به نام خدا

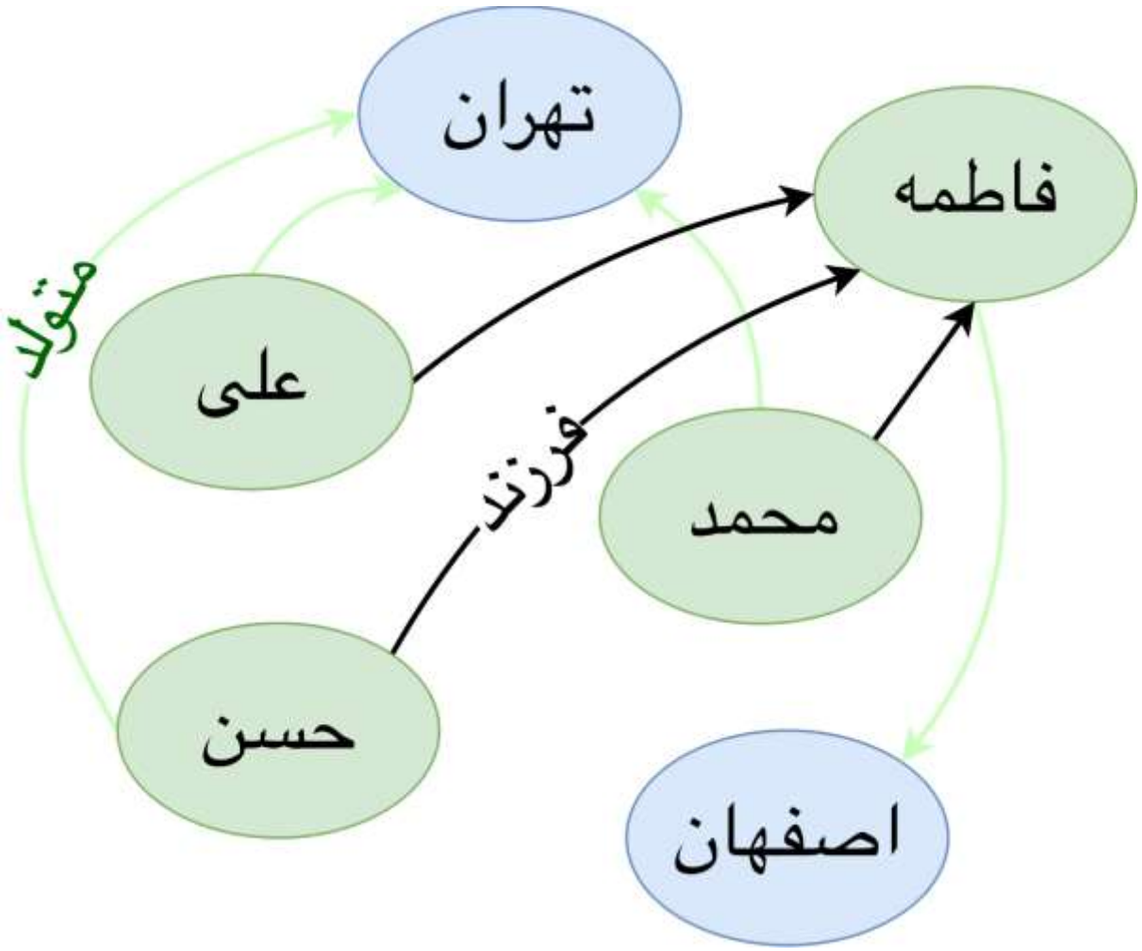
بهبود مدل‌های پیش‌بینی لینک مبتنی بر ویژگی‌های پنهان با استفاده از قوانین انجمنی

Improvement in Link Prediction Model based on Latent Factor using Association Rules

مسعود هاشمیان
استاد راهنما :دکتر یزدانی

فهرست

پایگاه‌دانش



پیش‌بینی لینک:

- تایید یا رد کردن یک حقیقت از روی اطلاعات موجود در گراف داده

پایگاه‌دانش ناهمگون

- گرافی که گره‌های آن موجودیت‌ها و یال‌های آن گونه‌های مختلف رابطه را مشخص می‌کنند.

(head, relation, tail)

پایگاه‌دانش

Freebase

- قسمتی از گراف دانش گوگل (GKG)
- داده‌ها:
  - ۸۰ میلیون موجودیت
  - ۲۰ هزار رابطه مختلف
  - ۱.۲ میلیارد حقیقت
- نمونه
  - (Barack Obama, place_of_birth, Hawai)
  - (Albert Einstein, follows_diet, Veganism)
  - (San Francisco, contains, Telegraph Hill)

مساله‌ی پیش‌بینی لینک

چالش‌ها:

- تعداد ابعاد بالا (موجودیت‌ها، تعداد رابطه‌ها)
- بشدت خلوت (sparse) (تعداد لینک‌های صحیح بسیار کم است)
- داده‌های نویز دار و ناکامل (عدم وجود یا اشتباه بودن رابطه‌ها و موجودیت‌ها)

راه حل؟

- انتقال مساله به فضای برداری با تعداد ابعاد پایین
- یافتن لینک‌های صحیح ناموجود در پایگاه دانش

فهرست

روش‌های یادگیری آماری

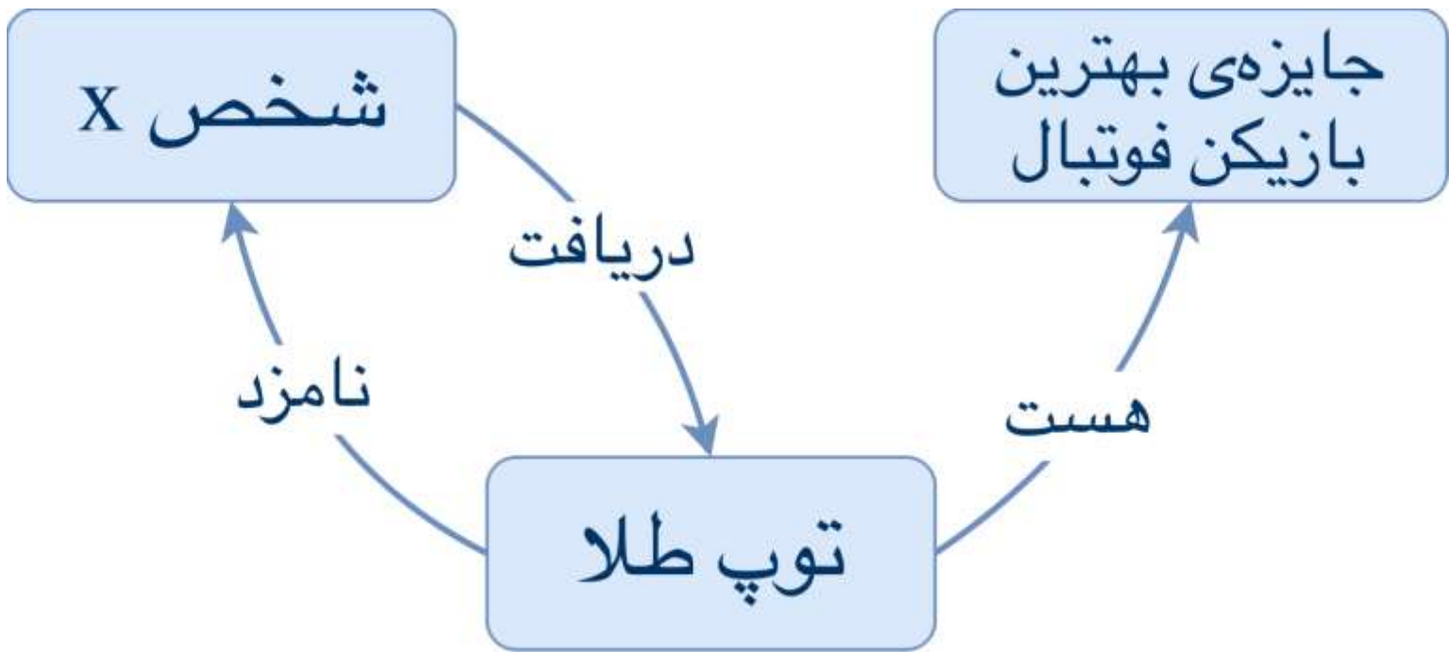
یادگیری آماری رابطه‌ها (Statistical Relational Learning):

