Hazy

Lixing Lian, Cheng Ren
Outline

- Motivation & Goal
- Framework & Design
- Examples
- Future Work
- Conclusion
Two Trends that Drive Hazy

- Data in a large number of formats
  - (text, audio, video, OCR, sensor data, etc.)
- Arms race to deeply understand data

Statistical tools attack both 1. and 2.

Hazy = statistical + data management
Hazy’s Thesis

- The next breakthrough in data analysis
doesn’t necessarily mean a new data analysis algorithm...
- ...but may be in the ability to rapidly combine, deploy, and maintain existing algorithms.
Hazy’s Goal

- Making big-data analytics-driven systems easier to build and maintain.
- Find common patterns when deploying statistical tools on data.
  - Programming abstractions
  - Infrastructure abstractions
Hazy's Programming Abstractions and Infrastructure Abstractions

- programming abstractions
- infrastructure abstractions
- applications
- algorithms
- implementations
Programming abstractions

- Enable developers to try many algorithms for the same data set.
- One algorithm improves, all applications using that algorithm automatic improve.
Infrastructure abstractions

- No need to reinvent or reengineer the wheel when adding a new algorithm to the system
- One component of the infrastructure improved, all algorithms benefit automatically
Markov logic

- Easily represent common statistical models: logistic regression and conditional random fields
- Build more sophisticated statistical models
Markov Logic by Example

Rules

\[ \text{wrote}(s,t) \land \text{advisedBy}(s,p) \Rightarrow \text{wrote}(p,t) \]

//students’ papers tend to be co-authored by advisors

\[ \text{advisedBy}(s,p) \land \text{advisedBy}(s,q) \Rightarrow p = q \]

\[ \text{advisedBy}(s,p) \Rightarrow \text{professor}(p) \]

...

Evidence

\[ \text{wrote}(\text{Tom, Paper1}) \]

\[ \text{wrote}(\text{Tom, Paper2}) \]

\[ \text{wrote}(\text{Jerry, Paper1}) \]

\[ \text{professor}(\text{John}) \]

...
Markov Logic by Example

\[ \text{wrote}(s, t) \land \text{advisedBy}(s, p) \rightarrow \text{wrote}(p, t) \]

Step 1: Grounding

\begin{align*}
\text{wrote}(\text{Tom}, \text{P1}), \text{advisedBy}(\text{Tom}, \text{Jerry}) & \rightarrow \text{wrote}(\text{Jerry}, \text{P1}) \\
\text{wrote}(\text{Tom}, \text{P1}), \text{advisedBy}(\text{Tom}, \text{Bob}) & \rightarrow \text{wrote}(\text{Bob}, \text{P1}) \\
\text{wrote}(\text{Bob}, \text{P1}), \text{advisedBy}(\text{Bob}, \text{Jerry}) & \rightarrow \text{wrote}(\text{Jerry}, \text{P1})
\end{align*}

Find the field and extract data

advisee | advisor
---|---
Tom | Jerry
Tom | Bob
Grounding via SQL in Tuffy

Program Transformed into many SQL queries (Bottom-up)

\[
\text{wrote}(s, t) \land \text{advisedBy}(s, p) \rightarrow \text{wrote}(p, t)
\]

```
SELECT w1.id, a.id, w2.id
FROM wrote w1, advisedBy a, wrote w2
WHERE w1.person = a.advisee AND w1.paper = w2.paper
AND a.advisor = w2.person AND ...
```

An RDBMS
Grounding: Top-down vs. Bottom-up

Comparison of In-Memory Grounding of Markov Logic with In-RDBMS Grounding

\[
\text{wrote}(s,t) \land \text{advisedBy}(s,p) \Rightarrow \text{wrote}(p,t)
\]

\text{wrote(person,paper)}: \text{known} \quad \text{advisedBY (advisee,advisor)}: \text{unknown}

For each person \(s\):
   For each paper \(t\):
      If \(!\text{wrote}(s,t)\): continue
      For each person \(p\):
         If \(\text{wrote}(p,t)\): continue
         Emit grounding \(<s,t,p>\)

