KNOWLEDGE VAULT: A WEB-SCALE APPROACH TO PROBABILISTIC KNOWLEDGE FUSION

Navid Hashemi

Spring 2015

OVERVIEW

- Introduction
- Components of Knowledge Vault
- Extraction Methods
- Knowledge Fusion
- Summary

INTRODUCTION

- Knowledge Vault:

1. combines extractions from Web content (obtained via analysis of text, tabular data, page structure, and human annotations)
2. supervised machine learning methods for fusing these distinct information sources
3. substantially bigger than any previously published structured knowledge repository
4. features a probabilistic inference system that computes calibrated probabilities of fact correctness

PREVIOUS APPROACHES

- Text-based extractions which can be very noisy

NEW APPROACH

- Web-scale probabilistic knowledge base, fusing web contents with prior knowledge

INTRODUCTION

- Current large-scale knowledge bases, still far from complete
- Freebase: Largest Open Source Knowledge Base
  - 71% of people in Freebase have no known place of birth
  - 75% have no known nationality
- Previous approaches primarily relied on direct contributions from human volunteers and integration of existing repositories of structured knowledge
- Yield head content, frequently mentioned properties of frequently mentioned entities

KNOWLEDGE VAULT (KV)

- stores information in the form of RDF triples
  - (subject, predicate, object)
  - </m/02mjmr, /place_of_birth /m/02hrh0_>
  - /m/02mjmr: Freebase ID for Barak Obama
  - /m/02hrh0_: Freebase ID for Austin

- Associated with each such triple is a confidence score, representing the probability that KV "believes" the triple is correct

MAIN CONTRIBUTIONS OF KV

1. It combines noisy extractions from the Web together with prior knowledge

   - analogous to techniques used in speech recognition
   - overcome errors due to the extraction process, as well as errors in the sources themselves (mutual improvements)

   - Extractor: Barack Obama was born in Kenya
   - Prior Model: Obama is the US President

   - Mistake with somebody else with this name
   - Erroneous statement on a spammy website

MAIN CONTRIBUTIONS OF KV

2. Much bigger than other comparable KBs

   - KV has 1.6B triples, of which 324M have a confidence of 0.7 or higher
   - 271M have a confidence of 0.9 or higher
   - 38 times more than the largest previous comparable system: DeepDive

MAIN CONTRIBUTIONS OF KV

3. Perform a detailed comparison of the quality and coverage of different extraction methods
   - Use multiple extraction sources and systems

- Extracting facts from a large variety of sources of Web data, including:
  - Free text
  - HTML DOM trees
  - HTML Web tables
  - Human annotations of Web pages

COMPONENTS OF KV

1. **Extractors**: extract triples from a huge number of Web sources & assigns a confidence score to an extracted triple
   - representing uncertainty about the identity of the relation and its corresponding arguments

2. **Graph-based Priors**: learn the prior probability of each possible triple, based on triples stored in an existing KB

3. **Knowledge Fusion**: computes the probability of a triple being true, based on agreement between different extractors and priors

WEIGHTED LABELED GRAPH

- To represent Knowledge Vault:
  - Construct a weighted labeled graph, which we can view as a very sparse E*P*E 3D matrix G
  - G(s, p, o) = 1 if there is a link of type p from **s** to **o**
  - G(s, p, o) = 0 otherwise

- We want to compute Pr(G(s, p, o) = 1 |.) for candidate (s, p, o) triples, where the probability is conditional on different sources of information
  1. Using extraction: condition on text features about the triple
  2. Using graph-based priors: condition on known edges in the Freebase graph
  3. In knowledge fusion: condition on both text extractions and prior edges

LOCAL CLOSED WORLD ASSUMPTION

- (s, p, o) triples that are in Freebase => label is true
- triples that do not occur in Freebase => label is false

- Define O(s, p) as the set of existing object values for a given s and p
- if $(s, p, o) \in O(s, p)$ => triple is correct
- if $(s, p, o) \notin O(s, p)$ but $|O(s, p)| > 0$ => triple is incorrect
  - We assumed KB is locally complete for this Subject-Predicate pair

- More sophisticated methods of training => future work

EXTRACTION SOURCES

1. Text Documents (TXT)
2. HTML Trees (DOM)
3. HTML Tables (TBL)
4. Human Annotated Pages (ANO)

TEXT DOCUMENTS (TXT)

