Efficient Random Walk Inference with Knowledge Bases

Ni Lao
Carnegie Mellon University
2012-7-11

Committee: William W. Cohen (Chair),
Teruko Mitamura, Tom Mitchell,
C. Lee Giles (Pennsylvania State University)
Knowledge itself is power.

--Francis Bacon

KB as an edge-labeled graph

an algorithm which tries to achieve something,
e.g. IR/IE/QA/MT
Link Prediction
-- a generic relational learning task

Given

a directed edge-labeled graph
a relation type r
a source node s (also called a query)

Find

the set of nodes G, so that r(s,t) for each t in G
What is the profession of Charlotte Brontë?
Consider Friends/Family

Patrick Brontë

Charlotte Brontë

HasFather

Writer

Profession
Consider Behaviors

IsA⁻¹ is the inverse of IsA
Wrote⁻¹ is the inverse of Wrote
Consider Literatures/Publications

Charlotte Brontë

Writer

Painter

Mentioned

Mentioned⁻¹

Profession
these are interesting papers

a paper stream
a paper river
new development of an interesting topic

write

cite

a paper river
new papers of my favorite author

a paper river
social recommendation

scientist who have similar interests

write
read
read
read
read

a paper river
Relational learning is a subfield of artificial intelligence, that learns with expressive logical or relational representations.

Relational Learning Goals

- **expressive**: define features expressing sequences of relations on a graph
- **robust**: combine many such features when making decisions
- **scalable**: efficiently discover and calculate such features
Why is relational learning computationally challenging?

Exponentially many path types

Exponentially many path instantiations

Our solution: feature metrics

Our solution: sampling
**Thesis Outline**

### Algorithms

- **Ch. 2**: Path Ranking Algorithm (Lao & Cohen, MLJ 2010)
- **Ch. 3**: knowledge base inference (Lao+, EMNLP 2011)
- **Ch. 4**: literature recommendation (Lao & Cohen, DILS 2012)
- **Ch. 5**: efficient RW (Lao & Cohen, KDD 2010)
- **Ch. 6**: distributed computing
- **Ch. 7**: more expressive features (submitted)

### Applications

- **Ch. 2**: Path Ranking Algorithm (Lao & Cohen, MLJ 2010)
- **Ch. 3**: knowledge base inference (Lao+, EMNLP 2011)
- **Ch. 4**: literature recommendation (Lao & Cohen, DILS 2012)
- **Ch. 5**: efficient RW (Lao & Cohen, KDD 2010)
- **Ch. 6**: relation extraction from parsed text (Lao+, EMNLP 2012)
- **Ch. 7**: coordinate term extraction

### Future Work

- **Ch. 8**: future work
Outline

Motivation
- the problem
- previous work
- idea
- contribution

Algorithms
- Path Ranking Algorithm (PRA)
- efficient RW
- distributed computing
- more expressive features

Applications
- knowledge base inference
- literature recommendation
- relation extraction from parsed text
- coordinate term extraction
Inductive Logic Programming

e.g.
First Order Inductive Learner--FOIL (Quinlan, ECML’93)

High precision Horn clauses

HasFather(a,b) \&\& Profession(b,y) \rightarrow Profession(a,y)

expressive

not robust

not scalable

experimental comparison later
Undirected Graphical Models
-- combine logics with GM

e.g.
Markov Logic Networks (Kok & Domingos, ICML’05)
Relational CRFs (Lao+, NIPS’10)

Horn clauses
smokes(A) & Friends(A,B) → smokes(B)

as CRFs features
smokes(A) & Friends(A,B) & !smokes(B)

expressive
robust
not scalable
Random Walk with Restart
-- ignore logic

e.g. Tong+, ICDM’06

Charlotte Brontë
Patrick Brontë
HasFather
Writer
Profession
Mentioned

Jane Eyre
Wrote
Novel
A Tale of Two Cities
IsA
Charles Dickens
Wrote⁻¹
Profession

not expressive
robust
scalable

experimental comparison later
Outline

Motivation

- the problem
- previous work
- idea
- contribution

Algorithms

- Path Ranking Algorithm (PRA)
- efficient RW
- distributed computing
- more expressive features

