Processing and Analyzing Web-Scale Knowledge Graphs

Sunil K. Narang
Sr. Software Eng.
Shared Platform Group - Microsoft
Overview

• Introduction

• Knowledge Bases and Graph

• Sample Problem

• Summary
AI Tasks

- **Machine Perception:**
  - Detect and encode physical information. Speech, Image, Video, gesture, touch.

- **Machine Cognition:**
  - Make sense of encoded information. **Knowledge**, Reasoning/Inference, Memorization/Learning, Decision, Actions, Interaction/Conversation

Semantic Understanding of the data

Data boom
Knowledge Graphs
Knowledge Base (KB) Graph

• Captures world knowledge by storing properties of billions of entities, as well as relations among them
Entity: Nodes in the KB Graph

- Digital proxy to maintain stronger binding between real world and digital world.
- Represented as nodes in graph
- Enabler for transferable and connected knowledge
### Facts: Edges in KB Graph

<table>
<thead>
<tr>
<th>Subject</th>
<th>Property</th>
<th>Object</th>
<th>Contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.freebase.com/m/02mjmr">http://www.freebase.com/m/02mjmr</a></td>
<td>people/person/weight</td>
<td>&quot;80.0&quot;^^decimal</td>
<td>(Blob)</td>
</tr>
<tr>
<td><a href="http://www.freebase.com/m/02mjmr">http://www.freebase.com/m/02mjmr</a></td>
<td>people/person/siblings</td>
<td><a href="http://www.freebase.com/m/0n4sqkl">http://www.freebase.com/m/0n4sqkl</a></td>
<td>(Blob)</td>
</tr>
<tr>
<td><a href="http://www.freebase.com/m/02mjmr">http://www.freebase.com/m/02mjmr</a></td>
<td>people/person/sibling</td>
<td><a href="http://www.freebase.com/m/0n4sqkk">http://www.freebase.com/m/0n4sqkk</a></td>
<td>(Blob)</td>
</tr>
<tr>
<td><a href="http://www.freebase.com/m/02mjmr">http://www.freebase.com/m/02mjmr</a></td>
<td>people/person/sexual_orientation</td>
<td><a href="http://www.freebase.com/m/04tk8n9">http://www.freebase.com/m/04tk8n9</a></td>
<td>(Blob)</td>
</tr>
<tr>
<td><a href="http://www.freebase.com/m/02mjmr">http://www.freebase.com/m/02mjmr</a></td>
<td>people/person/profession</td>
<td><a href="http://www.freebase.com/m/0cbd2">http://www.freebase.com/m/0cbd2</a></td>
<td>(Blob)</td>
</tr>
<tr>
<td><a href="http://www.freebase.com/m/02mjmr">http://www.freebase.com/m/02mjmr</a></td>
<td>people/person/profession</td>
<td><a href="http://www.freebase.com/m/0fj9f">http://www.freebase.com/m/0fj9f</a></td>
<td>(Blob)</td>
</tr>
<tr>
<td><a href="http://www.freebase.com/m/02mjmr">http://www.freebase.com/m/02mjmr</a></td>
<td>people/person/place_of_birth</td>
<td><a href="http://www.freebase.com/m/02nhh0">http://www.freebase.com/m/02nhh0</a></td>
<td>(Blob)</td>
</tr>
<tr>
<td><a href="http://www.freebase.com/m/02mjmr">http://www.freebase.com/m/02mjmr</a></td>
<td>people/person/parent</td>
<td><a href="http://www.freebase.com/m/03qccxj">http://www.freebase.com/m/03qccxj</a></td>
<td>(Blob)</td>
</tr>
<tr>
<td><a href="http://www.freebase.com/m/02mjmr">http://www.freebase.com/m/02mjmr</a></td>
<td>people/person/networth</td>
<td><a href="http://www.freebase.com/m/0jwhxpt">http://www.freebase.com/m/0jwhxpt</a></td>
<td>(Blob)</td>
</tr>
<tr>
<td><a href="http://www.freebase.com/m/02mjmr">http://www.freebase.com/m/02mjmr</a></td>
<td>people/person/nationality</td>
<td>location.sovereign.United States</td>
<td>(Blob)</td>
</tr>
<tr>
<td><a href="http://www.freebase.com/m/02mjmr">http://www.freebase.com/m/02mjmr</a></td>
<td>people/person/member_of</td>
<td><a href="http://www.freebase.com/m/03w60w">http://www.freebase.com/m/03w60w</a>_</td>
<td>(Blob)</td>
</tr>
<tr>
<td><a href="http://www.freebase.com/m/02mjmr">http://www.freebase.com/m/02mjmr</a></td>
<td>people/person/marriage</td>
<td><a href="http://www.freebase.com/m/02nagvlv">http://www.freebase.com/m/02nagvlv</a></td>
<td>(Blob)</td>
</tr>
<tr>
<td><a href="http://www.freebase.com/m/02mjmr">http://www.freebase.com/m/02mjmr</a></td>
<td>people/person/height</td>
<td>&quot;1.85&quot;^^decimal</td>
<td>(Blob)</td>
</tr>
<tr>
<td><a href="http://www.freebase.com/m/02mjmr">http://www.freebase.com/m/02mjmr</a></td>
<td>people/person/gender</td>
<td>people.gender.male</td>
<td>(Blob)</td>
</tr>
<tr>
<td><a href="http://www.freebase.com/m/02mjmr">http://www.freebase.com/m/02mjmr</a></td>
<td>people/person/friends</td>
<td><a href="http://www.freebase.com/m/01l9p">http://www.freebase.com/m/01l9p</a></td>
<td>(Blob)</td>
</tr>
</tbody>
</table>

