Towards Exploiting Implicit Human Feedback for Improving RDF2vec Embeddings

Ahmad Al Taweel, Heiko Paulheim
Motivation

- Learning about KG entities w/ existing ML tools
  - most ML tools expect vectors, not nodes
  - we need a vector representation of entities
  → wanted: a transformation from nodes to sets of features

A Brief Excursion to word2vec

- A vector space model for *words*
- Introduced in 2013

- Each word becomes a vector
  - similar words are close
  - relations are preserved
  - vector arithmetics are possible

https://www.adityathakker.com/introduction-to-word2vec-how-it-works/
A Brief Excursion to word2vec

- Assumption:
  - Similar words appear in similar *contexts*

  {Bush, Obama, Trump} was elected president of the United States
  United States president {Bush, Obama, Trump} announced…
  …

- Idea
  - Train a network that can predict a word from its context (CBOW)
    or the context from a word (Skip Gram)

Mikolov et al.: Efficient Estimation of Word Representations in Vector Space. 2013
A Brief Excursion to word2vec

- Skip Gram: train a neural network with one hidden layer
- Use output values at hidden layer as vector representation
- Observation:
  - *Bush, Obama, Trump* will activate similar context words
  - i.e., their output weights at the projection layer have to be similar

Mikolov et al.: Efficient Estimation of Word Representations in Vector Space. 2013
From word2vec to RDF2vec

- Word2vec operates on *sentences*, i.e., sequences of words
- Idea of RDF2vec
  - First extract “sentences” from a graph
  - Then train embedding using RDF2vec
- “Sentences” are extracted by performing random graph walks:
  
  Year Zero — artist — Nine Inch Nails — member — Trent Reznor

- Experiments
  - RDF2vec can be trained on large KGs (DBpedia, Wikidata)
  - 300-500 dimensional vectors outperform other propositionalization strategies

Ristoski & Paulheim: RDF2vec: RDF Graph Embeddings for Data Mining. ISWC, 2016
From word2vec to RDF2vec

- RDF2vec example
  - similar instances form clusters
  - direction of relations is stable

Ristoski & Paulheim: RDF2vec: RDF Graph Embeddings for Data Mining. ISWC, 2016
Biased Graph Walks

• Maybe *random* walks are not such a good idea
  – They may give too much weight on less-known entities and facts
    • Strategies:
      – Prefer edges with more frequent predicates
      – Prefer nodes with higher indegree
      – Prefer nodes with higher PageRank
      – …
    – They may cover less-known entities and facts too little
    • Strategies:
      – The opposite of all of the above strategies

• Bottom line of experimental evaluation:
  – Not one strategy fits all

Cochez et al.: Biased Graph Walks for RDF Graph Embeddings. WIMS, 2017
A New Signal for Bias

• Existing biased graph walk strategies
  – use *internal* knowledge
    • e.g., property frequencies, PageRank, ...
    – i.e., signals only from the knowledge graph

• Simulating *human* walks instead of random walks
  – *biased* walk transition probabilities similar to *human* walk probabilities

• Problem
  – we don’t know how a human navigates through a knowledge graph
A New Signal for Bias

• Problem
  – we don’t know how a human navigates through a knowledge graph
• But
  – we know how humans navigate through Wikipedia

A New Signal for Bias

• Use transition probabilities from Wikipedia clickstream data
  – assuming each DBpedia entity corresponds to a Wikipedia page

• Discard non-Wikipedia pages and non-DBpedia-entities
  – incoming from other Web pages, e.g., Google search
  – outgoing to other Web pages, i.e., clicking on external links
  – outgoing to Wikipedia pages which are non-DBpedia entities, e.g., discussion pages
(Simplified) Example

- **Bad Witch**
  - artist (3,998)

- **Pretty Hate Machine**
  - artist (4,529)

- **The Downward Spiral**
  - artist (4,491)

- **The Fragile**
  - artist (2,494)

- **Nine Inch Nails**
  - band Member (8,989)
  - genre (979)
  - probability 62.4%
  - probability 6.8%

- **Atticus Ross**
  - band Member (4,439)
  - probability 30.8%

- **Trent Reznor**
  - band Member (4,529)
Evaluations

• We compare
  – Classic RDF2vec embeddings
  – Best performing RDF2vec embeddings w/ internal bias, i.e.
    • predicate frequency
    • PageRank
    • Inverse PageRank

• Evaluation setup
  – Classification tasks
  – Regression tasks
  – Content-based movie recommenders

  according to WIMS 2017 paper

  check out resource track paper:
  GEval: a Modular and Extensible Evaluation Framework for Graph Embedding Techniques
Evaluation Results

- Classification tasks (accuracy)
  - best result on 1/3 tasks
  - marginally worse than plain RDF2vec on 2/3 tasks
  - improvements over strategies with internal bias