تولید مدل‌های آماری برای داده‌های رابطه‌ای

- خصوصیات گراف (graph feature)
- مدل تصادفی مارکف (Markov random field)
- ویژگی‌های پنهان (Latent factor)

مدل ویژگی‌های پنهان

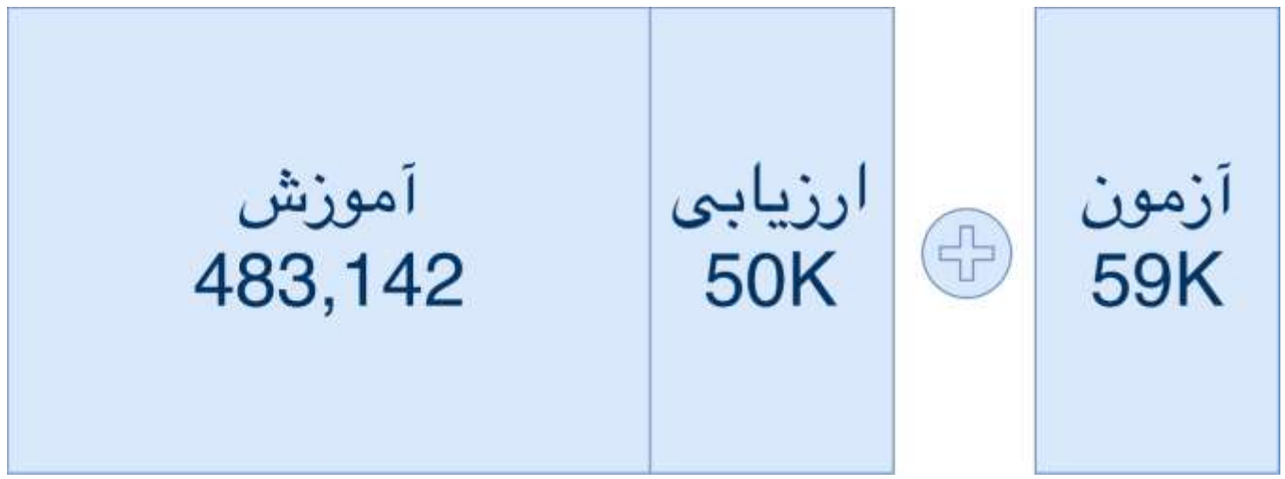
- هر موجودیت و نوع رابطه به صورت برداری از ویژگی‌ها تعریف می‌شوند که ویژگی‌های پنهان نام دارد.
- به طور مستقیم از روی داده‌ها تصمیم‌گیری نمی‌شود و از ویژگی‌های پنهان هر موجودیت یا رابطه تصمیم‌گیری انجام می‌شود.
- مثال:



پایگاه دانش

Freebase 15k

- یک نمونه نرمال از پایگاه‌دانش اصلی
- داده‌ها:
  - ۱۴,۹۵۱موجودیت
  - ۱,۳۴۵رابطه مختلف
  - ۵۹۲،۲۱۳ حقیقت



توزیع مجموعه داده‌ها در فرایند آموزش

(WALL-E, has_genre, Fantasy)

WALL-E has the genre?!

(WALL-E, has_genre, ?)

1- Animations
2- Computer Animation
3- Comedy film
4- Adventure film
5- Science Fiction
6- Fantasy
7- Stop motion
8- Satire
…

=>

- ✓ Mean rank
- ✓ hit@10

روش‌های مبتنی بر ویژگی‌های پنهان

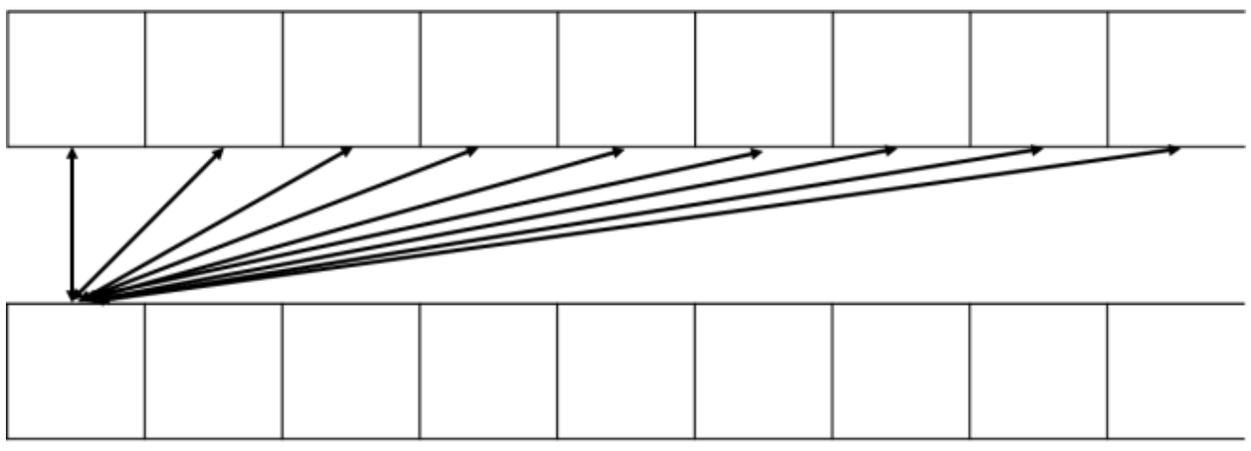
- Bilinear
  - RESCAL
- Multi-layer Perceptions
  - (NTN) Neural Tensor network
- Latent distance models
  - Structured Embedding
  - Translating Embedding (TransE)
  - Translating on Hyperplane (TransH)