Applications

- knowledge base inference
- literature recommendation
- relation extraction from parsed text
- coordinate term extraction
Relational Classification
-- combine logics with RWs

e.g. Path Ranking algorithm (Lao & Cohen, MLJ’10)

\[ P(\text{Charlotte} \rightarrow \text{Writer}; <\text{HasFather,IsA}>) \]
\[ P(\text{Charlotte} \rightarrow \text{Writer}; <\text{Mention,Mention}^{-1},\text{IsA}>) \]
\[ \ldots \]

\[ P(\text{Charlotte} \rightarrow \text{Painter}; <\text{HasFather,IsA}>) \]
\[ P(\text{Charlotte} \rightarrow \text{Painter}; <\text{Mention,Mention}^{-1},\text{IsA}>) \]
\[ \ldots \]
Contribution

Apply relational learning at scales not possible before

made possible by

a family of easy-to-learn features
fast random walk
distributed computing
Outline

**Motivation**
- the problem
- previous work
- idea
- contribution

**Algorithms**
- Path Ranking Algorithm (PRA)
- efficient RW
- distributed computing
- more expressive features

**Applications**
- knowledge base inference
- literature recommendation
- relation extraction from parsed text
- coordinate term extraction
Path Ranking Algorithm (PRA) (Lao & Cohen, MLJ 2010)

\[
score(s, t) = \sum_{\pi \in B} P(s \rightarrow t; \pi) \theta_{\pi}
\]

e.g. \( \pi = \langle \text{Mention}, \text{Mention}^{-1}, \text{IsA} \rangle \)

expressive

robust

a weight
Random Walk Calculation

$$\text{score}(s,t) = \sum_{\pi \in B} P(s \rightarrow t; \pi) \theta_{\pi}$$

Dynamic Programming

$$P(s \rightarrow t; \pi) = \sum_{z} P(s \rightarrow z; \pi') P(z \rightarrow t; r)$$

scalable

later about how to do it x100
more efficiently using sampling

e.g.
$$\pi' = \langle \text{Mention, Mention}^{-1} \rangle$$
$$r = \text{Profession}$$
Feature Selection with Labeled Data

\[
\text{score}(s,t) = \sum_{\pi \in B} P(s \rightarrow t; \pi) \theta_{\pi}
\]

given training query set \{ (s_i, G_i) \}

\[
\text{hits}(f) = \sum_i I \left[ \sum_{j \in G_i} f(s_i, t_j) \right] \geq h
\]

\[
\text{accuracy}(f) = \frac{1}{N} \sum_i \left[ \frac{\sum_{j \in G_i} f(s_i, t_j)}{\sum_j f(s_i, t_j)} \right] \geq a
\]

\( I() \): the indicator function

\( N \): total number of queries
Estimating $\theta$

$$score(s, t) = \sum_{\pi \in B} P(s \rightarrow t; \pi) [\theta_{\pi}]$$

for a relation $r$

- generate positive and negative node pairs $\{(s_i, t_i)\}$

for each $(s_i, t_i)$ generate $(x_i, y_i)$

- $x_i$ is a vector of RW features of different paths $\pi$
- $y_i$ is a binary label $r(s_i, t_i)$

estimate $\theta$ by L1/L2 regularized (elastic-net) logistic regression
Outline

Motivation
- the problem
- previous work
- idea
- contribution

Algorithms
- Path Ranking Algorithm (PRA)
- efficient RW
- distributed computing
- more expressive features

Applications
- knowledge base inference
- literature recommendation
- relation extraction from parsed text
- coordinate term extraction
## Knowledge Base Inference