- standard methods for relation extraction from text => much larger scale than previous systems

1. Run standard methods for entity recognition and linking them together
2. Find seed set of entity pairs that have this predicate
   - For example, if the predicate is **married_to**, the pairs could be (BarackObama, MichelleObama) and (BillClinton, HillaryClinton)

HTML TREES (DOM)

- Document Object Model: a cross-platform and language-independent convention for representing and interacting with objects in HTML, XHTML and XML documents
  - A different way to extract information from Web pages is to parse their DOM trees
  - Nodes of every document are organized in a tree structure

- Data Sources:
  1. text pages
  2. deep web sources

HTML TABLES (TBL)

- 570M tables on the Web that contain relational information

- Fact extraction techniques developed for text and trees do not work well for tables
  - relation between two entities is usually contained in the column header, rather than being close by in the text/ tree

1. Perform entity linkage
2. Identify the relation that is expressed in each column of table
   1. Reasoning about which predicate each column could correspond to
   2. Matching to Freebases

- Discard ambiguous columns

HUMAN ANNOTATED PAGES (ANO)

- Webpages where the web master has added manual annotations and ontologies
- Schema.org => collection of schemas that webmasters can use to markup HTML pages in ways recognized by major search providers
- Not fully implemented yet in Knowledge Vault

FUSING THE EXTRACTORS

- A simple way to combine our extracted data is to construct a feature vector f(t) for each extracted triple t = (s, p, o), and then to apply a binary classifier to compute Pr(t = 1 | f(t))

- Feature vector composed of two numbers for each extractor:
  - The square root of the number of sources that the extractor extracted this triple from
    - to reduce the effect of very commonly expressed facts
  - The mean score of the extractions from this extractor, averaging over sources