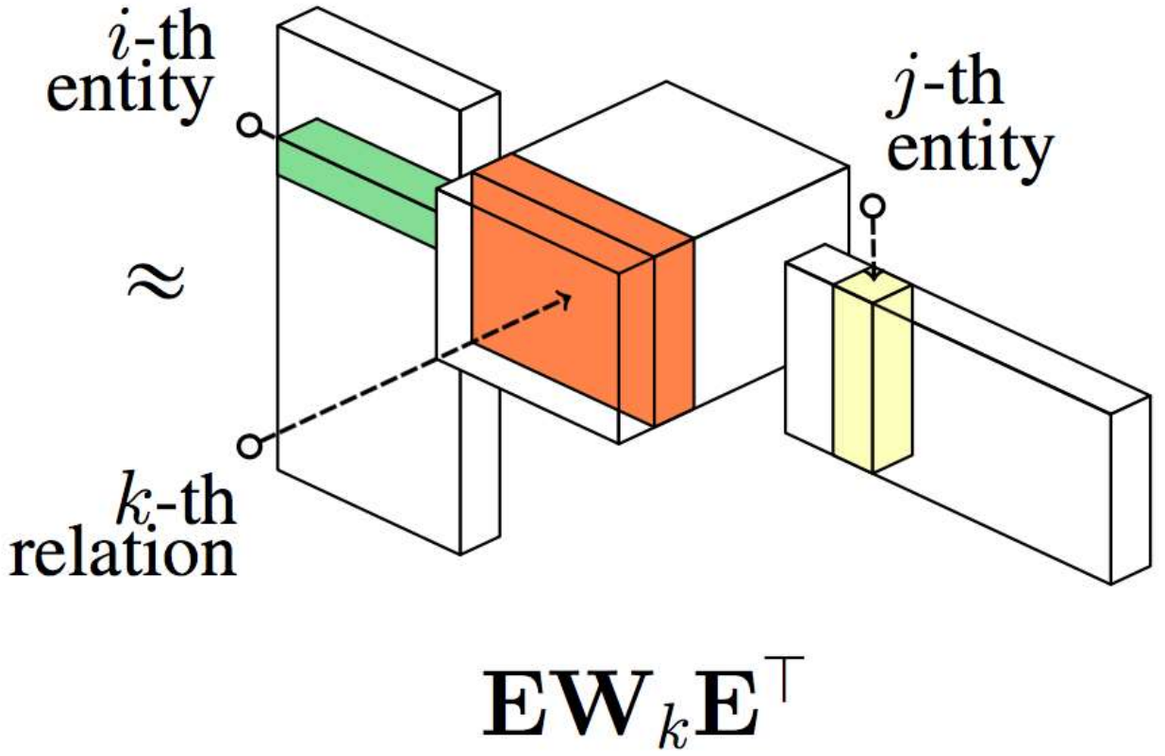
روش رسکال (دوخطی)

- هر جفت ویژگی از دو موجودیت را در فضای رابطه‌ی مورد نظر مورد بررسی قرار می‌دهد.
- تابع امتیاز:

$$f_{ijk}^{\text{RESCAL}} := \mathbf{e}_i^\top \mathbf{W}_k \mathbf{e}_j = \sum_{a=1}^{H_e} \sum_{b=1}^{H_e} w_{abk} e_{ia} e_{jb}$$

$\mathbf{W}_k \in \mathbb{R}^{H_e \times H_e}$ is a weight matrix





روش‌های فاصله‌ی پنهان (LDM)

- احتمال رابطه بین موجودیت‌ها، از روی فاصله‌ی آنها در فضای برداری مشخص می‌شود.
- تابع امتیاز در روابط تک-رابطه‌ای پس از بازنمایی موجودیت‌ها در فضای برداری به صورت:

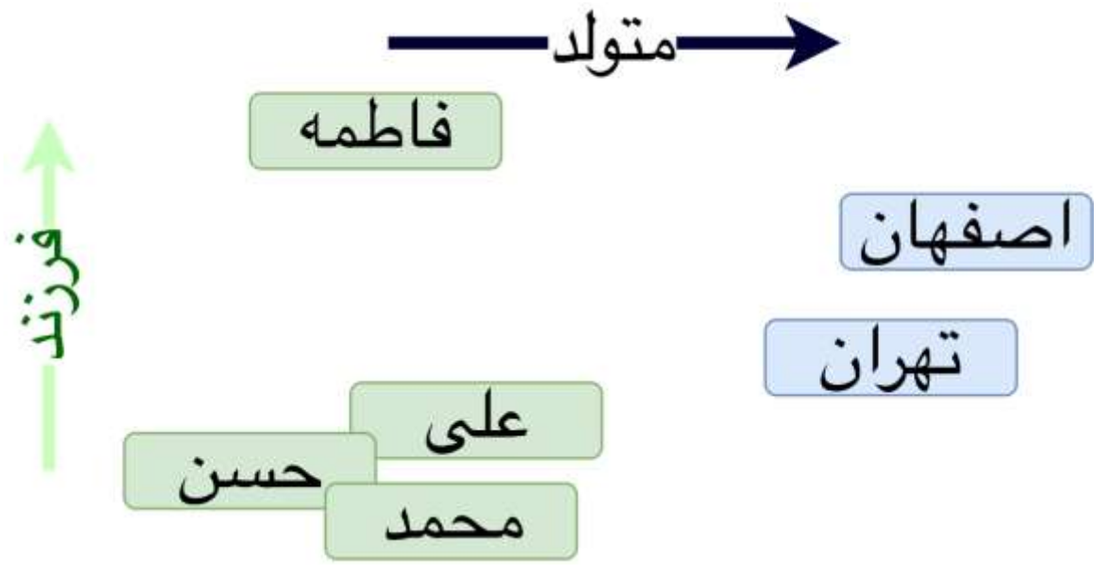
$$f(e_i, e_j) = -d(\mathbf{e}_i, \mathbf{e}_j)$$

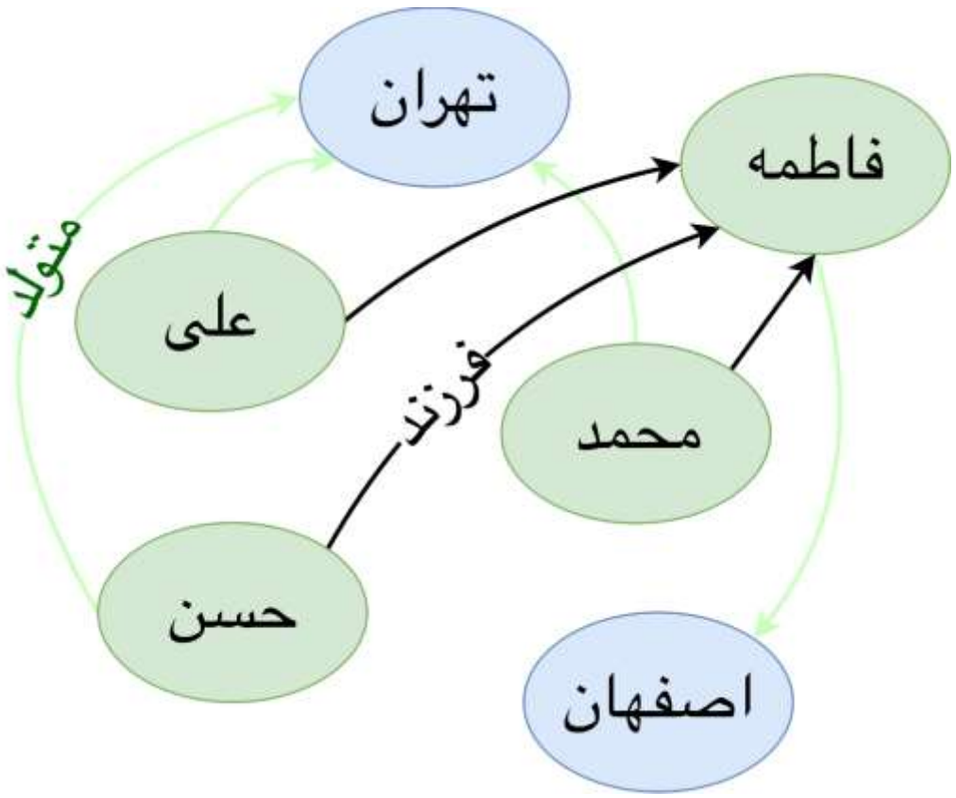
TransE

- برای کاهش تعداد پارامترها رابطه‌ها را بجای ضرب ماتریسی، یک آفست در کنار موجودیت اول در نظر می‌گیریم.