### Three types of Nodes:
1. Primary Entity,
2. Relationship Entities,
3. Literal Values
APPLICATIONS IN MICROSOFT PRODUCTS

Entity Pane Experiences

Smart Search

Contextual Experiences

Office Insights

Cortana Experiences

Knowledge Widget (Beta)
SEARCH SCENARIOS

Entity Pane

Facts

Related Entities
SEARCH SCENARIOS CONTD.

Query Formulation
(Page-0, SERP)

Monetization
(Ads in Apps)
RESEARCH PROBLEMS IN KB

**Growth:** Knowledge graphs need to grow to catch up with the world.
- Entity/Ontology matching: connect graphs (a.k.a. conflation)
- Knowledge Extraction: extract new entities and relations from web/text/feeds
- **Link prediction/inference:** add more link.

**Data Quality:** Knowledge graphs are not always correct!
- Conflation resolution: merge duplicate entities, split wrongly merged ones
- **Inconsistency detection:** remove false facts.
- Other issues: freshness, coverage etc.
SAMPLE RESEARCH PROBLEMS IN KB

**Consumption**: how to make it easier to access knowledge?
- Semantic understanding: interpret the meaning of queries with entity intent
- Question answering: compute answers using the knowledge graph
- Document-stamping: Identify dominant entities in documents

**Intelligence**: can AI emerge from knowledge graphs?
- Automatic reasoning and planning
- Generalization and abstraction
Challenges

**Mammoth graph:**
Billions of entities, 10-15 facts on average per entity.
Reading the graph once can take hours. Joins are very expensive and can only be used for few hops.

**Diversity of the graph:**
20+ segments 5000+ types 15000+ properties. Imbalance distribution makes learning hard.

**Data Skews:** Uneven fan-in (Example US is a country of birth of millions of entities) and fan-out (Example: academy award winners).
Challenges in processing KBs Contd.

**Resource Constraints:**
Limited computational resources are shared across teams. Complex multi-step experiments lead to low productivity.

**Economy of improvement:**
**Big Consumption skew:** a few hundred thousand entities account of 90% of the clicks.
Human judges can fix 10 defects per hour at 3 cent each.
The rule based models and editorial correction triumph over ML due to high precision and low cost.
Link Prediction

• Principle:
  • Compute probability $P(\text{fact} | \text{other KB facts})$

• Applications:
  • **Predict new facts:** $\text{Nationality}(Natasha Obama, ?)$
    • Mine Inference rules: $\text{ BornInCity}(a, b) \land \text{CityInCountry}(b, c) \Rightarrow \text{ Nationality}(a, c)$
    • Inconsistency detection: $\text{date_of_birth}(a,x) \land \text{date_of_death}(a,y) \Rightarrow x < y$

• Methods:
  • Rule Based Methods (e.g., random walk)
  • Knowledge graph embedding
Part I: Rule Based Learning

• Define one-hop facts as atoms \( r(s,o) = (s,r,o) \)
• Identify complex patterns \( P = r_1(s_1,o_1)^r_2(s_2,o_2)^...^r_N(s_N,o_N) \)
• Identify rules \( P_i \rightarrow P_j \)
Algo 1: Path Ranking Algorithm (PRA)

For each relation $r$ connecting two entities, find other ways to connect. The quality depends on precision and support of the paths.

For example, in the knowledge base:
- Barack Obama is married to Michelle Obama.
- Michelle Obama is the parent of Malia Obama.
- Malia Obama is associated with the Michigan Wolverines in the context of education/college and SchoolSports/team/School.

