</td>
<td>$h^T W_r t, W_r \in \mathbb{R}^{k \times k}$</td>
<td>$O(n_e k + n_r k^2)$</td>
</tr>
<tr>
<td>Single Layer</td>
<td>$u_r^T f(W_{rh} h + W_{rt} t + b_r)$</td>
<td>$O(n_e k + n_r (sk + s))$</td>
</tr>
<tr>
<td></td>
<td>$u_r, b_r \in \mathbb{R}^s, W_{rh}, W_{rt} \in \mathbb{R}^{s \times k}$</td>
<td></td>
</tr>
<tr>
<td>NTN (Socher et al. 2013)</td>
<td>$u_r^T f(h^T W_r t + W_{rh} h + W_{rt} t + b_r)$</td>
<td>$O(n_e k + n_r (sk^2 + 2sk + 2s))$</td>
</tr>
<tr>
<td></td>
<td>$u_r, b_r \in \mathbb{R}^s, W_r \in \mathbb{R}^{k \times k \times a}, W_{rh}, W_{rt} \in \mathbb{R}^{s \times k}$</td>
<td></td>
</tr>
<tr>
<td>TransH</td>
<td>$|(h - w_r^T h w_r) + d_r - (t - w_r^T t w_r)|_2^2$</td>
<td>$O(n_e k + 2n_r k)$</td>
</tr>
</tbody>
</table>
Existing work

Typically evaluate on Freebase and WordNet

<table>
<thead>
<tr>
<th>Data</th>
<th>WN18</th>
<th>FB15K</th>
<th>WN11</th>
<th>FB13</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Rel</td>
<td>18</td>
<td>1,345</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>#Ent</td>
<td>40,943</td>
<td>14,951</td>
<td>38,696</td>
<td>75,043</td>
</tr>
<tr>
<td>#Train</td>
<td>141,442</td>
<td>483,142</td>
<td>112,581</td>
<td>316,232</td>
</tr>
<tr>
<td>#Valid</td>
<td>5,000</td>
<td>50,000</td>
<td>2,609</td>
<td>5,908</td>
</tr>
<tr>
<td>#Test</td>
<td>5,000</td>
<td>59,071</td>
<td>10,544</td>
<td>23,733</td>
</tr>
</tbody>
</table>

Wang et al. (2008)
## Application 1: Triples completion

<table>
<thead>
<tr>
<th>Dataset</th>
<th>WN18</th>
<th>FB15k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MEAN</td>
<td>HITS@10</td>
</tr>
<tr>
<td>Unstructured (Bordes et al. 2012)</td>
<td>315</td>
<td>304</td>
</tr>
<tr>
<td>RESCAL (Nickel, Tresp, and Kriege 2011)</td>
<td>1,180</td>
<td>1,163</td>
</tr>
<tr>
<td>SE (Bordes et al. 2011)</td>
<td>1,011</td>
<td>985</td>
</tr>
<tr>
<td>SME (Linear) (Bordes et al. 2012)</td>
<td>545</td>
<td>533</td>
</tr>
<tr>
<td>SME (Bilinear) (Bordes et al. 2012)</td>
<td>526</td>
<td>509</td>
</tr>
<tr>
<td>LFM (Jenatton et al. 2012)</td>
<td>469</td>
<td>456</td>
</tr>
<tr>
<td>TransE (Bordes et al. 2013b)</td>
<td>263</td>
<td>251</td>
</tr>
<tr>
<td>TransH (unif.)</td>
<td>318</td>
<td>303</td>
</tr>
<tr>
<td>TransH (bern.)</td>
<td>400.8</td>
<td>388</td>
</tr>
</tbody>
</table>

