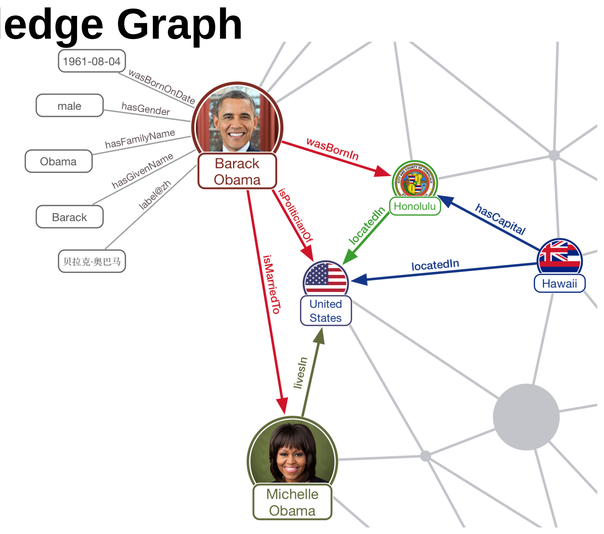
Link Prediction of an Academic Knowledge Graph

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

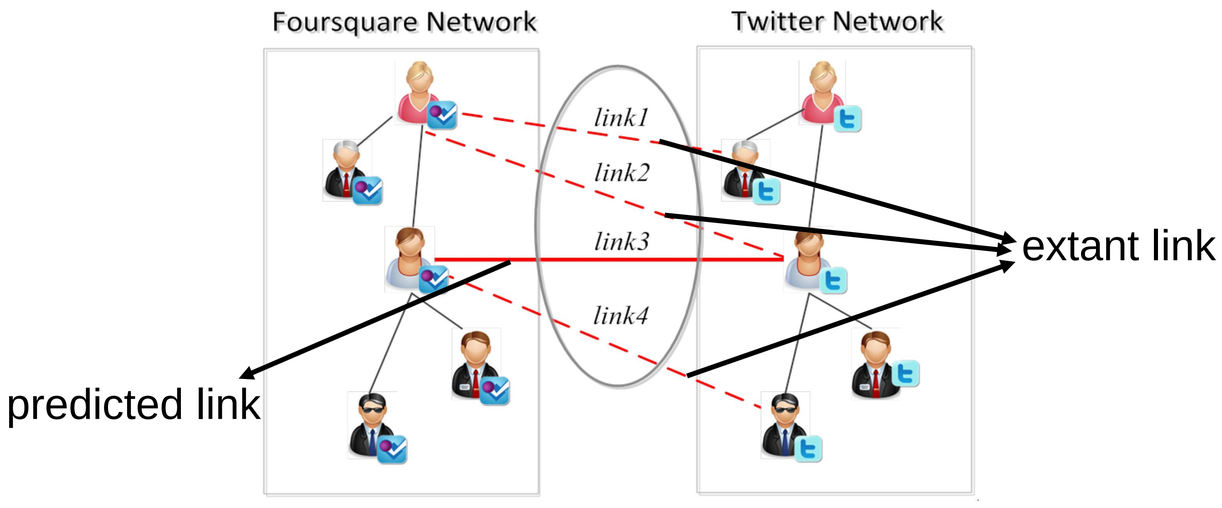
Mengdie Zhou, Shanghai Jiao Tong University

Background

Knowledge Graph



Link Prediction



Terminologies

Head - an anonymous id unique to an entity, which is the head of a triplet.
Relation - a string unique to a relation, which is the relation of a triplet.
Tail - an anonymous id unique to an entity, which is the tail of a triplet.

Task

Raw Data

Head	Relation	Tail
AAFEBBA8	paper_is_i	6506F4E8
E88BED0D	author_is_	DC6494B5
C63D435B	paper_is_i	191A225A
01658F27	author_is_	7.70E+103
EA8E0DE4	author_is_	2A37E79A
730564D0	paper_is_i	AA5F22A3
75C64E93	author_is_	5AA03FF0
66BA130D	author_is_	8B92C65D
9FCCF65B	author_is_	81D85941
13468B1E	author_is_	935C9037
CE35EAF0	author_is_	419B27FF