*(Lao, Mitchell, Cohen, EMNLP 2010)*

### Example NELL relations

<table>
<thead>
<tr>
<th>TARGET RELATION</th>
<th>$N_Q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>athletePlaysForTeam</td>
<td>498</td>
</tr>
<tr>
<td>athletePlaysInLeague</td>
<td>892</td>
</tr>
<tr>
<td>athletePlaysSport</td>
<td>1119</td>
</tr>
<tr>
<td>stadiumLocatedInCity</td>
<td>254</td>
</tr>
<tr>
<td>teamHomeStadium</td>
<td>186</td>
</tr>
<tr>
<td>teamPlaysInCity</td>
<td>135</td>
</tr>
<tr>
<td>teamPlaysInLeague</td>
<td>341</td>
</tr>
<tr>
<td>teamPlaysSport</td>
<td>339</td>
</tr>
<tr>
<td>teamMember</td>
<td>142</td>
</tr>
<tr>
<td>companiesHeadquarteredIn</td>
<td>393</td>
</tr>
<tr>
<td>publicationJournalist</td>
<td>68</td>
</tr>
<tr>
<td>producedBy</td>
<td>134</td>
</tr>
<tr>
<td>competesWith</td>
<td>226</td>
</tr>
<tr>
<td>hasOfficeInCity</td>
<td>398</td>
</tr>
<tr>
<td>teamWonTrophy</td>
<td>149</td>
</tr>
<tr>
<td>worksFor</td>
<td>363</td>
</tr>
</tbody>
</table>

### NELL (Never Ending Language Learner) v165

- 353 relations
- 0.7M nodes (concepts)
- 1.7M edges
PRA Uses Broad Coverage Features

AthletePlaysSport(HinesWard, ?)
PRA Has Much Higher Recall and Is Much Faster

Mechanical Turk evaluate new beliefs of 8 functional relations

PRA trains in an hour vs. FOIL trains in a few days
Outline

Motivation
- the problem
- previous work
- idea
- contribution

Algorithms
- Path Ranking Algorithm (PRA)
- efficient RW
- distributed computing
- more expressive features

Applications
- knowledge base inference
- literature recommendation
- relation extraction from parsed text
- coordinate term extraction
Biology Literatures

Databases

Yeast: 0.8M nodes, 3.5M edges
Fly: 0.7M nodes, 16.9M edges

mesh descriptors/qualifiers 29k
chemical 14k
author 71k
journal 1.0k
institute 6k
year 64

title word 40k
gene 5.6k
Cites 0.3M

1.1M
0.3M
160k
0.5M
39k
Before

Relates to 1.6K

Application
Recommendation Tasks

Literature Recommendation

year, author → papers a user is going to read

training data --- 1 user over 20 years

(collected from Dr. John Woolford’s computer)
PRA Combines Dozens of Recommendation Strategies
Reading Recommendation

Mean Reciprocal Rank

Maximum Path Length (L)

- PRA
- RWR
- RWR(no training)
- Community-based
- Content-based
Motivation
- the problem
- previous work
- idea
- contribution

Algorithms
- Path Ranking Algorithm (PRA)
- efficient RW

Applications
- knowledge base inference
- literature recommendation
- relation extraction from parsed text
- coordinate term extraction
Efficient Random Walks

Exact calculation of random walks results in non-zero probabilities for many internal nodes

(Lao & Cohen, KDD 2010)

1 billion nodes

A few nodes that we care about

Charlotte

query node

Writer

Painter

Zebra

7/13/2012
Idea: a few random walkers (particles) are enough to distinguish good target nodes from bad ones.
Compare Speedup Approaches

Gene Recommendation on Fly Data (N=2k)

- Finger Printing
- Particle Filtering
- Fixed Truncation
- Beam Truncation

10x ~ 100x faster with little loss of quality

exact random walks

exact random walks
Outline

Motivation
- the problem
- previous work
- idea
- contribution

Algorithms
- Path Ranking Algorithm (PRA)
- efficient RW
- distributed computing

Applications
- knowledge base inference
- literature recommendation
- relation extraction from parsed text
- more expressive features
- coordinate term extraction
Relation Extraction

(21M concepts, 70M edges)

Freebase

Entity Resolution

Coreference Resolution

News Corpus (60M)

Application
Can PRA scale?