PROBABILITY ESTIMATES

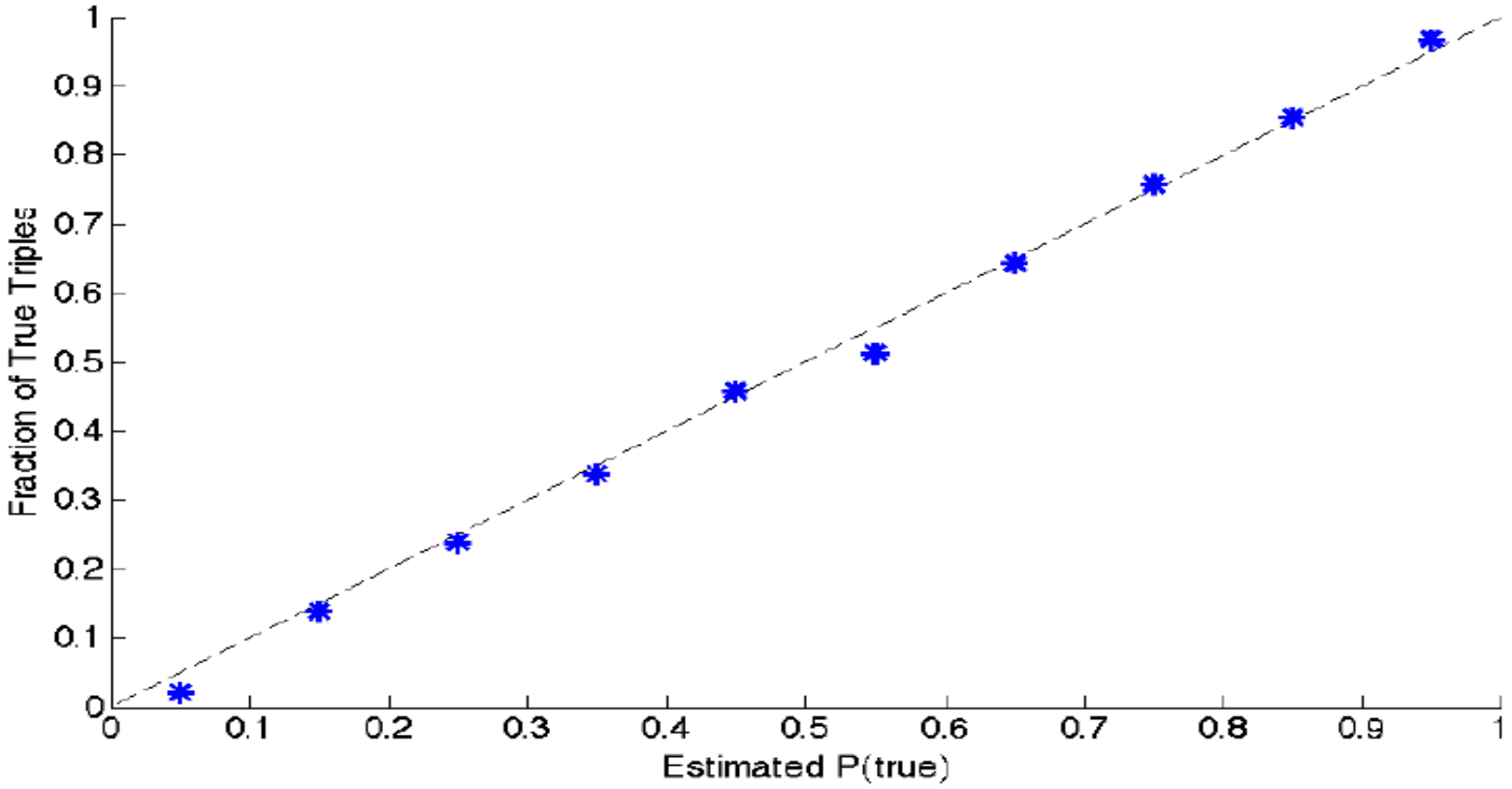


Figure 1: True probability vs estimated probability for each triple in KV.

ADDING MORE EVIDENCE

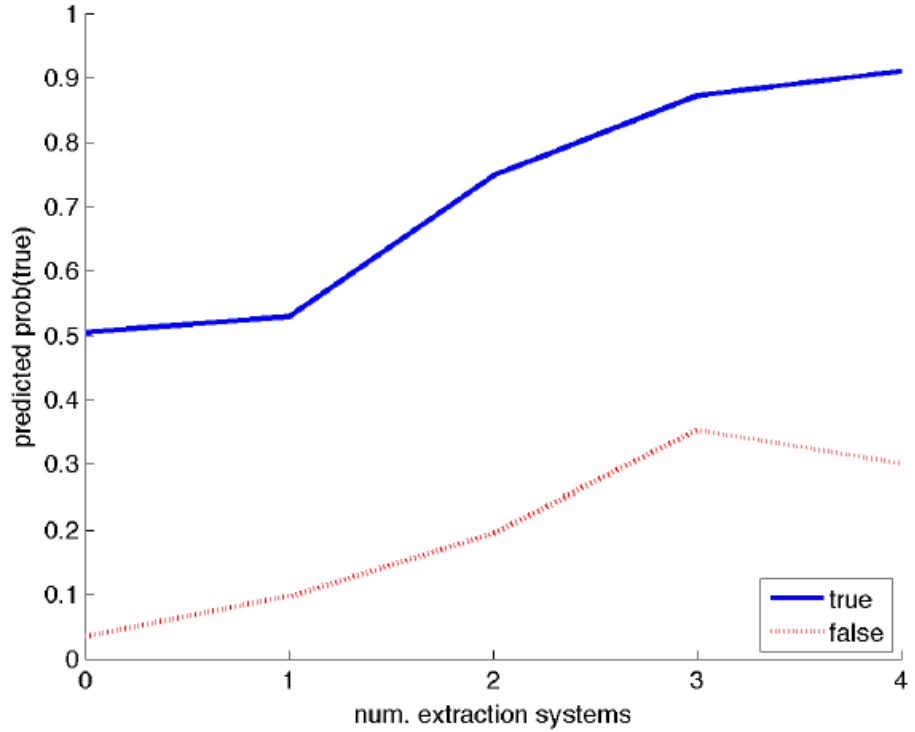


Figure 2: Predicted probability of each triple vs. the number of systems that predicted it. Solid blue line: correct (true) triples. Dotted red line: incorrect (false) triples.