Link prediction

Predict Possible Tail

QueryId	Head	Relation	ExpectedTail
1	1ADE9CA3	work_in	1B57B3EA 1760D569 11869FC8
2	8B8C45F4	work_in	4D0E689A 1AC09AFE 6D3EE95D
3	42977719	work_in	ABD7DD86 FF99AEE5 A2281DDA
4	1DB623B8	work_in	045A90E8 C662C902 AC68FE02
5	E7CA4516	work_in	0349257F 6DB82332 83898A5D
6	32EF15AE	work_in	D8CAC16A 6E449A19 990C5E45
7	14EF175A	work_in	A53FE20E 63B11C49 362E311A
8	D483BCC9	work_in	F17B8472 28122D91 2404DBE9
9	4DDD16C	work_in	0FCD3B42 C70682BA A386ED37
10	784EDD65	work_in	7C2D2639 54464FB4 95CCEED5

Predict top 3 possible tails, given head and corresponding relation

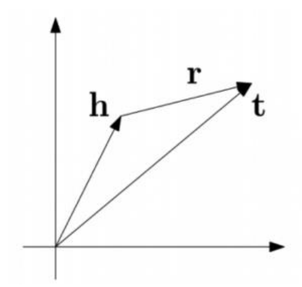
Challenges

Text Prediction: not a traditional vector that used in classifiers.
Unique Triplet: all 3 attributes of (head,relation,tail) should be considered when predicting.
Rank of Prediction: after predicting expected tails, the rank of tails also matter in this case.

Our Approach

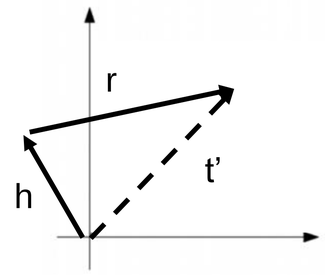
Translating Embedding(TransE)

Vectorize Triplet



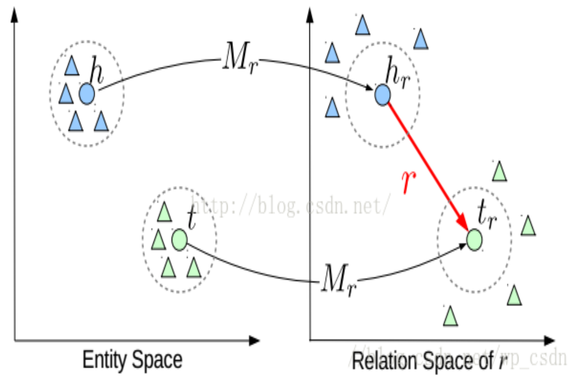
Predict

Using vector h,r to predict new tail t'



Learning Entity and Relation Embeddings(TransR)

Vector Space: The TransR regard that vector of relation lies on different space of that of entity.
Mapping Metrix: Metrix M_r transfers the vector in entity space to relation space.



Voter

Mechanism: The performance of different classifiers differs with regard to their prerequisites of data**.** When combined together using voter, the result can be compensated.

$$scores(e_i) = \sum_j r_j(e_i)$$

- e_i is expected entity
- r_j is a function that calculate the score of e_i in classifier r_j
- $$r_j(e_i) = \begin{cases} 2 - i & i\ is\ the\ rank\ of\ e_i\ in\ r_j \\ 0 & where\ e_i\ \ is\ not\ in r_j \end{cases}$$

Experiments

Setup

Dataset: given by Kaggle, concludes:train.csv, test.csv
Evaluation: Submissions are evaluated according to the Mean Average Precision @ 3 (MAP@3):

$$MAP@3 = \frac{1}{|U|} \sum_{u=1}^{|U|} \sum_{k=1}^{min(3,n)} P(k)$$

Compared Models

- N: the dimension of vector space.
- Margin: the distance between true triplets and corrupted triplets.

Results

Approach	Public accuracy scores			Private accuracy socres		
Margin	m=1	m=2	m=3	m=1	m=2	m=3
TransE (n=150)	0.23304	0.25853	0.28487	0.23364	0.24179	0.28208
TransE (n=200)	0.25107	0.28984	0.23909	0.25709	0.28935	0.24661
TransE (n=215)	0.24587	0.28687	0.28815	0.25298	0.28883	0.28829
TransE (n=225)	0.25684	0.28724	0.28333	0.26201	0.28802	0.29247
TransR(n=200)	0.22035	0.23453		0.23112	0.23298	
number of mixer	num=5	num=6	num=7	num=5	num=6	num=7
vote	0.40308	0.41179	0.41257	0.41431	0.41431	0.41298