Alchemy: C++ loops in RAM

runtime: hours

SELECT \(w1.id, a.id, w2.id\)
FROM \(\text{wrote} w1, \text{advisedBy} a, \text{wrote} w2\)
WHERE \(w1.\text{truth} \text{AND NOT} w2.\text{truth}\)
\text{AND} \(w1.\text{person} = a.\text{advisee}\)
\text{AND} \(w1.\text{paper} = w2.\text{paper}\)
\text{AND} \(a.\text{advisor} = w2.\text{person}\)

Tuffy: SQL in RDBMS

runtime: seconds
(a) input data sources
- Macrostrat taxonomy
- Freebase location lat-longs
- Google search results
- 20k geology papers
- document annotations

(b) feature extraction
- NLP parsing
- pattern matching
- dictionary matching
- mention extraction
- heuristic coreference

(c) statistical processing

\[
\begin{align*}
\text{IsPerson}(x) &\not\rightarrow \text{IsFormation}(x) \\
\text{IsFormation}(x) \land \text{IsLocation}(y) &\land \text{InSameSentence}(x, y) \\
& \rightarrow \text{LocatedIn}(x, y)
\end{align*}
\]

(d) output probabilistic predictions

<table>
<thead>
<tr>
<th>prob.</th>
<th>formation</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.98</td>
<td>Barnette</td>
<td>Texas</td>
</tr>
<tr>
<td>0.97</td>
<td>Husky</td>
<td>Pie River</td>
</tr>
<tr>
<td>0.84</td>
<td>Atoka</td>
<td>Alberta</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

Hazy infrastructure
machine learning
statistical inference
Example 1: DeepDive

- Enrich Wikipedia with structured data that is extracted from both unstructured sources
DeepDive

DeepDive’s Origin

- Build a system that is able to read the Web and answer questions.

- Machine Reading: “List members of the Brazilian Olympic Team in this corpus with years of membership”
DeepDive

What about Barack Obama?
- wife is Michelle Obama
- went to Harvard Law School
- ...
DeepDive

What about Barack Obama?
* wife is Michelle Obama
* went to Harvard Law School
* ...

Billions of webpages
Billions of tweets
Billions of events
Billions of videos
Billions of photos
Billions of blogs
Sample Relations about Barack Obama, Elon Musk, and Microsoft Extracted by Deepdive

- attended the school
  Harvard Law School [16]

- married to
  Michelle Obama [1678]

- is a child of
  Barack Obama, Sr. [8]

- is a parent of
  Malia Obama [18]
  Sasha Obama [8]
  (Dive)

- founded
  SpaceX [35]
  PayPal [4]
  Tesla Motors [3]

- has the title
  Architect [56]
  Chief executive of. [28]
  Executive Officer [14]
  Entrepreneur [12]
  (Dive)

- has the top member
  Steve Ballmer [14]
  Bill Gates [6]

- may have subsidiaries
  Bungie [98]
  Massive Incorporated [91]
  Tellme Networks [71]
  Razorfish (company) [51]
  Skype [48]
  MSNBC [39]
  (Dive)
DeepDive

Given a name, collects all the information related to this name and display together.
DeepDive

Demo

DeepDive: Demo

Tasks it performs:
- Web Crawling
- Information Extraction
- Deep Linguistic Processing
- Audio/Video Transcription
- Tera-byte Parallel Joins

Some Information:
- 50TB Data
- 500K Machine hours
- 500M Webpages
- 400K Videos
- 7Bn Entity Mentions
- 114M Relationship Mentions

Declare graphical models at Web scale
Example 2: GeoDeepDive

- [http://hazy.cs.wisc.edu/demo/geo/](http://hazy.cs.wisc.edu/demo/geo/)

- The goal is to help geo-scientists extract data that is buried in the text, tables, and figures of journal articles and web sites, sometimes called dark data.

- Extends a database called Macrostrat.
Future work

- Assisted Development
  - expertise, experience of data and algorithms
- New Data Platforms
  - Hadoop environment
Conclusion

- Key technical hypothesis: A large fraction of the processing performed by applications that use and analyze these new sources of data can be captured using a small handful of primitives.

- Hazy group is building several applications

- More information:
  
  http://hazy.cs.wisc.edu/hazy/