$$f_{ijk}^{\text{TransE}} := -d(\mathbf{e}_i + \mathbf{r}_k, \mathbf{e}_j)$$

$$d(sub, rel, obj) = ||\mathbf{s} + \mathbf{r} - \mathbf{o}||_2^2$$





مقیاس‌پذیری

Method	#Params	On FB15K
RESCAL	$O(n_e d + n_r d^2)$	88M (d=250)
MLP (NTN)	$O(n_e d + n_r d^3)$	165M (d=50)
SE	$O(n_e d + 2n_r d^2)$	8M (d=50)
TransE	$O(n_e d + n_r d)$	0.8M (d=50)

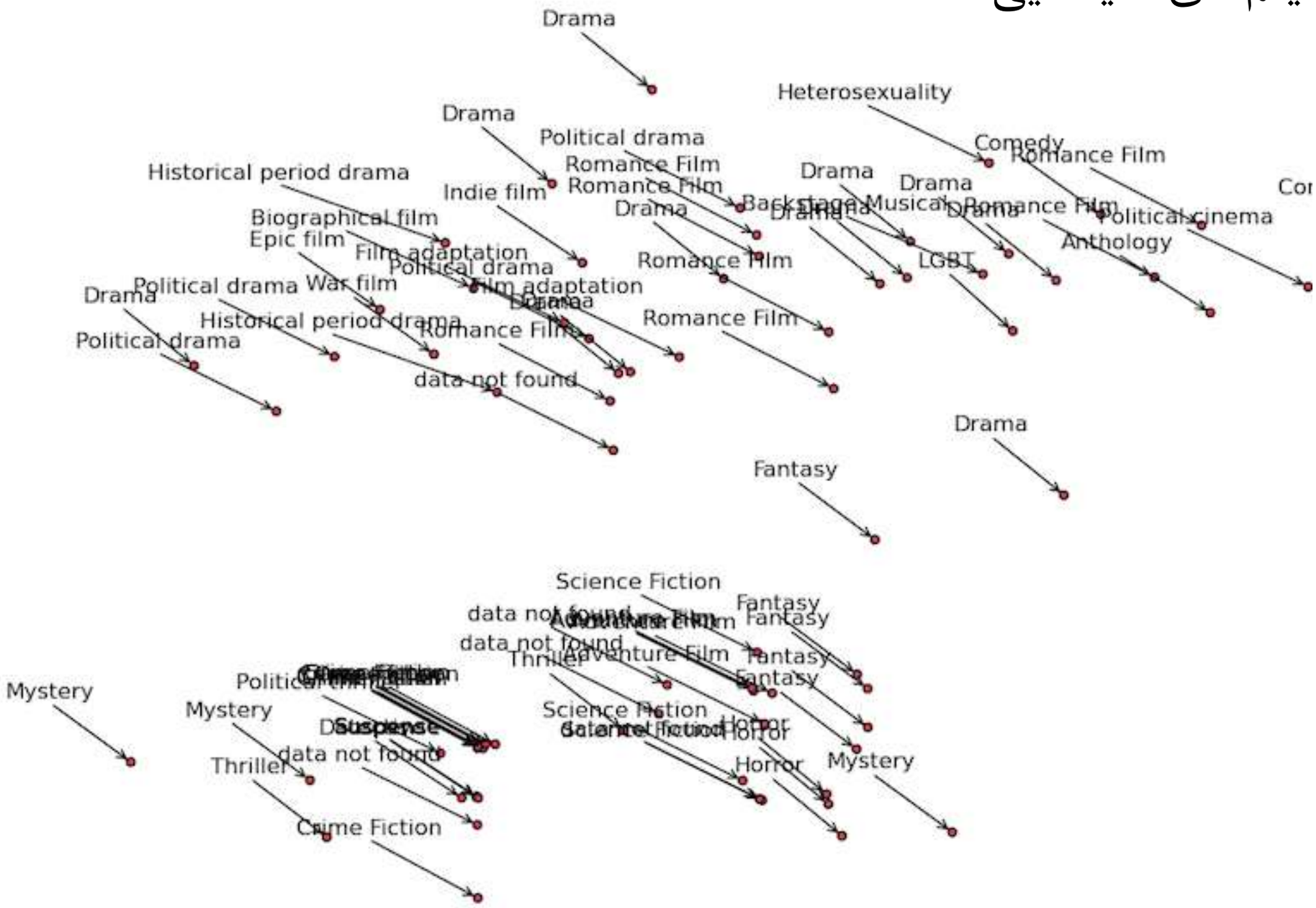