Random Walk Inference and Learning in A Large Scale Knowledge Base. Lao et al 2011 EMNLP
Algo 2: Inconsistency Detection

- Literals value consist of few syntactic types:
  - DateTime: DOB, DOD etc.
  - Numbers (Decimals): height, weight, size etc.
  - String: Name, address etc.
  - Text: description, quotes, snippets etc.
  - Enums: gender, EntityType, Etc
Algo 2: Inconsistency Detection

- Literals value consist of few syntactic types:
  - DateTime: DOB, DOD etc.
  - Numbers (Decimals): height, weight, size etc.
  - String: Name, address etc.
  - Text: description, quotes, snippets etc.
  - Enums: gender, EntityType, Etc
Inconsistency Rules

• Easy to Model:
  • Expected range of a literal
  • Expected order of a pair of literal (DOD > DOB) for same entity
  • Uniqueness of a literal for an entity
  • Correlation between multiple values of a literal
Knowledge Base Embedding

• Each entity in a KB is represented by an $R^d$ vector

• Predict whether $(e_1, r, e_2)$ is true by $f_r(v_{e_1}, v_{e_2})$

Recent work on KB embedding

• Tensor decomposition
  • RESCAL [Nickel+, ICML-11], TRESCAL [Chang+, EMNLP-14]

• Neural networks
  • SME [Bordes+, AISTATS-12], NTN [Socher+, NIPS-13], TransE [Bordes+, NIPS-13]
Tensor Decomposition: Knowledge Base Representation (1/2)

- Collection of subj-pred-obj triples – \((e_1, r, e_2)\)

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obama</td>
<td>BornIn</td>
<td>Hawaii</td>
</tr>
<tr>
<td>Bill Gates</td>
<td>Nationality</td>
<td>USA</td>
</tr>
<tr>
<td>Bill Clinton</td>
<td>SpouseOf</td>
<td>Hillary Clinton</td>
</tr>
<tr>
<td>Satya Nadella</td>
<td>WorkAt</td>
<td>Microsoft</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

\(n: \# \text{entities}, m: \# \text{relations}\)
A zero entry means either:
- Incorrect (false)
- Unknown

\( \mathbf{X} \)

\[ R_k : \text{BornIn} \]

\( \chi_k \)  

\( k \)-th slice
Tensor Decomposition Objective

- Objective: \( \frac{1}{2} \left( \sum_k \| x_k - A R_k A^T \|^2_F \right) + \frac{1}{2} \left( \| A \|^2_F + \sum_k \| R_k \|^2_F \right) \)

- Reconstruction Error
- Regularization

\( x_k \) \( A \) \( R_k \) \( A^T \) RESCAL [Nickel+, ICML-11]
Measure the Degree of a Relationship

\[ f_{\text{BornIn}}(\text{Obama}, \text{Hawaii}) = A_{\text{Obama}} \cdot \mathcal{R}_{\text{BornIn}} A_{\text{Hawaii}}^T \]
Typed Tensor Decomposition – TRESCAL
[Chang+ EMNLP-14]

• Relational domain knowledge
  • Type information and constraints
  • Only legitimate entities are included in the loss

• Benefits of leveraging type information
  • Faster model training time
  • Highly scalable to large KB
  • Higher prediction accuracy
Typed Tensor Decomposition Objective

- Reconstruction error: \( \frac{1}{2} \sum_k \| \mathbf{X}_k - \mathbf{A} \mathcal{R}_k \mathbf{A}^T \|_F^2 \)

Diagram:
- \( \mathbf{X}_k \) (people)
- \( \mathbf{A} \)
- \( \mathcal{R}_k \)
- \( \mathbf{A}^T \)
- Relation: born-in
Typed Tensor Decomposition Objective

- Reconstruction error: \( \frac{1}{2} \sum_{k} \| \mathcal{X}'_k - \mathbf{A}_{kl} \mathbf{R}_k \mathbf{A}_{kr}^T \|_F^2 \)
Training Procedure –
Alternating Least-Squares (ALS) Method

Fix $\mathcal{R}_k$, update $\mathbf{A}$

Fix $\mathbf{A}$, update $\mathcal{R}_k$
Training Procedure –
Alternating Least-Squares (ALS) Method

\[ A \leftarrow \left[ \sum_k \mathcal{X}_k' A_{kr} R_k^T + \mathcal{X}_k'^T A_{kl} R_k \right] \left[ \sum_k B_{kr} + C_{kl} + \lambda I \right]^{-1} \]

where \( B_{kr} = R_k A_{kr}^T A_{kr} R_k \), \( C_{kl} = R_k A_{kl}^T A_{kl} R_k \).