<table>
<thead>
<tr>
<th>Strategy/Dataset</th>
<th>Cities</th>
<th>Metacritic Movies</th>
<th>Metacritic Albums</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NB</td>
<td>KNN</td>
<td>SVM</td>
</tr>
<tr>
<td>Uniform SG 200w 200v 4d</td>
<td>73.25</td>
<td>72.90</td>
<td>76.32</td>
</tr>
<tr>
<td>Uniform SG 500w 200v 4d</td>
<td>59.25</td>
<td>67.51</td>
<td>73.01</td>
</tr>
<tr>
<td>Uniform SG 500w 200v 8d</td>
<td>71.65</td>
<td>75.52</td>
<td>72.82</td>
</tr>
<tr>
<td>Uniform SG 500w 500v 8d</td>
<td>87.63</td>
<td>70.15</td>
<td>82.70</td>
</tr>
<tr>
<td>Predicate frequency weight SG 200w 200v 4d</td>
<td>72.15</td>
<td>70.77</td>
<td>73.37</td>
</tr>
<tr>
<td>Page-Rank weight SG 200w 200v 4d</td>
<td>74.16</td>
<td>72.73</td>
<td>67.21</td>
</tr>
<tr>
<td>Inverse Page-Rank Weight SG 200w 200v 4d</td>
<td>73.98</td>
<td>69.37</td>
<td>74.60</td>
</tr>
<tr>
<td>Click-Stream weight CBOW 200w 200v 4d</td>
<td>60.25</td>
<td>64.90</td>
<td>74.32</td>
</tr>
<tr>
<td>Click-Stream weight SG 200w 200v 4d</td>
<td>74.68</td>
<td>71.20</td>
<td>77.94</td>
</tr>
<tr>
<td>Click-Stream weight SG 500w 200v 4d</td>
<td>61.22</td>
<td>73.90</td>
<td>78.39</td>
</tr>
<tr>
<td>Click-Stream weight SG 500w 500v 8d</td>
<td>88.63</td>
<td>69.70</td>
<td>81.62</td>
</tr>
</tbody>
</table>
Evaluation Results

• Regression tasks (RMSE)
  – both plain RDF2vec and internal bias variants are outperformed
Evaluation Results

• Results on recommender task
  – item-based knn
  – both plain RDF2vec and internal bias variants are outperformed

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform SG 200w 200v 4d</td>
<td>0.05128</td>
<td>0.02466</td>
<td>0.03330</td>
</tr>
<tr>
<td>Uniform SG 500w 200v 4d</td>
<td>0.04852</td>
<td>0.03024</td>
<td>0.03725</td>
</tr>
<tr>
<td>Uniform SG 500w 200v 8d</td>
<td>0.04279</td>
<td>0.02612</td>
<td>0.03243</td>
</tr>
<tr>
<td>Uniform SG 500w 500v 8d</td>
<td>0.02692</td>
<td>0.01624</td>
<td>0.02025</td>
</tr>
<tr>
<td>Predicate frequency weight SG 200w 200v 4d</td>
<td>0.01946</td>
<td>0.0960</td>
<td>0.03236</td>
</tr>
<tr>
<td>Page-Rank weight SG 200w 200v 4d</td>
<td>0.03251</td>
<td>0.01828</td>
<td>0.02340</td>
</tr>
<tr>
<td>Inverse Page-Rank weight SG 200w 200v 4d</td>
<td>0.03924</td>
<td>0.02369</td>
<td>0.02954</td>
</tr>
<tr>
<td>Click-Stream weight CBOW 200w 200v 4d</td>
<td>0.03162</td>
<td>0.01348</td>
<td>0.01890</td>
</tr>
<tr>
<td>Click-Stream weight SG 200w 200v 4d</td>
<td>0.05261</td>
<td>0.03625</td>
<td>0.04292</td>
</tr>
<tr>
<td>Click-Stream weight SG 500w 200v 4d</td>
<td>0.04622</td>
<td>0.02573</td>
<td>0.03305</td>
</tr>
<tr>
<td>Click-Stream weight SG 500w 500v 8d</td>
<td>0.02489</td>
<td>0.01925</td>
<td>0.02170</td>
</tr>
</tbody>
</table>
Discussion and Outlook

• Wikipedia clickstream data adds a valuable signal
  – most results improve
  – (albeit by a small margin)

• Future work
  – obtain similar signals for DBpedia and other graphs
    • straight forward: Wikipedia-based graphs (YAGO, CaLiGraph)
    • less obvious: Wikidata, NELL, …
  – incorporate human signal in other embedding methods
    • e.g., weighted sums in TransE etc.

\[ \mathcal{L} = \sum_{(h, t, t') \in S} \sum_{(h', \ell, t') \in S'} \left[ \gamma + d(h + \ell, t) - d(h' + \ell, t') \right]_+ \]

• rdf2vec.org collects further implementations, variants, applications of RDF2vec
Towards Exploiting Implicit Human Feedback for Improving RDF2vec Embeddings

Ahmad Al Taweel, Heiko Paulheim