Freebase15k: n_e=15k, n_r=1.3k

پراکندگی موجودیت‌ها

- پراکندگی ۵,۰۰۰ موجودیت در روش TransE

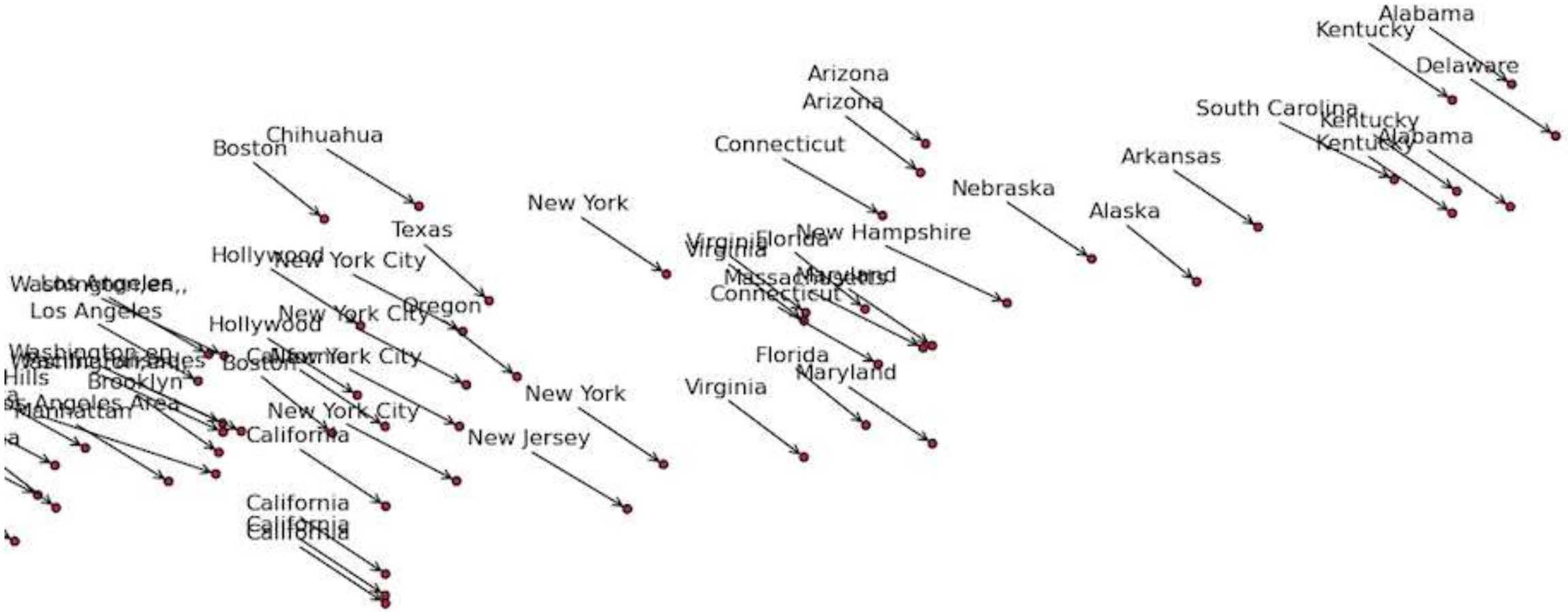
پراکندگی موجودیت‌ها

- ژانر فیلم‌های سینمایی

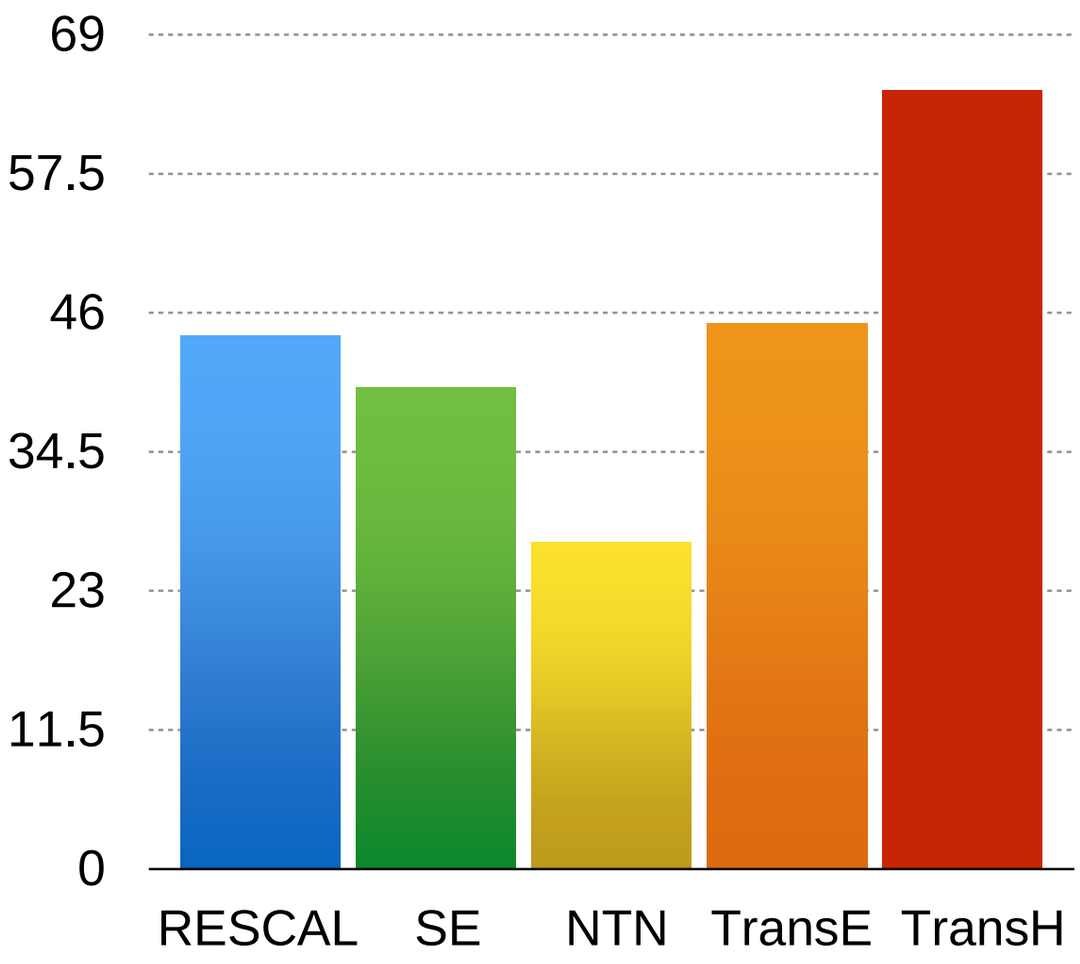


پراکندگی موجودیت‌ها

- ایالت‌های آمریکا



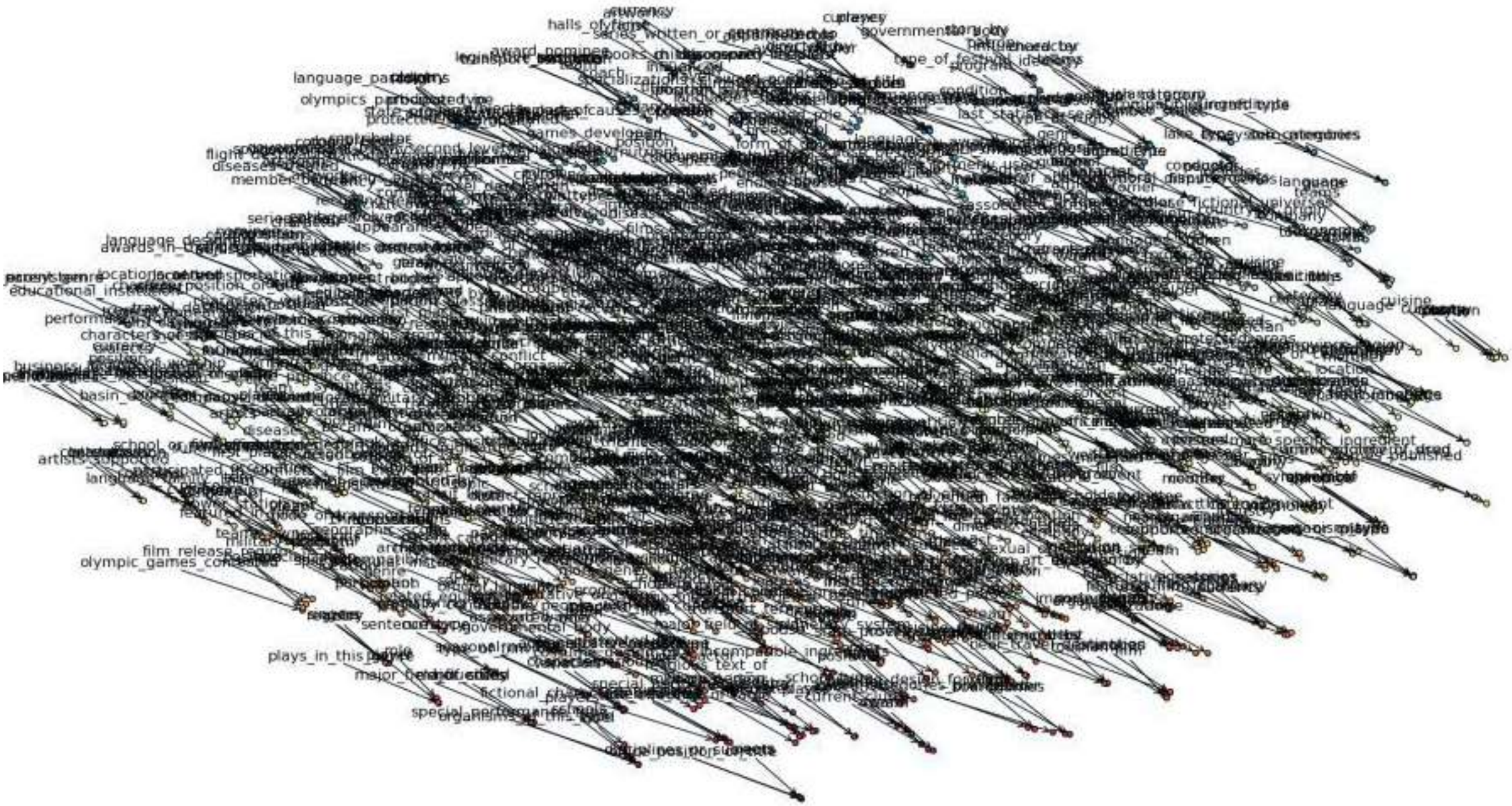
نتایج



	Hit@10	Mean
RESCAL	42.1%	683
NTN	27%	164
SE	39.8%	162
TransE	45.1%	125
TransH	64.4%	87