ADDING MORE EVIDENCE

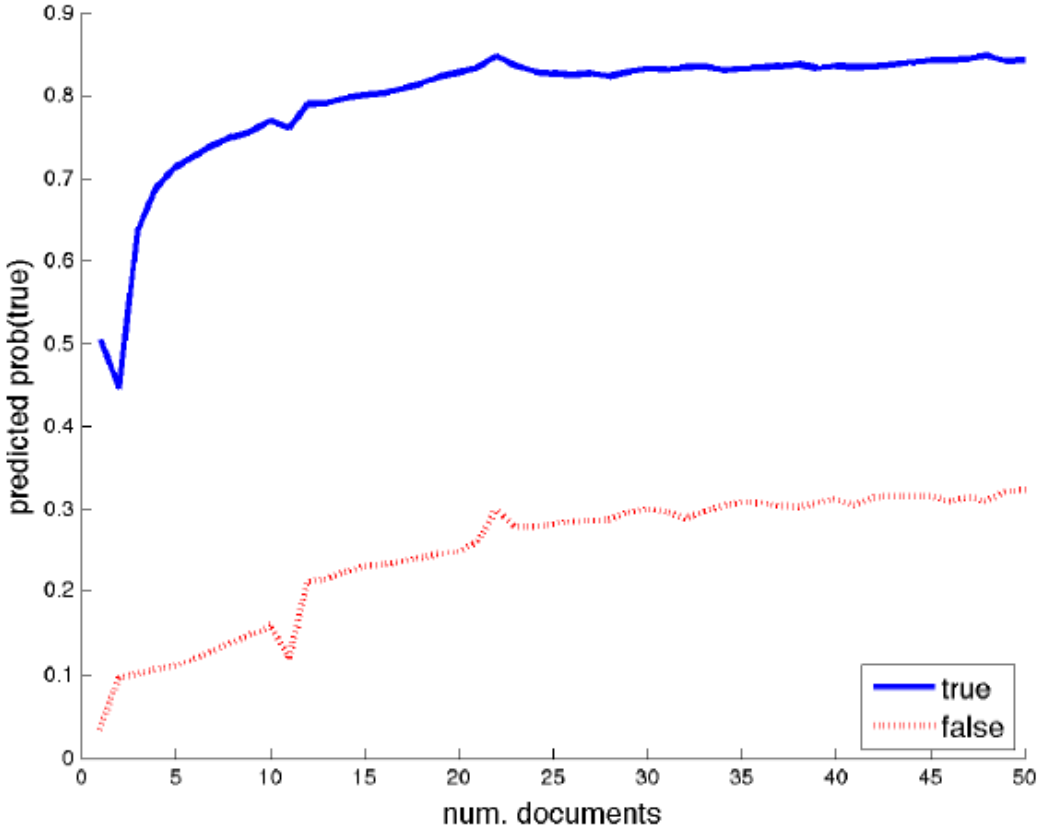


Figure 3: Predicted probability of each triple vs. the number of unique web sources that contain this triple (axis truncated at 50 for clarity).

GRAPH BASE PRIORS

- Facts extracted from the Web can be unreliable => use prior knowledge (Freebase)
- Assign a probability to triples came from Freebase => even if there is no corresponding evidence for a fact on the Web
- Like link-prediction in graphs
  - Observe a set of existing edges
  - Want to predict which other edges are likely to exist

- Two different approaches to solve the problem of Link Prediction:
  - Path Ranking Algorithms
  - Neural Network Model

PATH RANKING ALGORITHM (PRA)

- Start with a set of pairs of entities that are connected by some predicate p
- Performs a random walk on the graph, starting at all the subject (source) nodes
- Paths that reach the object (target) nodes are considered successful
- Example: algorithm learns that pairs (X,Y) which are connected by a "marriedTo" edge often also have a path of the form:

$$X \xrightarrow{\text{parentOf}} Z \xleftarrow{\text{parentOf}} Y$$

PATH RANKING ALGORITHM (PRA)

- paths that PRA learns can be interpreted as rules:

F1	P	R	W	Path
0.03	1	0.01	2.62	/sports/drafted-athlete/drafted,/sports/sports-league-draft-pick/school
0.05	0.55	0.02	1.88	/people/person/sibling-s, /people/sibling-relationship/sibling, /people/person/education, /education/education/institution
0.06	0.41	0.02	1.87	/people/person/spouse-s, /people/marriage/spouse, /people/person/education, /education/education/institution
0.04	0.29	0.02	1.37	/people/person/parents, /people/person/education, /education/education/institution
0.05	0.21	0.02	1.85	/people/person/children, /people/person/education, /education/education/institution
0.13	0.1	0.38	6.4	/people/person/place-of-birth, /location/location/people-born-here, /people/person/education, /education/education/institution
0.05	0.04	0.34	1.74	/type/object/type, /type/type/instance, /people/person/education, /education/education/institution
0.04	0.03	0.33	2.19	/people/person/profession, /people/profession/people-with-this-profession, /people/person/education, /education/education/institution