Can PRA learn syntactic-semantic rules?
Large number of queries

- e.g. 0.3M/2M persons have known profession

**Solution:** map/reduce to explore path, generate training samples, calculate gradient, and do predictions for each query

Large text graph

- e.g. 60M documents

**Solution:** each node keeps the Freebase graph in memory relevant sentences are loaded/unloaded for each query
Combine Syntax with Semantics

\[ <M, \text{conj}, M^{-1}, \text{Profession}> \]

Freebase

Ian McDougall

Simon Philips

Professor

M

M

conj

subj

collaborated

"McDougall and Simon Phillips collaborated..."
Combine Text with Semantics

<\text{M, WORD, CW}^{-1}, \text{profession}^{-1}, \text{profession}>
Text and KB Work Better Together

Tested by existing knowledge in Freebase

Mean Reciprocal Rank

with closed world assumption

Profession
Nationality
Parent

Freebase-only
Text-only
Freebase+Text

7/13/2012
Highly Accurate New Beliefs

manually evaluated new beliefs

Precision

<table>
<thead>
<tr>
<th></th>
<th>Profession</th>
<th>Nationality</th>
<th>Parents</th>
</tr>
</thead>
<tbody>
<tr>
<td>p@100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p@1k</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p@10k</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7/13/2012
Outline

Motivation
- the problem
- previous work
- idea
- contribution

Algorithms
- Path Ranking Algorithm (PRA)
- efficient RW
- distributed computing
- more expressive features

Applications
- knowledge base inference
- literature recommendation
- relation extraction from parsed text
- coordinate term extraction
Coordinate Term Extraction Task

(Minkov & Cohen, ECML 2010)

parsed MUC-6 corpus
153k nodes, 748K edges
30 queries
given 4 person names as seeds, find other persons

Tokens

Words/POSs

Tokens

W: word
POS: part of speech
nnp: singular proper nouns
vbd: verb, past tense
nsubj: subject of a verb
Good Paths Are Quite Long

<W⁻¹,nsubj,W,W⁻¹,nsubj⁻¹,W>

Tokens

Words

Tokens

find entities with similar behaviors

SteveJobs

founded

W

W

W

W

nsubj

nsubj

SteveJobs

founded

BillGates

founded

BillGates

founded

nsubj

nsubj
Good Paths Are Quite Long

![Graph showing the relationship between Max Path Length and Mean Average Precision. The x-axis represents Max Path Length (3 to 6), and the y-axis represents Mean Average Precision (0.00 to 0.20). The graph includes lines and markers indicating the trend.]
Forward Search Is Wasteful

Find paths that connect s and t
Bidirectional Search Is More Efficient

The challenge is to calculate $P(s \rightarrow t; \pi)$.
Forward vs. Backward RWs

Forward

\[ P(s \rightarrow t; \pi) = \sum P(s \rightarrow z; \pi')P(z \rightarrow t; r) \]

\[ \text{evaluate } P(s \rightarrow t; \pi) \text{ for many } t \]

Backward

\[ P(t \leftarrow s; \pi) = \sum P(t \leftarrow z; \pi')P(z \leftarrow s; r) \]

\[ \text{evaluate } P(s \rightarrow t; \pi) \text{ for many } s \]
Bidirectional Search with RW

\[ P(s \rightarrow t; \pi_1 \pi_2) = \sum_z P(s \rightarrow z; \pi_1)P(t \leftarrow z; \pi_2) \]
Bidirectional Search Is Much Faster

Path Finding Time (s) vs. Max Path Length

- 2F+1B
- 3F+1B
- 3F
- 4F
- 2F+2B
- 3F+2B
- 5F
- 4F+2B
- 3F+3B
- 4F+1B

Exceed 20Gb memory limit

1000x faster
Need for Lexicalized Paths

Task: find person entities

\[ P(\text{vbd} \rightarrow t \mid <\text{POS}^{-1}, \text{nsubj}^{-1}, W>) \]
Evaluate Lexicalized Paths