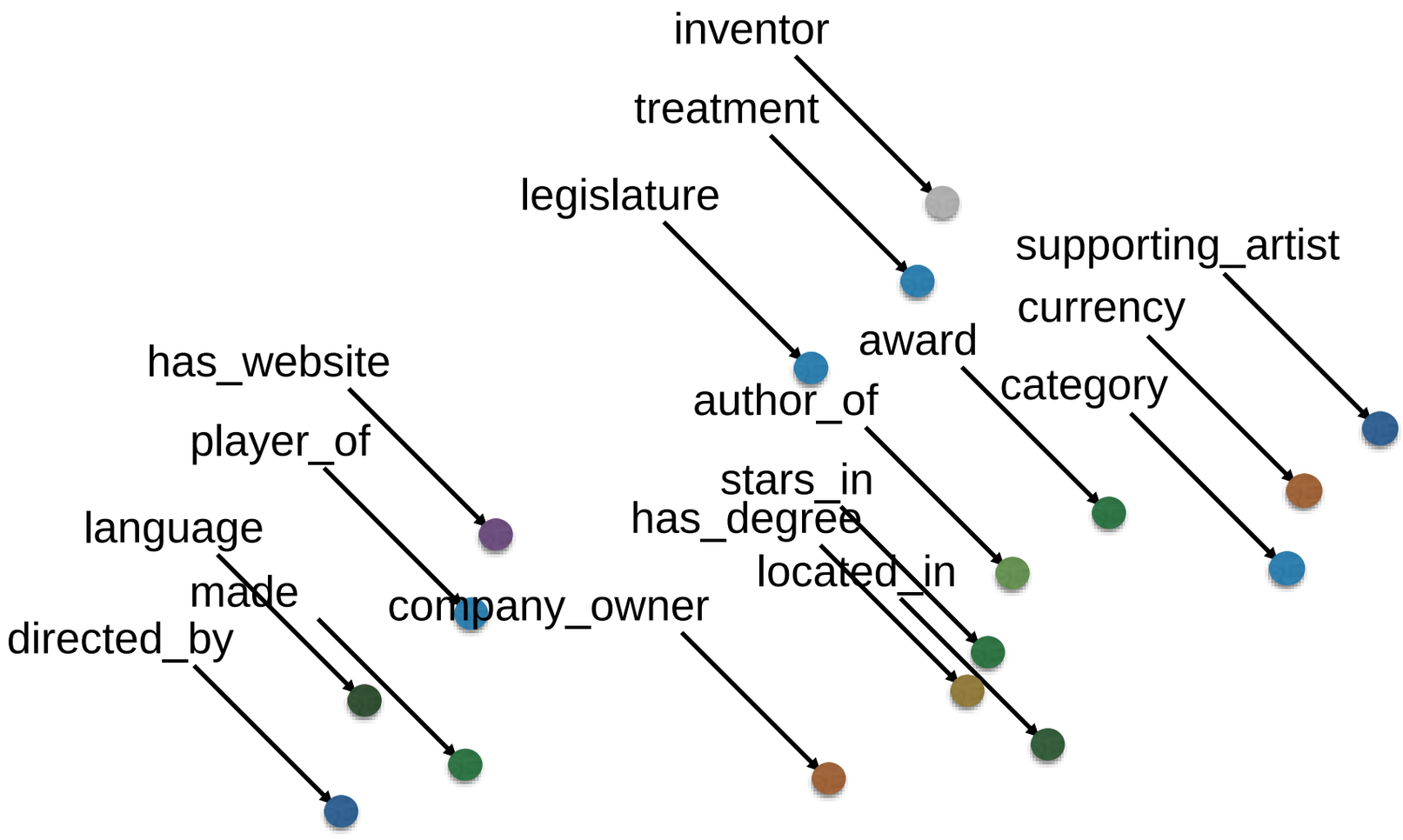
پراکندگی رابطه‌ها

- پراکندگی همه رابطه‌ها در روش TransE



پراکندگی رابطه‌ها

- بی‌ربط بودن رابطه‌های نزدیک به هم

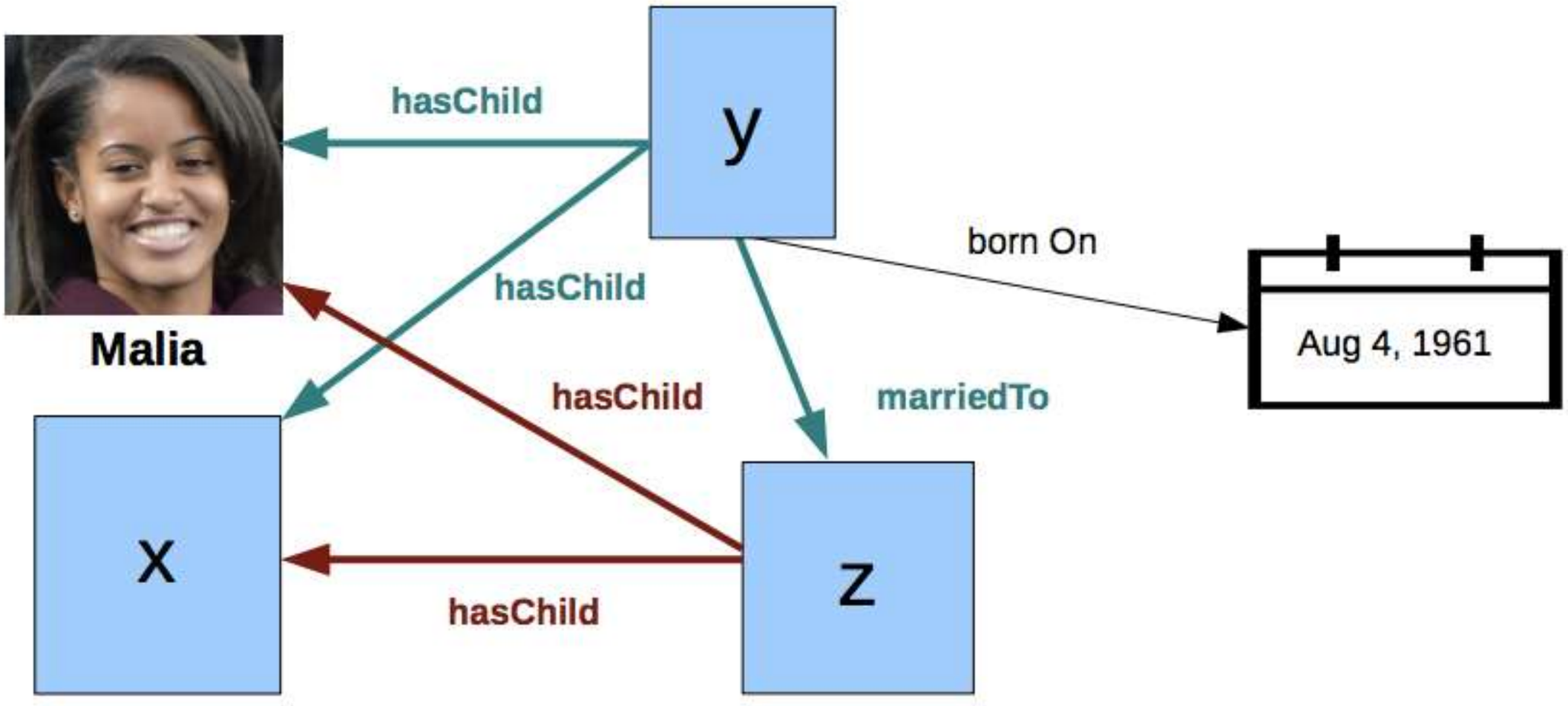


فهرست

استفاده از قوانین انجمنی

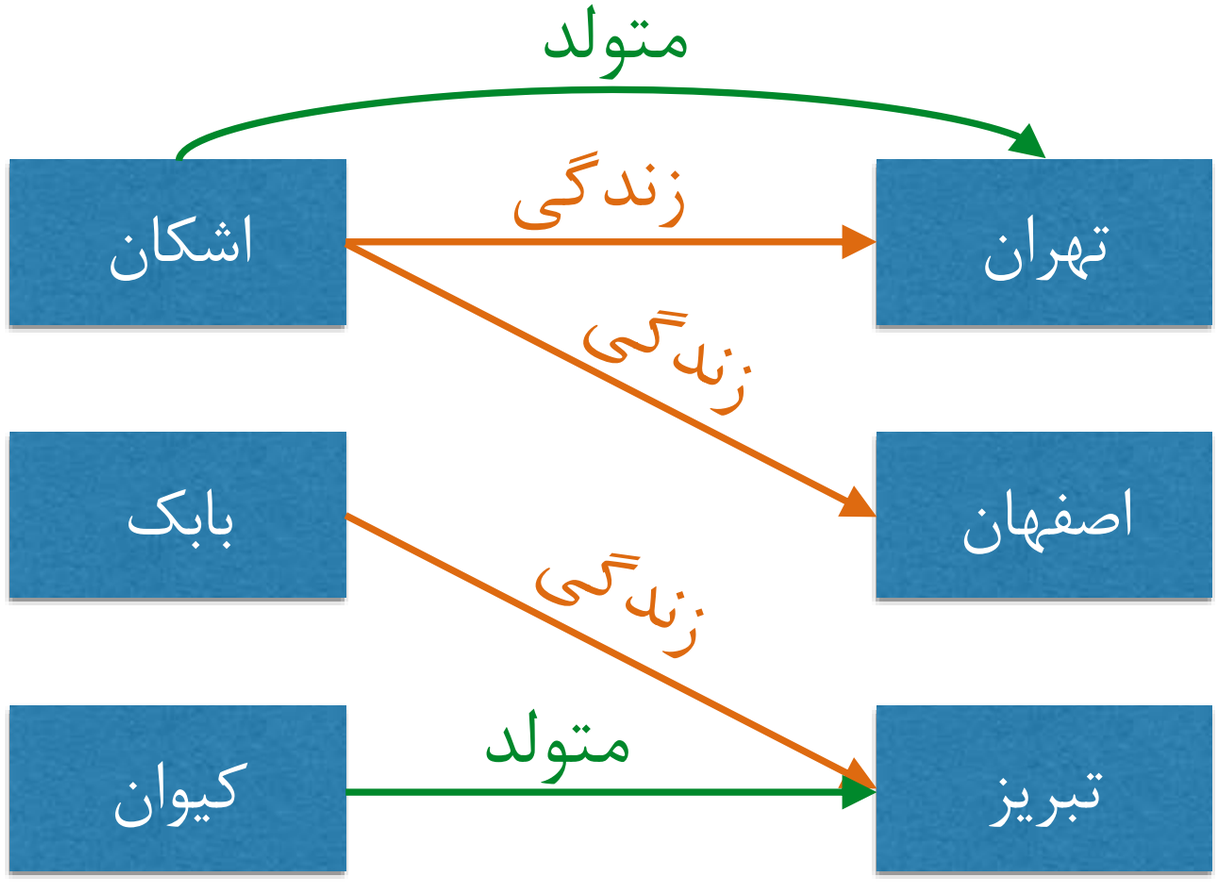
- استخراج قوانین پرتکرار نهفته در روابط موجود در پایگاه دانش با درصد اطمینان منطقی
- چالش‌ها:
  - اطلاعات موجود در پایگاه‌های دانش کامل نیست
  - فرض جهان باز (OWA) بودن مساله
