UNIVERSITÄT MANNHEIM

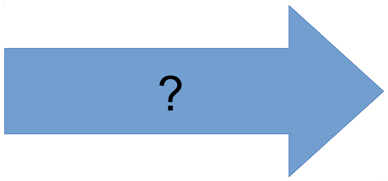
Towards Exploiting Implicit Human Feedback for Improving RDF2vec Embeddings

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



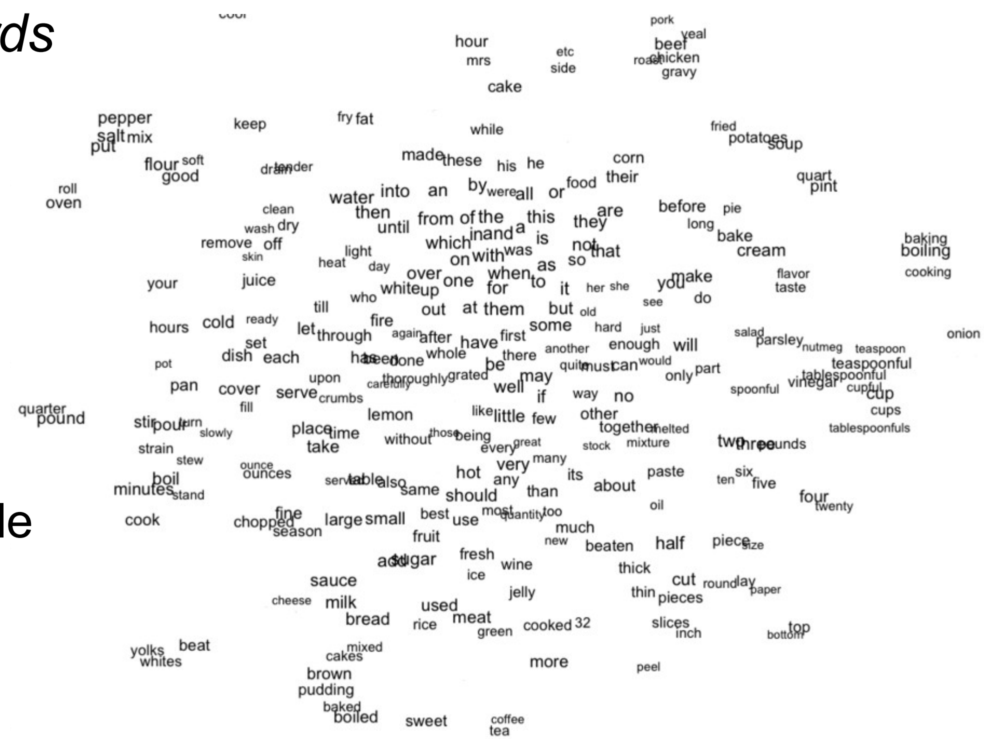
Row No.	attribute_1	attribute_2	attribute_3	attribute_4	attribute_5	attribute_6	attribute_7	attribute_8	attribute_9	attribute_10	attribute_11
1	0.020	0.037	0.043	0.021	0.095	0.099	0.154	0.160	0.311	0.211	0.161
2	0.045	0.052	0.084	0.069	0.118	0.258	0.216	0.348	0.334	0.287	0.492
3	0.026	0.058	0.110	0.108	0.097	0.228	0.243	0.377	0.560	0.619	0.633
4	0.010	0.017	0.062	0.021	0.021	0.037	0.110	0.128	0.060	0.126	0.088
5	0.076	0.067	0.048	0.039	0.059	0.065	0.121	0.247	0.356	0.446	0.415
6	0.029	0.045	0.028	0.017	0.038	0.099	0.120	0.183	0.210	0.304	0.299
7	0.032	0.096	0.132	0.141	0.167	0.171	0.073	0.140	0.208	0.351	0.179
8	0.052	0.055	0.084	0.032	0.116	0.092	0.103	0.061	0.146	0.284	0.280
9	0.022	0.037	0.048	0.048	0.065	0.059	0.075	0.010	0.068	0.149	0.116
10	0.016	0.017	0.035	0.007	0.019	0.067	0.106	0.070	0.096	0.025	0.080
11	0.004	0.006	0.015	0.034	0.031	0.028	0.040	0.027	0.032	0.045	0.049
12	0.012	0.031	0.017	0.031	0.036	0.010	0.018	0.058	0.112	0.084	0.055
13	0.008	0.009	0.005	0.025	0.034	0.055	0.053	0.096	0.101	0.124	0.110
14	0.009	0.006	0.025	0.049	0.120	0.159	0.139	0.099	0.096	0.190	0.190
15	0.012	0.043	0.060	0.045	0.060	0.035	0.053	0.034	0.105	0.212	0.164
16	0.030	0.061	0.065	0.092	0.162	0.229	0.218	0.203	0.146	0.085	0.248
17	0.035	0.012	0.019	0.047	0.074	0.118	0.168	0.154	0.147	0.291	0.233
18	0.019	0.061	0.038	0.077	0.139	0.081	0.057	0.022	0.104	0.119	0.124
19	0.027	0.009	0.015	0.028	0.041	0.076	0.103	0.114	0.079	0.152	0.168
20	0.013	0.015	0.064	0.173	0.257	0.256	0.295	0.411	0.498	0.592	0.583
21	0.047	0.051	0.082	0.125	0.178	0.307	0.301	0.236	0.383	0.376	0.302
22	0.066	0.058	0.084	0.037	0.046	0.077	0.077	0.113	0.235	0.184	0.287
23	0.010	0.048	0.030	0.030	0.065	0.108	0.236	0.238	0.007	0.188	0.146
24	0.011	0.015	0.014	0.008	0.021	0.106	0.102	0.044	0.093	0.073	0.074