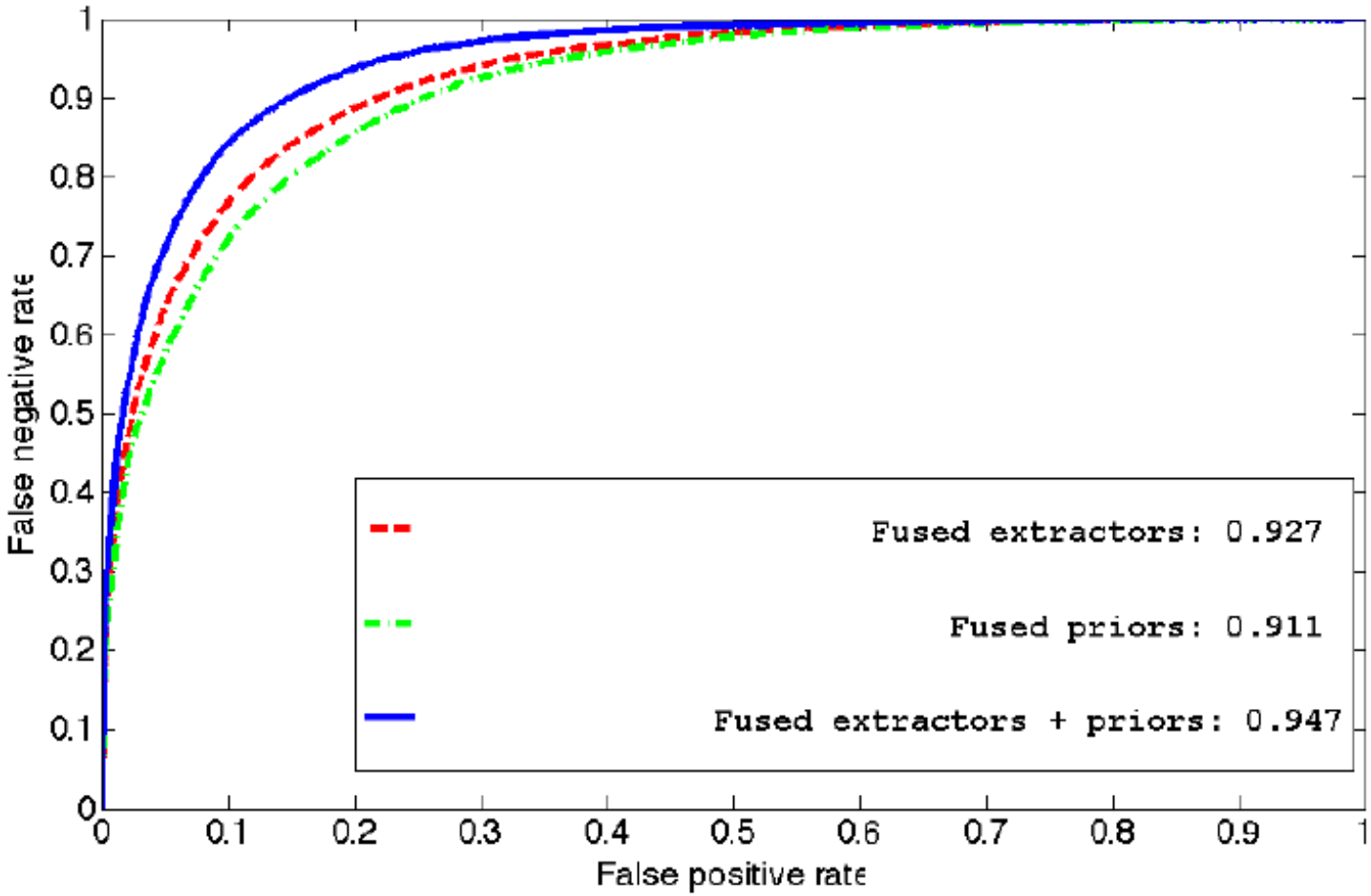
Table 3: Some of the paths learned by PRA for predicting where someone went to college. Rules are sorted by decreasing precision. Column headers: F1 is the harmonic mean of precision and recall, P is the precision, R is the recall, W is the weight given to this feature by logistic regression.