\[
\text{vec}(R_k)
\leftarrow \left( A_{kr}^T A_{kr} \otimes A_{kl}^T A_{kl} + \lambda I \right)^{-1} \times \text{vec}(A_{kl}^T \mathcal{X}_k' A_{kr})
\]
## Relation Operators

<table>
<thead>
<tr>
<th>Relation representation</th>
<th>Scoring Function</th>
<th># Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector (TransE)</td>
<td>$</td>
<td></td>
</tr>
<tr>
<td>(Bordes+ 2013)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matrix (Bilinear)</td>
<td>$a^T M_r b$</td>
<td>$O(n_r \times k^2)$</td>
</tr>
<tr>
<td>(Bordes+ 2012, Collobert &amp; Weston 2008)</td>
<td>$u^T f ( M_{r1} a + M_{r2} b)$</td>
<td></td>
</tr>
<tr>
<td>Tensor (NTN)</td>
<td>$u^T f (a^T T_r b + M_{r1} a + M_{r2} b)$</td>
<td>$O(n_r \times k^2 \times d)$</td>
</tr>
<tr>
<td>(Socher+ 2013)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagonal Matrix</td>
<td>$a^T diag(M_r) b$</td>
<td>$O(n_r \times k)$</td>
</tr>
<tr>
<td>(RelDot) (Yang+ 2015)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$n_r$: #predicates, $k$: #dimensions of entity vectors, $d$: #layers
Neural Knowledge Base Embedding

\[ a = g(Wx_a) \]
\[ b = g(Wx_b) \]
Mining Horn-clause Rules

• Can relation embedding capture relation composition?
  \[ \text{BornInCity}(a, b) \land \text{CityInCountry}(b, c) \Rightarrow \text{Nationality}(a, c) \]

• Embedding-based Horn-clause rule extraction
  • For each relation \( r \), find a chain of relations \( r_1 \cdots r_n \), such that:
    \[ \text{dist}(M_r, M_1 \circ M_2 \circ \cdots \circ M_n) < \theta \]
  • \( r_1(e_1, e_2) \land r_2(e_2, e_3) \cdots \land r_n(e_n, e_{n+1}) \rightarrow r(e_1, e_{n+1}) \)

• Advantages vs. Inductive Logic Programming
  • Search the relation space instead of instance space
Aggregated Precision of Top Length-2 Rules

- AMIE [Galárraga+, WWW-2013] is an association rule-mining approach for large-scale KBs.
- Data: FB15k-401
- Execution time:
  - AMIE: 9 min.
  - EmbedRule: 2 min.
SEMANTIC LINKING TASK

- Given a piece of user content (e.g., search query, document, email, tabular data, etc.), link the content or portions of the content to a semantic dimension.

- For example:
  - link “top gun” in the search query “top gun seafood close to Bellevue” to the restaurant entity of that name, and
  - link “4.7” in the query “Body Glove Apple iPhone 6 4.7 Satin Case” to the display_size property.
Solution: Knowledge and Text Jointly Embedding

Jointly embedding

- Entity / word / phrase
  → Continuous vector in the same space

- Relation → Operator

\[
\min_{\{e_i\},\{r_i\},\{w_i\}} L_k(KG) + L_t(\text{Text}) + L_{kt}(KG, \text{Text})
\]

Knowledge Model  Text Model  Alignment Model

- Preserve semantic relations/similarities
  - Relations in knowledge graph
  - Similarities (co-occurrences) in text

Jianwen Zhang et al., Knowledge Graph and Text Jointly Embedding, EMNLP 2014
Performance

Table 4: **Triplet classification**: accuracy (%) over various types of triplets.

<table>
<thead>
<tr>
<th>Type</th>
<th>$e - e$</th>
<th>$w - e$</th>
<th>$e - w$</th>
<th>$w - w$</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>respectively</td>
<td>93.4</td>
<td>52.1</td>
<td>51.4</td>
<td>71.0</td>
<td>77.5</td>
</tr>
<tr>
<td>jointly (anchor)</td>
<td>94.4</td>
<td>67.0</td>
<td>66.7</td>
<td>79.8</td>
<td>81.9</td>
</tr>
<tr>
<td>jointly (name)</td>
<td>94.5</td>
<td>80.5</td>
<td>80.0</td>
<td>89.0</td>
<td>87.7</td>
</tr>
<tr>
<td>jointly (anchor+name)</td>
<td>95.0</td>
<td>82.0</td>
<td>81.5</td>
<td>90.0</td>
<td>88.8</td>
</tr>
</tbody>
</table>

respectively: TransE + Skip Gram
Open Question

• Can the problem above be formulated as interpolation problem and solved using GSP?

• Can we obtain a better embedded space representation using GSP formulation?
Questions?