Ristoski & Paulheim: A Comparison of Propositionalization Strategies for Creating Features from Linked Open Data. LD4KD, 2014

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…

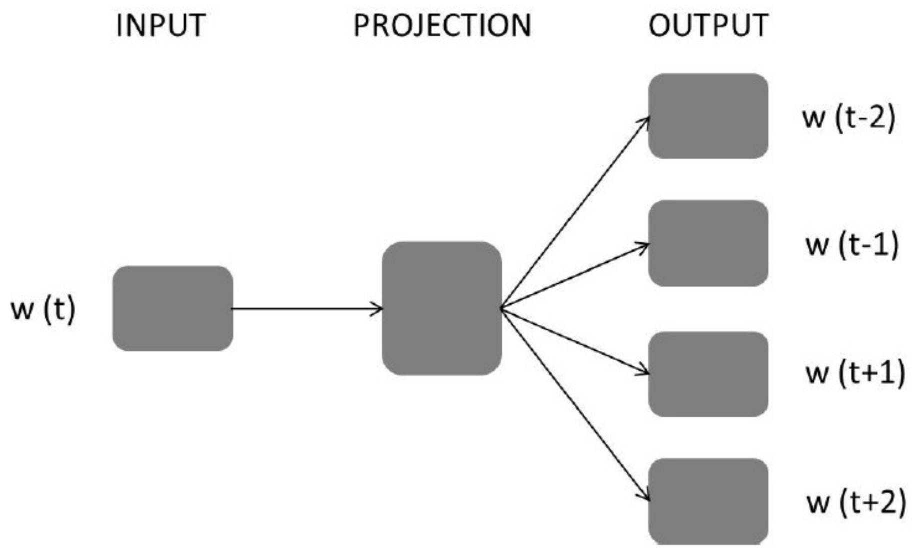
…

- 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:

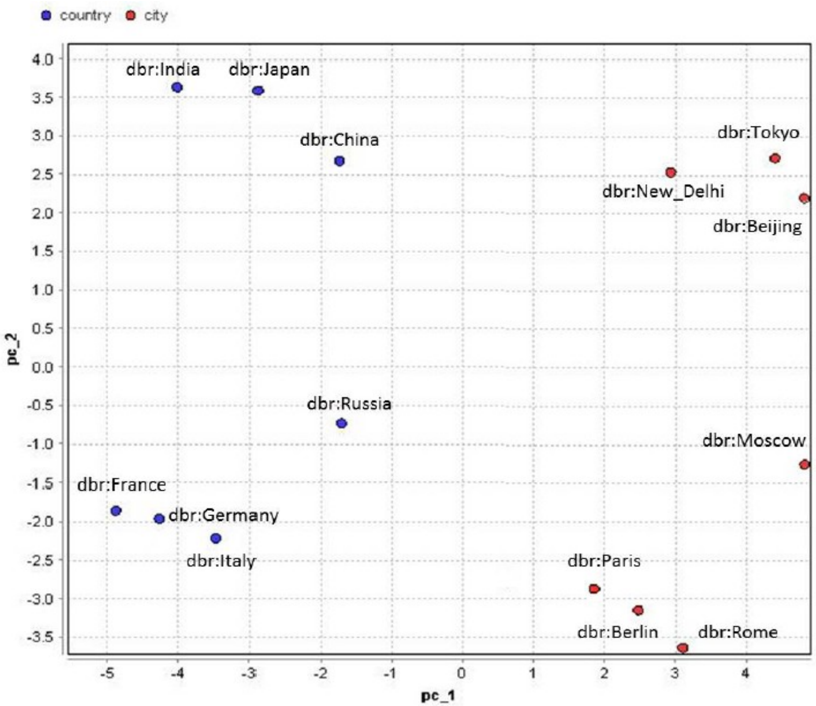
Year Zero $\xrightarrow{\text{artist}}$ Nine Inch Nails $\xrightarrow{\text{member}}$ Trent Reznor

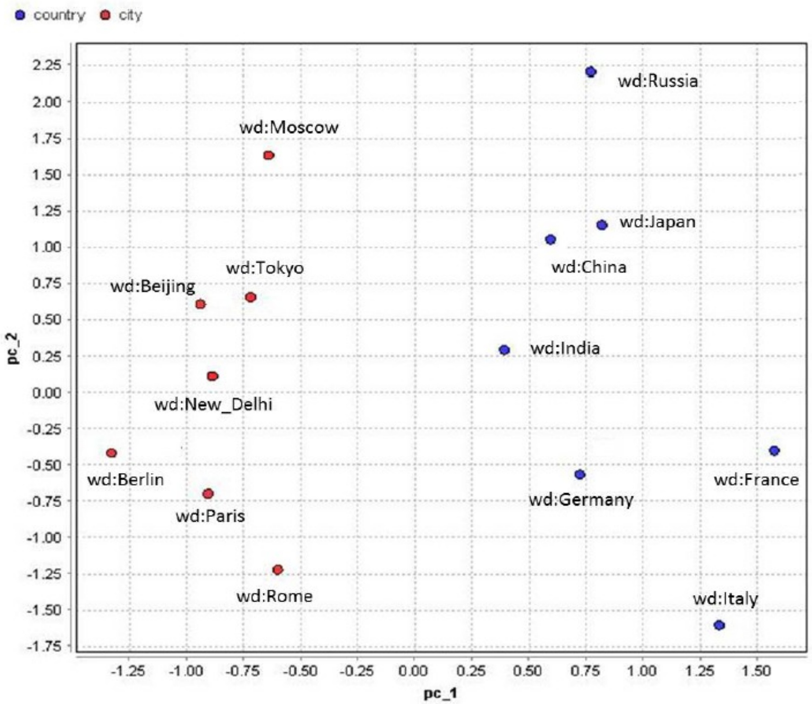
- Experiments
  - RDF2vec can be trained on large KGs (DBpedia, Wikidata)
  - 300-500 dimensional vectors outperform other propositionalization strategies