  - معیار اطمینان از درستی قانون‌های بدست آمده
  - حجم بالای اطلاعات پایگاه‌دانش

قوانین انجمنی



hasChild(y, x), marriedTo(y, z) => hasChild(z, x)

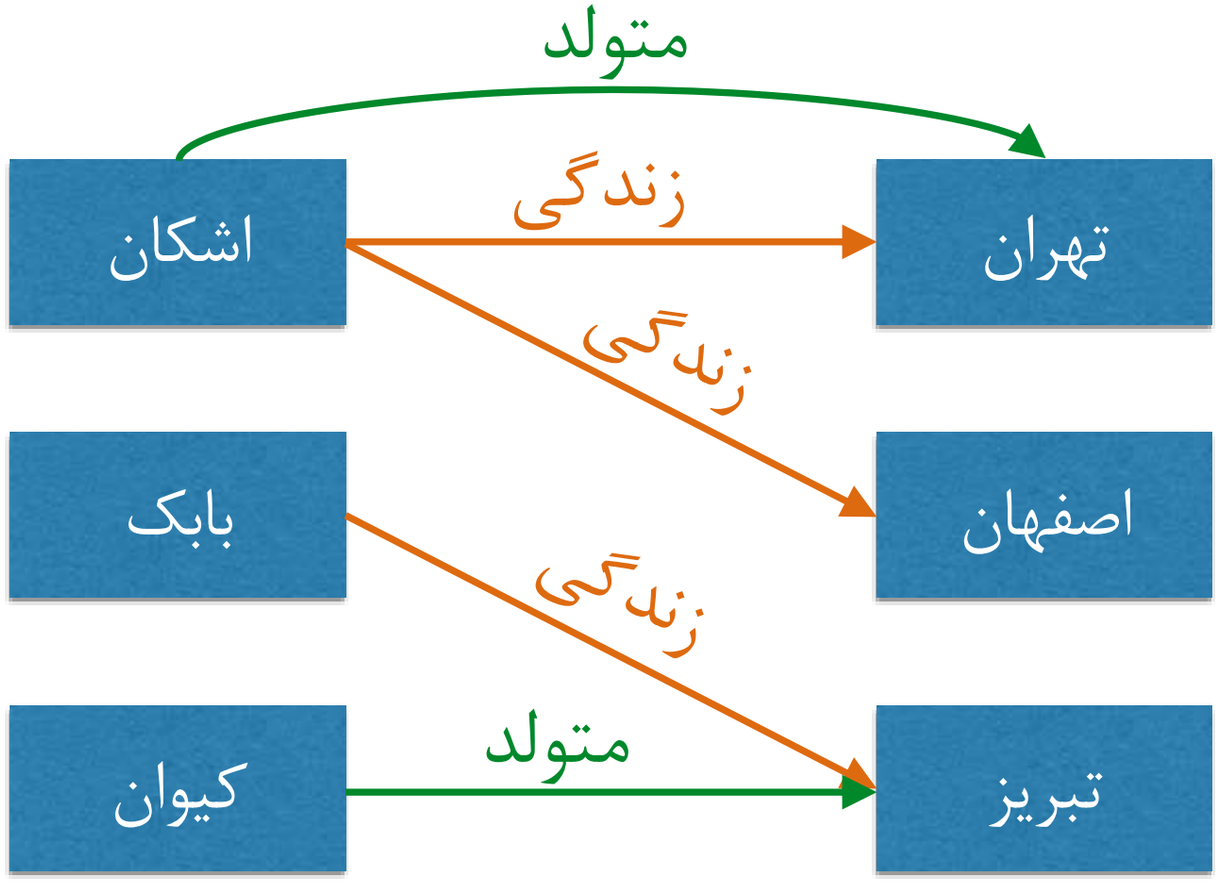
معیار اطمینان استاندارد



زندگی(الف، ب) ← متولد(الف، ب)

$$conf(\overrightarrow{B} \Rightarrow r(x,y)) := \frac{supp(\overrightarrow{B} \Rightarrow r(x,y))}{\#(x,y) : \exists z_1, ..., z_m : \overrightarrow{B}} = \frac{1}{3}$$

