Applications of Rule Mining in Knowledge Bases

Luis Galárraga

November 3rd, 2014
PIKM, Shanghai
Knowledge Bases (KBs)
KBs in action

Barack Obama - Wikipedia, the free encyclopedia
Barack Hussein Obama II (US /bəˈrɑːk oʊˈbɑːrəmaɪ/ UK /ˈbærək əˈbɑːrəmaɪ/; born August 4, 1961) is the 44th and current President ... Early life and career - Legislative career - Presidential campaigns
en.wikipedia.org/wiki/Barack_Obama • 2014-10-28

Barack Obama - Official Site
“You and I, as citizens, have the power to set this country’s course.” —President Obama. Take Action
www.barackobama.com •

Images of barack obama

Barack Obama - Biography - Lawyer, U.S. Representative, ...
Some popular KBs
Rule Mining in KBs

Barack Obama

Malia

Sasha

Michelle

marriedTo

born On

Aug 4, 1961

hasChild
Rule Mining in KBs

X

Malia

Sasha

Y

Z

born On Aug 4, 1961
Rule Mining in KBs

Barack Obama
- hasChild: Malia
- hasChild: Sasha
- marriedTo: Michelle

Malia
- hasChild: Barack Obama

Sasha
- hasChild: Barack Obama

Michelle
- marriedTo: Barack Obama

Born On: Aug 4, 1961
Rule Mining in KBs

- Malia
- X
- Y
- Z
- Aug 4, 1961

Relationships:
- hasChild: Malia → X, Malia → Z, Y → X, Y → Z
- marriedTo: Y → Z
- born On: Z → Aug 4, 1961
Rule Mining in KBs

hasChild(y, x), marriedTo(y, z) => hasChild(z, x)
Rule Mining in KBs

KBs are often incomplete

Elvis Presley

Lisa Marie

marriedTo

Priscilla

hasChild
Rule Mining in KBs

Rules can be used to make predictions

hasChild(y, x), marriedTo(y, z) => hasChild(z, x)
Rule Mining in KBs

Missing information is counter-evidence under the Closed World Assumption
Rule Mining in KBs

KBs operate under the Open World Assumption

Elvis Presley

Lisa Marie

Priscilla

[UNKNOWN]
Partial Completeness Assumption (PCA)

Malia

Sasha

Michelle
Partial Completeness Assumption (PCA)
PCA for Rule Mining

\[
\text{hasChild}(y, x), \text{marriedTo}(y, z) \Rightarrow \text{hasChild}(z, x)
\]
PCA for Rule Mining

hasChild(y, x), marriedTo(y, z) => hasChild(z, x)

<table>
<thead>
<tr>
<th>Hits</th>
<th>Misses</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
PCA for Rule Mining

\[\text{hasChild}(y, x), \text{marriedTo}(y, z) \Rightarrow \text{hasChild}(z, x)\]

- **Hits**: 2
- **Misses**: 0
PCA for Rule Mining

(hasChild(y, x), marriedTo(y, z) => hasChild(z, x))

Hits: 2
Misses: 0
hasChild(y, x), marriedTo(y, z) => hasChild(z, x)

Hits: 2
Misses: 1
PCA for Rule Mining

Elvis Presley

hasChild

Lisa Marie

marriedTo

Priscilla

hasChild(y, x), marriedTo(y, z) => hasChild(z, x)

Hits

2

Misses

1
PCA for Rule Mining

hasChild(y, x), marriedTo(y, z) \implies hasChild(z, x)

Hits
2

Misses
1
PCA for Rule Mining

hasChild(y, x), marriedTo(y, z) => hasChild(z, x)

Hits
2

Misses
1
PCA for Rule Mining

\[(\text{hasChild}(y, x), \text{marriedTo}(y, z)) \Rightarrow \text{hasChild}(z, x)\]

- **Hits**: 2
- **Misses**: 1

**Standard Confidence**: \(\frac{2}{4} = 50\%\)

**PCA Confidence**: \(\frac{2}{3} = 66.67\%\)
AMIE: Association Rule Mining Under Incomplete Evidence

• AMIE is a system that learns **closed Horn rules**
  
  \[ \text{hasChild}(y, x), \text{marriedTo}(y, z) \Rightarrow \text{hasChild}(z, x) \]

• It performs exhaustive search based on:
  
  - Minimum support threshold
  - Mining operators
  - Monotonicity of support for pruning
  - Optimized in-memory database
  - Confidence gain is used to prune the output.

hasChild(y, x), marriedTo(y, z) $\Rightarrow$ hasChild(z, x)
Add dangling atom \((O_D)\)

```
marriedTo influences ....
```

marriedTo influences ...

Add dangling atom \((O_D)\)
Add dangling atom ($O_D$)

marriedTo influences ...

Add closing atom ($O_C$)

hasChild supervises ...

hasChild

marriedTo

$r$

$z$

$y$

$z$

$y$

$z$

$y$

$z$

$x$

$x$

$x$

$x$
Add dangling atom ($O_D$)

marriedTo
influences
...