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

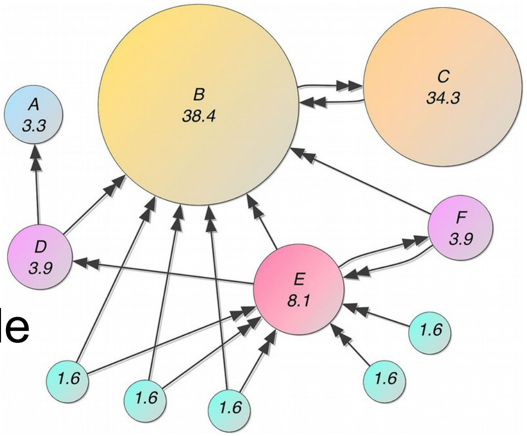




Ristoski & Paulheim: RDF2vec: RDF Graph Embeddings for Data Mining. ISWC, 2016

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

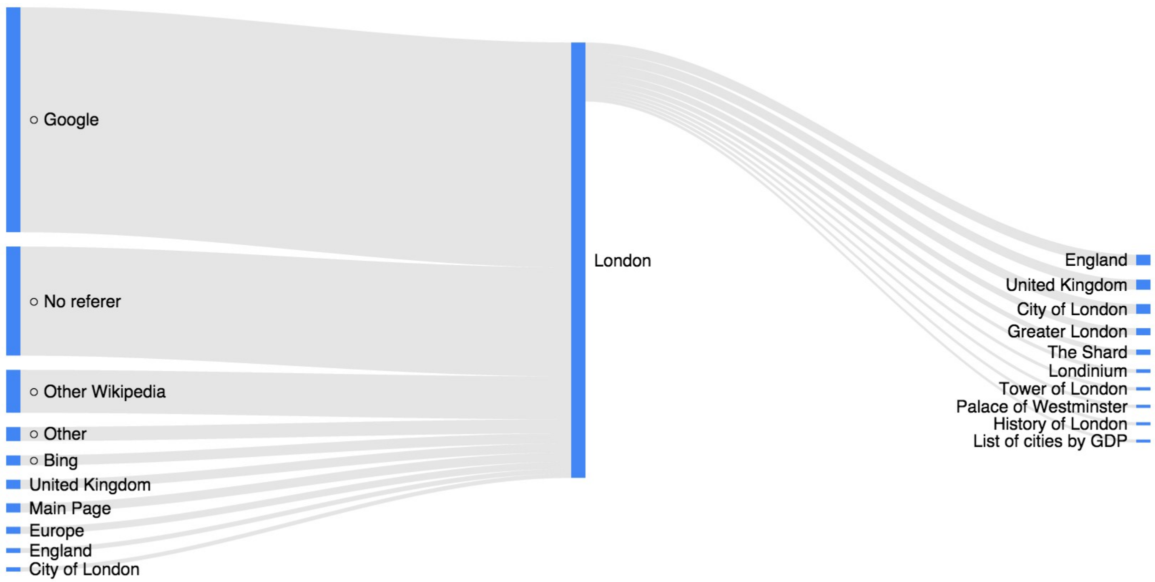
A New Signal for Bias

- 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

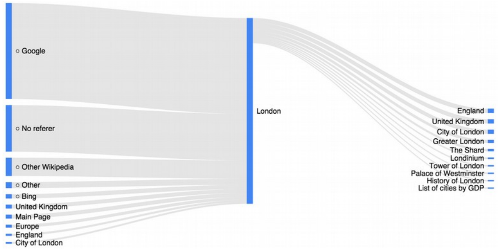
- Problem
  - we don't know how a human navigates through a knowledge graph
- But
  - we know how humans navigate through Wikipedia