- First rule says: a person X is likely to have attended school S if X was drafted from sports team T, and T is from school S
- Second rule says: a person is likely to attend the same school as their sibling

NEURAL NETWORK MODEL

- An alternative approach to building the prior model is to view the link prediction problem as matrix completion
- original KB can be viewed as a very sparse E*P*E 3D matrix G
  - E is the number of entities
  - P is the number of predicates
  - G(s, p, o) = 1 if there is a link of type p from s to o
  - G(s, p, o) = 0 otherwise

FUSING THE PRIORS AND EXTRACTORS



FUSING THE PRIORS AND EXTRACTORS

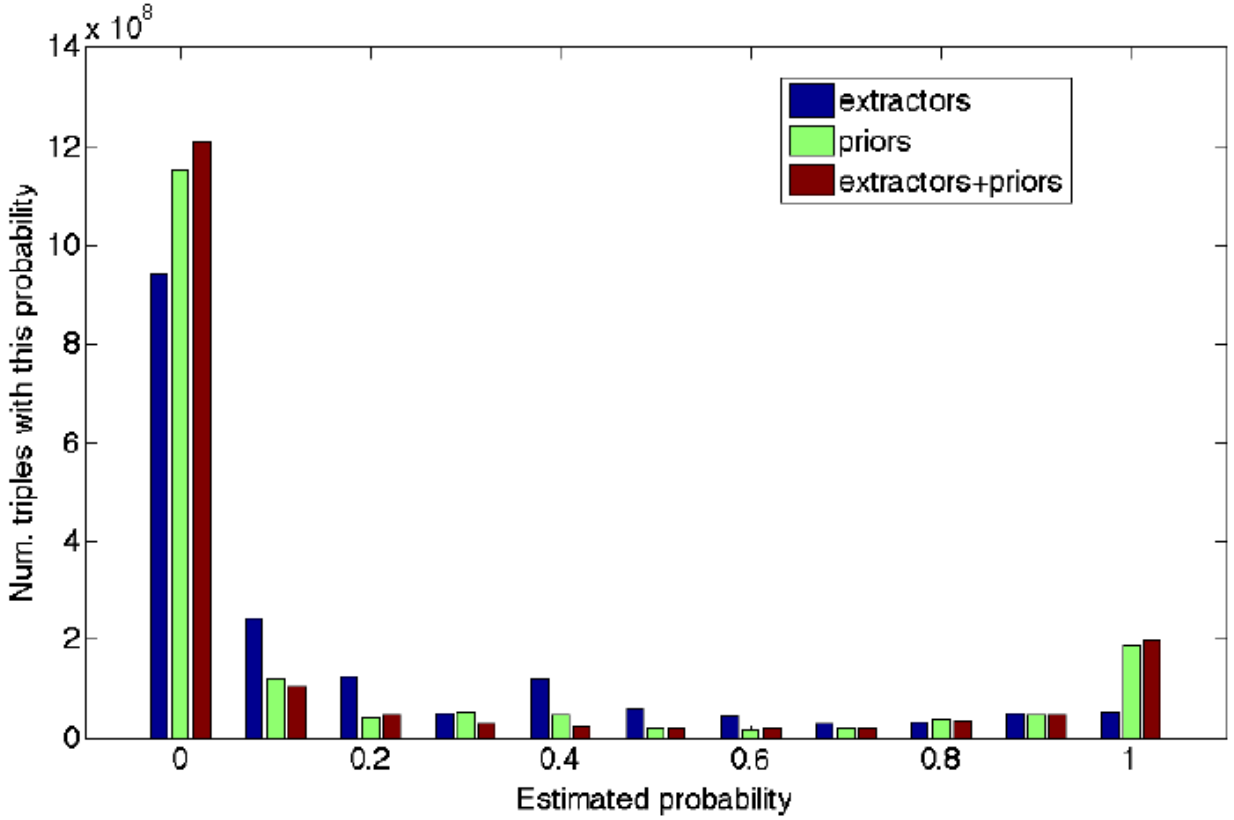


Figure 5: Number of triples in KV in each confidence bin.