Given an example \((s_i, t_i)\)

calculate \(P(z \rightarrow t_i; \pi)\) for many \(z\)
Person Name Extraction

~1000 correct answers

Mean Average Precision

L=6

RWR(no train)
RWR
FOIL
PRA
PRA+c2
PRA+c3
**Motivation**
- the problem
- previous work
- idea
- contribution

**Algorithms**
- Path Ranking Algorithm (PRA)
- efficient RW
- distributed computing
- more expressive features

**Applications**
- knowledge base inference
- literature recommendation
- relation extraction from parsed text
- coordinate term extraction
Future Work

Apply knowledge to NLP/IE/IR/CV tasks

$$\arg \max P(\text{decision} | \text{context}, \text{KB})$$

Entity Resolution

Coreference?

Dependency Trees

Text
Conjunctions of Paths
rules can have tree structures
with source/constant/target nodes as leaves
Future Work

Conjunctions of Paths

**forward** PCRW with multiple walkers

![Diagram](image)
Future Work

Conjunctions of Paths

**backward** PCRW with multiple walkers
Contribution

Apply relational learning at scales not possible before.

Leads to new applications!

Made possible by

- a family of easy-to-learn features (3 types)
- fast random walk (sampling)
- distributed computing
other work I did at CMU

Relational CRFs (Lao+, NIPS’10)
Question answering (Lao+, NTCIR’08)
Utility based retrieval evaluation (Yang+, SIGIR’07)
Future Work

KB extension

new relation types, new concepts

Unsupervised

\[ \arg \max_{\Delta KB} P(\text{corpus} \mid KB + \Delta KB) \]

Supervised

\[ \arg \max_{\Delta KB} P(\text{decisions} \mid \text{contexts}, KB + \Delta KB) \]
Directed Graphical Models

e.g.
Probabilistic Relational Models (Getoor+, ICML’01)

model structure restricted to DAG
cannot express features corresponding to chains
Coverage of top k triples

<table>
<thead>
<tr>
<th>Profession Triples</th>
<th>Unique Persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>1k</td>
<td>970</td>
</tr>
<tr>
<td>10k</td>
<td>8,726</td>
</tr>
<tr>
<td>100k</td>
<td>79,885</td>
</tr>
</tbody>
</table>
Repeatedly Combine Forward and Backward RWs

\[ <W^{-1}, \text{conj_and}, W, W^{-1}, \text{conj_and}, W, W^{-1}, \text{conj_and}, W, W> \]

- Forward search
- +Backward Search
- +Backward Search
## Summary of PRA

<table>
<thead>
<tr>
<th>Stage</th>
<th>Computation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Path Discovery</strong></td>
<td>given { (s_i, G_i) }, find { f ; acc(f) \geq a, hits(f) \geq h }</td>
</tr>
<tr>
<td><strong>Generate Training Samples</strong></td>
<td>generate { (s_i, t_i) } and { (x_i, y_i) }</td>
</tr>
<tr>
<td><strong>Logistic Regression Training</strong></td>
<td>( \theta = \arg \max_{\theta} \left[ \sum_i l_i(\theta) - \lambda_1 | \theta |_1 - \lambda_2 | \theta |_2^2 \right] )</td>
</tr>
<tr>
<td><strong>Prediction</strong></td>
<td>apply model to nodes ( s ) in \textit{domain}(r)</td>
</tr>
</tbody>
</table>
Need for Lexicalized Paths

Task=AthletePlaysInLeague

\[ P(mlb \rightarrow t; \phi) \]  
Bias toward MLB

\[ P\left( BostonBraves \rightarrow t; \left\{ \text{AthletePlaysForTeam}^{-1}, \text{AthletePlaysInLeague} \right\} \right) \]  
A prior over the leagues participated by Boston Braves university athletes
Need for Lexicalized Paths

Task=CompetesWith

\[ P(\text{google} \rightarrow t; \phi) \]

Bias toward Google

\[ P(\text{Google} \rightarrow t; \langle \text{CompetesWith, CompetesWith} \rangle) \]

Companies around Google
Knowledge Base Inference

16 tasks

MRR vs L=3

- RWR (no train)
- RWR
- PRA
- PRA+c1
- PRA+c2