https://meta.wikimedia.org/wiki/Research:Wikipedia_clickstream

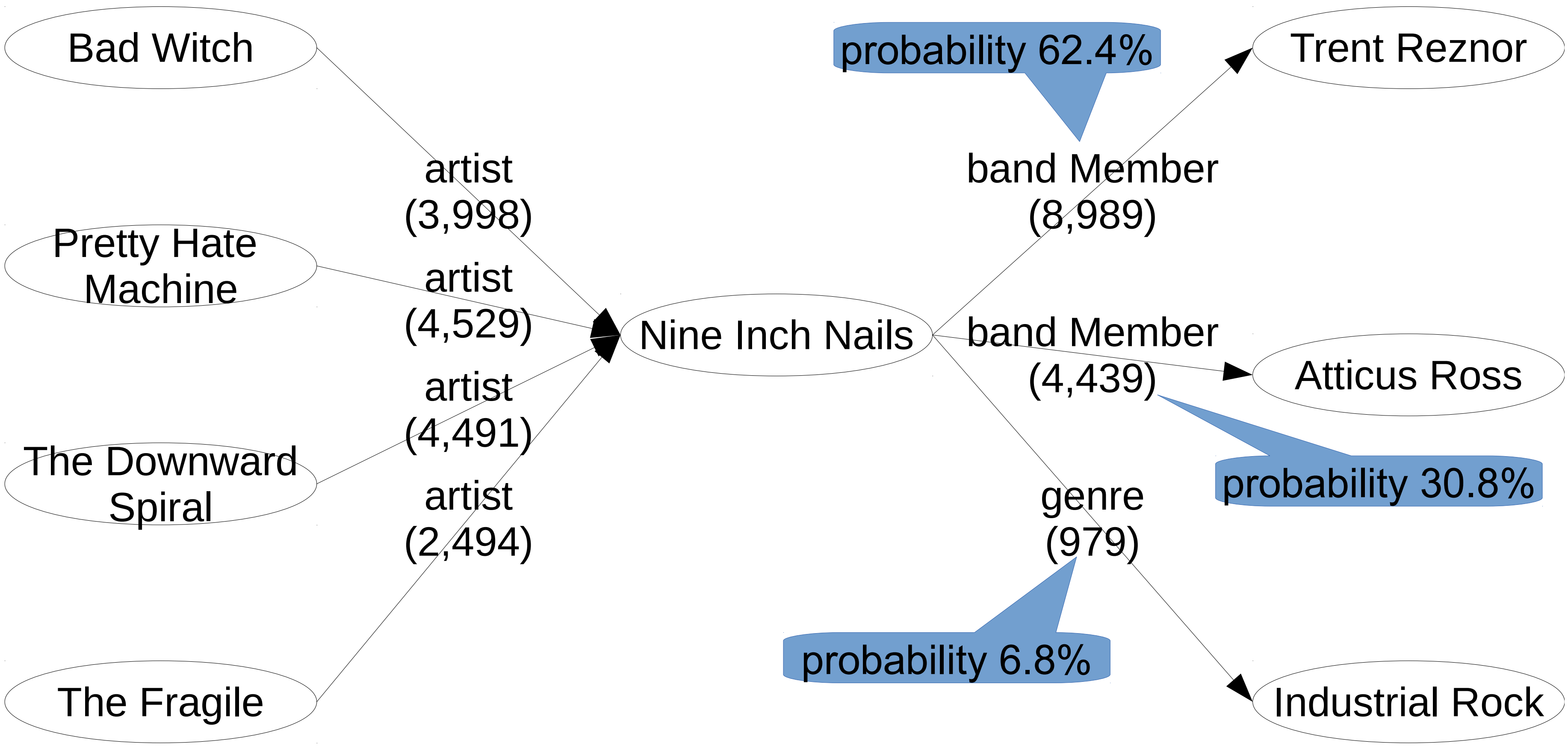
A New Signal for Bias

- 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



Evaluations

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

according to
WIMS 2017 paper

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

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

Strategy/Dataset	Cities				Metacritic Movies				Metacritic Albums			
	NB	KNN	SVM	C4.5	NB	KNN	SVM	C4.5	NB	KNN	SVM	C4.5
Uniform SG 200w 200v 4d	73.25	72.90	76.32	50.26	72.29	75.81	75.94	67.28	71.46	72.87	68.66	63.45
Uniform SG 500w 200v 4d	59.25	67.51	73.01	61.27	65.25	79.62	81.19	74.52	69.25	72.85	75.80	64.92
Uniform SG 500w 200v 8d	71.65	75.52	72.82	59.48	76.41	72.35	**83.51**	76.83	76.36	75.96	**81.72**	67.25
Uniform SG 500w 500v 8d	87.63	70.15	82.70	68.72	81.35	76.24	83.35	71.52	75.64	75.92	78.63	67.18
Predicate frequency weight SG 200w 200v 4d	72.15	70.77	73.37	51.25	74.60	75.21	76.13	72.42	69.82	71.90	73.53	62.22
Page-Rank weight SG 200w 200v 4d	74.16	72.73	67.21	62.83	71.32	69.19	79.70	74.43	69.61	71.98	74.47	68.11
Inverse Page-Rank Weight SG 200w 200v 4d	73.98	69.37	74.60	57.38	68.68	71.25	72.71	64.93	67.39	73.57	64.27	58.66
Click-Stream weight CBOW 200w 200v 4d	60.25	64.90	74.32	65.26	67.29	75.81	77.94	69.20	71.15	72.38	74.47	63.48
Click-Stream weight SG 200w 200v 4d	74.68	71.20	77.94	53.45	71.11	77.31	79.52	67.78	70.46	72.20	67.18	62.88
Click-Stream weight SG 500w 200v 4d	61.22	73.90	78.39	62.45	65.14	80.49	81.44	79.65	71.19	69.78	74.15	69.37
Click-Stream weight SG 500w 500v 8d	**88.63**	69.70	81.62	71.93	82.44	74.69	82.27	69.12	77.63	72.84	79.48	66.81

Evaluation Results