FUSING THE PRIORS AND EXTRACTORS

```
<Barry Richter  (/m/02ql38b),
/people/person/edu./edu/edu/institution,
Universty of Wisconsin-Madison (/m/01yx1b)>
```

- The (fused) extraction confidence for this triple was just **0.14**, since it was based on the following two rather indirect statements:

```
In the fall of 1989, Richter accepted a scholarship
to the University of Wisconsin, where he played for
four years and earned numerous individual accolades...

The Polar Caps' cause has been helped by the impact of
knowledgable coaches such as Andringa, Byce and former
UW teammates Chris Tancill and Barry Richter.
```

- Freebase: Barry Richter was born and raised in Madison, WI => increase in our prior belief => Final fused belief of **0.61**

EVALUATING LCWA

- Local Closed World Assumption => just an approximation of the Truth

- Freebase: list only top 5 actors for any given movie => False Negatives
- Freebase: contain some errors => False Positives

- More complete approach needed to work further, beyond local closed world assumptions

4 APPROACHES ON AUTOMATIC KB CONSTRUCTIONS

1. Built on Wikipedia and other structured Data Sources
   - YAGO, Dbpedia, Freebase

2. Use open information (schema-less) extraction techniques applied to the entire web
   - Reverb, OLLIE, PRISMATIC

3. Extract information from the entire web, but use a fixed ontology/ schema
   - NELL, ReadTheWeb, PROSPERA, DeepDive

4. Construct is-a hierarchies, as opposed to general KBs with multiple types of predicates
   - Probase

DIFFERENCE BETWEEN KV & EXISTING APPROACH

- knowledge vault is most similar to methods of the third kind, which extract facts, in the form of disambiguated triples, from the entire web
- main difference from this prior work is that we fuse together facts extracted from text with prior knowledge derived from the Freebase graph
- KV is a probabilistic database and support simple queries such as:
  - BarakObama BornIn ?
  - Returns distribution over places where KV thinks Obama was born
  - More sophisticated queries, JOIN/ SELECT => to be implemented
- KV represents uncertainty in the facts it has extracted

IMPROVEMENTS NEEDED TO BE DONE

- Modeling mutual exclusion between facts
- Modeling soft correlation between facts
- Values can be represented at multiple levels of abstraction
- Dealing with correlated sources
- Some facts are only temporarily true
- Adding new entities and relations
- Knowledge representation issues
- Inherent upper bounds on the potential amount of knowledge that we can extract

MODELING MUTUAL EXCLUSION BETWEEN FACTS

- Currently treat each fact as an **independent binary random** variable, that is either true or false
- In reality, many triples are correlated
- Functional relation such as **born-in =>** only one true value

- Barak Obama was born in Austin, Texas, US => these are not mutually exclusive => need more sophisticated approach

MODELING SOFT CORRELATION BETWEEN FACTS

- Number of children: usually between 0 and 5
- Date of Birth: 15-50 years earlier than their children birth dates
- Previous approaches: use Gaussian models to represent correlations among numerical values => more accurate approach needed to be integrated to KV

CORRELATED SOURCES

- Our belief in a triple increase as we see it from more sources
- Current approach: counting each domain only once
- More sophisticated copy detection approach is needed

SOME FACTS ARE ONLY TEMPORARILY TRUE

- Current CEO of Google: Larry Page
- Google's CEO from 2001 to 2011: Eric Schmidt

ADDING NEW ENTITIES AND RELATIONS

- Many entities exists on the Web, but are not in Freebase
- Some relations could not be mapped to Freebase schema

KNOWLEDGE REPRESENTATION ISSUES

- RDF => good option for factual assertions
- How about representing the difference between Jazz music and Blues?!

WEB IS NOT ENOUGH TO GATHER ALL KNOWLEDGE

- Goal of KV: large scale repository of all human knowledge
- Some crowd-sourcing techniques to acquire knowledge

SUMMARY

- A Web-scale probabilistic knowledge base, called Knowledge Vault is introduced
- In contrast to previous works, we fuse together multiple extraction sources with prior knowledge derived from an existing KB
- The resulting knowledge base is about 38 times bigger than existing automatically constructed KBs
- The facts in KV have associated probabilities, which we show are well-calibrated, so that we can distinguish what we know with high confidence from what we are uncertain about
- In the future, we hope to continue to scale KV, to store more knowledge about the world

THANK YOU..