Wang et al. (2008)
### Application 2: Triples classification

<table>
<thead>
<tr>
<th>Dataset</th>
<th>WN11</th>
<th>FB13</th>
<th>FB15k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distant Model</td>
<td>53.0</td>
<td>75.2</td>
<td>-</td>
</tr>
<tr>
<td>Hadamard Model</td>
<td>70.0</td>
<td>63.7</td>
<td>-</td>
</tr>
<tr>
<td>Single Layer Model</td>
<td>69.9</td>
<td>85.3</td>
<td>-</td>
</tr>
<tr>
<td>Bilinear Model</td>
<td>73.8</td>
<td>84.3</td>
<td>-</td>
</tr>
<tr>
<td>NTN</td>
<td>70.4</td>
<td>87.1</td>
<td>66.5 (≈ 40h)</td>
</tr>
<tr>
<td>TransE (unif.)</td>
<td>75.85</td>
<td>70.9</td>
<td>79.7 (≈ 5m)</td>
</tr>
<tr>
<td>TransE (bern.)</td>
<td>75.87</td>
<td>81.5</td>
<td>87.3 (≈ 5m)</td>
</tr>
<tr>
<td>TransH (unif.)</td>
<td>77.68</td>
<td>76.5</td>
<td>80.2 (≈ 30m)</td>
</tr>
<tr>
<td>TransH (bern.)</td>
<td>78.80</td>
<td>83.3</td>
<td>87.7 (≈ 30m)</td>
</tr>
</tbody>
</table>

*Wang et al. (2008)*
Code availability

Code for replicating experiments can be found at https://github.com/glorotxa/SME; implemented using both theano/tensorflow backend.

Unclear how to extend to new, sparse data, how to scale to much bigger KGs.
Application 3: Featurizing nodes

E.g. Convering ‘locations’ into feature vectors

Relevant for toponym resolution, building rich graphs...


https://github.com/mayankkejriwal/Geonames-embeddings
Features encode spatial proximity

But could encode much else, lots of room for new research!
In unusual domains...

Not clear to what extent these techniques work (which objective function is better?)

Several questions persist

- How to acquire training data for triples classification?
- How to model the concept of an entity?
- How to regularize with sparse mentions?
Searching Knowledge Graphs
Motivation

Direct query execution does not succeed because of noise/incomplete data

What is the phone number provided in the ad that contains the email address brianna.elite@dumy-email.com, the price $600/hour, and the location manhattan midtown east 33rd & 2nd ave?
Keyword vs. faceted search
Faceted search form

Kejriwal, Szekely

127
Other kinds of search: many research opportunities

Clustering and network queries

Find all massage parlor ads linked either directly or indirectly to phone 12345678

Aggregate queries

Find the most common ethnicity in the massage parlor ads linked either directly or indirectly to phone 12345678
**Key technique: Query reformulation**

Traditional work tends to reformulate if original query returns *no results*. Does not work well in presence of noise or in interactive session.

Use *weighted combination* of reformulated queries instead, execute in NoSQL database.
Semantic Query Reformulation

Original query

Conservative

Relaxed

Keyword-only

Knowledge graph stored in NoSQL database

Ranked list of results

Execute

Weighted tree of semantically reformulated queries

\[ w_1, w_2, w_3, w_4 \ldots \]
Other applicable techniques

Entity set learning and expansion

E.g., can expand red to red, auburn, fiery...

Use machine learning to learn weights of reformulated sub-queries

Relatively unexplored research area: similar to learning to rank in IR community, but weights strategies, not features
Wrap up and review
Summary

Unusual domains
Interesting, fun, high social impact

Knowledge graphs
Wide-spectrum of representations

Knowledge graph construction
Many tools to help, easy to create interesting graphs

Case study
Tools make it possible to build a KG in a day

Knowledge graph completion
Inferring wrong and missing links

Entity resolution
Clustering/linking mentions referring to the same underlying entity

Probabilistic soft logic
Correcting outputs of KG construction and entity resolution through logical probabilistic rules

KGs in latent space
Performing inference in low-dimensional vector spaces

Searching knowledge graphs
Using semantic query reformulation to answer complex queries over noisy knowledge graphs
Thank you!