- Regression tasks (RMSE)
  - both plain RDF2vec and internal bias variants are outperformed

Strategy/Dataset	Cities			Metacritic Movies			Metacritic Albums		
	LR	KNN	M5	LR	KNN	M5	LR	KNN	M5
Uniform SG 200w 200v 4d	16.64	15.92	16.87	17.16	19.33	17.79	12.38	14.42	12.97
Uniform SG 500w 200v 4d	14.74	12.78	14.52	16.95	18.62	17.28	13.56	14.33	13.65
Uniform SG 500w 200v 8d	13.68	14.93	13.45	16.79	18.40	16.97	13.17	14.36	13.22
Uniform SG 500w 500v 8d	12.23	13.81	10.86	16.42	18.03	16.68	12.48	13.63	11.96
Predicate frequency weight SG 200w 200v 4d	16.54	17.83	17.56	18.28	20.90	19.72	14.31	16.88	13.44
Page-Rank weight SG 200w 200v 4d	14.74	14.57	16.14	17.63	20.81	16.86	12.57	15.72	12.56
Inverse Page-Rank weight SG 200w 200v 4d	14.87	16.59	14.93	16.10	18.44	16.16	11.56	12.93	11.51
Click-Stream weight CBOW 200w 200v 4d	15.13	13.64	16.13	19.76	20.91	19.48	13.52	14.29	13.58
Click-Stream weight SG 200w 200v 4d	15.48	16.16	15.24	17.31	19.27	16.78	12.28	13.75	12.21
Click-Stream weight SG 500w 200v 4d	13.96	16.53	14.82	**15.66**	16.15	15.86	11.70	12.50	11.77
Click-Stream weight SG 500w 500v 8d	12.25	12.57	**10.11**	15.72	16.89	15.80	10.79	11.67	**11.14**

Evaluation Results

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

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

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.

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

Bordes et al.: Translating Embeddings for Modeling Multi-relational Data. NIPS, 2013

Advertisement

- rdf2vec.org collects further implementations, variants, applications of RDF2vec

UNIVERSITÄT MANNHEIM

Towards Exploiting Implicit Human Feedback for Improving RDF2vec Embeddings

Ahmad Al Taweel, Heiko Paulheim