Add closing atom ($O_C$)

hasChild
supervises
...

marriedTo

$\therefore$
Add dangling atom \((O_D)\)

Add closing atom \((O_C)\)

\[
\text{marriedTo influences } \ldots
\]

\[
\text{marriedTo(y, z), marriedTo(y, z) } \Rightarrow \text{hasChild(z, x)}
\]
AMIE: Association Rule Mining Under Incomplete Evidence

Minimum support threshold

RDF KB

Tailored In-memory DB

Concurrent mining implementation
AMIE: Association Rule Mining Under Incomplete Evidence

PCA Confidence used to rank rules

Minimum support threshold

RDF KB

Tailored In-memory DB

Concurrent mining implementation
AMIE: Association Rule Mining Under Incomplete Evidence

Some rules mined by AMIE on YAGO:

\[
\begin{align*}
\text{isMarriedTo}(x, y) \land \text{livesIn}(x, z) & \Rightarrow \text{livesIn}(y, z) \\
\text{isCitizenOf}(x, y) & \Rightarrow \text{livesIn}(x, y) \\
\text{hasAdvisor}(x, y) \land \text{graduatedFrom}(x, z) & \Rightarrow \text{worksAt}(y, z) \\
\text{hasWonPrize}(x, \text{Gottfried Wilhelm Leibniz Prize}) & \Rightarrow \text{livesIn}(x, \text{Germany})
\end{align*}
\]
AMIE: Association Rule Mining Under Incomplete Evidence

AMIE finds rules in medium-size ontologies in a few minutes.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Facts</th>
<th>Runtime</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAGO2</td>
<td>1M</td>
<td>3.62min</td>
<td>138</td>
</tr>
<tr>
<td>YAGO2 (const)</td>
<td>1M</td>
<td>17.76min</td>
<td>18K</td>
</tr>
<tr>
<td>Dbpedia (2 atoms)</td>
<td>6.7M</td>
<td>2.89min</td>
<td>6.9K</td>
</tr>
</tbody>
</table>
AMIE: Association Rule Mining Under Incomplete Evidence

PCA confidence better for prediction than standard confidence.

![Graph showing comparison between PCA Confidence and Standard Confidence over aggregated predictions, top 30 rules.](image-url)
Rules for Ontology Schema Alignment

Rule mining can be used for data integration.
Rules for Ontology Schema Alignment

Use instance alignments to align the schemas
Rules for Ontology Schema Alignment

hasChild(x, y) <=> parent(y, x)

KB 1

Malia

hasChild

Barack Obama

sameAs

Sasha

KB 2

Malia

sameAs

parent

President Obama

sibling

Sasha

sameAs

parent
Rules for Ontology Schema Alignment

hasChild(y, x) hasChild(y, z) => sibling(x, z)
Rules for Ontology Schema Alignment

Run AMIE on a coalesce of the KBs

Malia

Barack Obama

Sasha

\[
\text{hasChild} \iff \text{parent}^1
\]

\[
\text{hasChild}(y, x) \text{ hasChild}(y, z) \implies \text{sibling}(x, z)
\]
ROSA rules

ROSA rules are a class of cross-ontology alignments

<table>
<thead>
<tr>
<th>Rule</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r(x, y) \Rightarrow r'(x, y)$</td>
<td>R-subsumption</td>
</tr>
<tr>
<td>$r(x, y) \Leftrightarrow r'(x, y)$</td>
<td>R-equivalence</td>
</tr>
<tr>
<td>$\text{type}(x, C) \Rightarrow \text{type}(x, C')$</td>
<td>C-subsumption</td>
</tr>
<tr>
<td>$r_1(x, y), r_2(y, z) \Rightarrow r'(x, z)$</td>
<td>2-hops translation</td>
</tr>
<tr>
<td>$r(x, z) r(y, z) \Rightarrow r'(x, y)$</td>
<td>Triangle alignment</td>
</tr>
<tr>
<td>$r_1(x, y), r_2(x, V) \Rightarrow r'(x, y)$</td>
<td>Specific R-subsumption</td>
</tr>
<tr>
<td>$r(x, V) \Rightarrow r'(x, V')$</td>
<td>Attribute-Value translation</td>
</tr>
<tr>
<td>$r_1(x, V1), r_2(x, V2) \Rightarrow r'(x, V')$</td>
<td>2-values translation</td>
</tr>
</tbody>
</table>

Rule Mining for canonicalization of relations

Open KBs express relations in multiple ways

Harvard Law School → is a graduate of → Barack Obama

Columbia University → earned degree from → Barack Obama
Rule Mining for canonicalization of relations

Problem for query answering

Harvard Law School

is a graduate of

Barack Obama

earned degree from

Columbia University

earned degree from
Rule Mining for canonicalization of relations

Barack Obama is a graduate of?

Harvard Law School

Columbia University

Barack Obama
Rule Mining for canonicalization of relations

Use rule mining to find equivalent relations

Luis Galárraga, Geremy Heitz, Kevin Murphy, Fabian Suchanek. Canonicalizing Open Knowledge Bases. In CIKM, 2014
Research outlook

• Numerical correlations

\[ \text{export}(x, y), \text{import}(x, z) \Rightarrow \text{cad}(x, 1.2 \times (z - y)) \]

• Probabilistic model to learn confidence of predictions
  - Multiple rules can predict a fact
  - Integrate soft and hard constraints