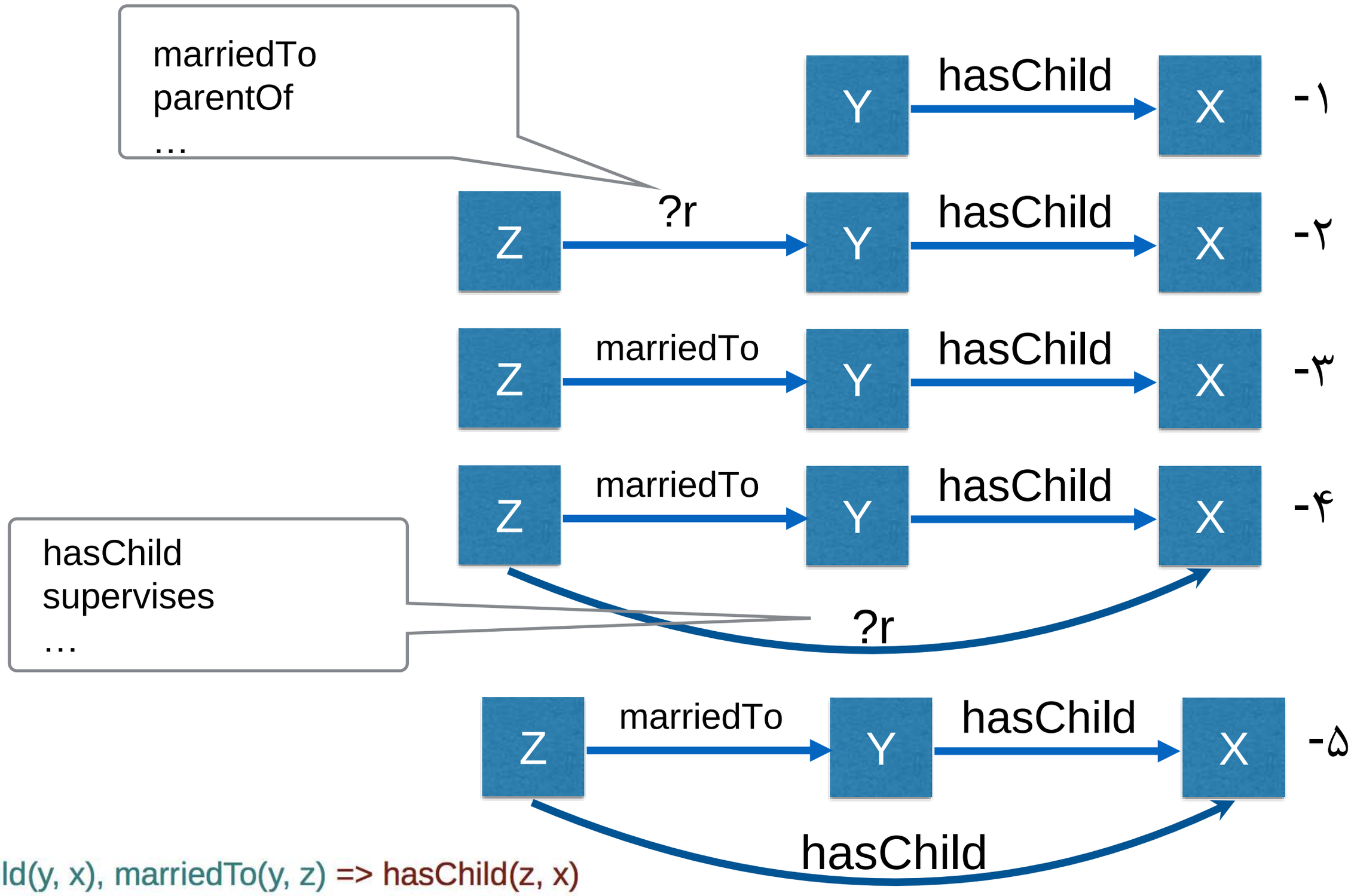
معیار اطمینان با فرض نیمه‌کامل



زندگی(الف، ب) ← متولد(الف، ب)

$$conf_{pca}(\overrightarrow{B} \Rightarrow r(x,y)) := \frac{supp(\overrightarrow{B} \Rightarrow r(x,y))}{\#(x,y) : \exists z_1, ..., z_m, y' : \overrightarrow{B} \wedge r(x,y')} = \frac{1}{2}$$

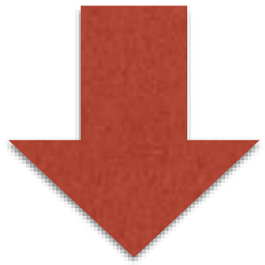
استخراج قوانین انجمنی



hasChild(y, x), marriedTo(y, z) => hasChild(z, x)

نمونه قوانین انجمنی

?b /award/awards_won ?a



?a /award/award_nomination/nominated_for ?b

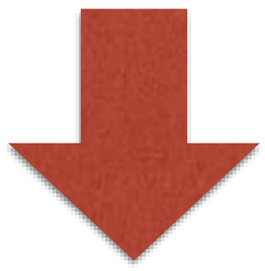
PCA Confidence = 0.83

نمونه قوانین انجمنی

?a /educational_institution/located_in ?f

and

?f /location/location/containedby ?b



?a /location/location/containedby ?b

PCA Confidence = 0.93

قوانین انجمنی

- دسته‌بندی قوانین

Body rule		Target rule	name
$r(x, y)$	=>	$r'(x, y)$	R-subsumption
$r(x, y)$	<=>	$r'(x, y)$	R-equivalence
$r_1(x, y), r_2(y, z)$	=>	$r'(x, z)$	2-hope translation
$r(x, z), r(y, z)$	=>	$r'(x, y)$	Triangle alignment
$r_1(x, y), r_2(x, v)$	=>	$r'(x, y)$	Specific R-sub

استفاده از قوانین انجمنی

- استفاده از قوانین انجمنی در آموزش TransE

- R-Subsumption

- $r(e_1,e_2) \Rightarrow r'(e_1,e_2)$

$$\| e_1 + r - e_2 \| \Rightarrow \| e_1 + r' - e_2 \|$$

استفاده از قوانین انجمنی

- استفاده از قوانین انجمنی در آموزش TransE

- 2-hope translation

- $r_1(e_1, e_2)$, $r_2(e_2, e_3)$ => $r'(e_1, e_3)$

$$\| e_1 + r_1 - e_2 \| \text{ and } \| e_2 + r_2 - e_3 \|$$
$$\Rightarrow \quad \| e_1 + r' - e_3 \|$$

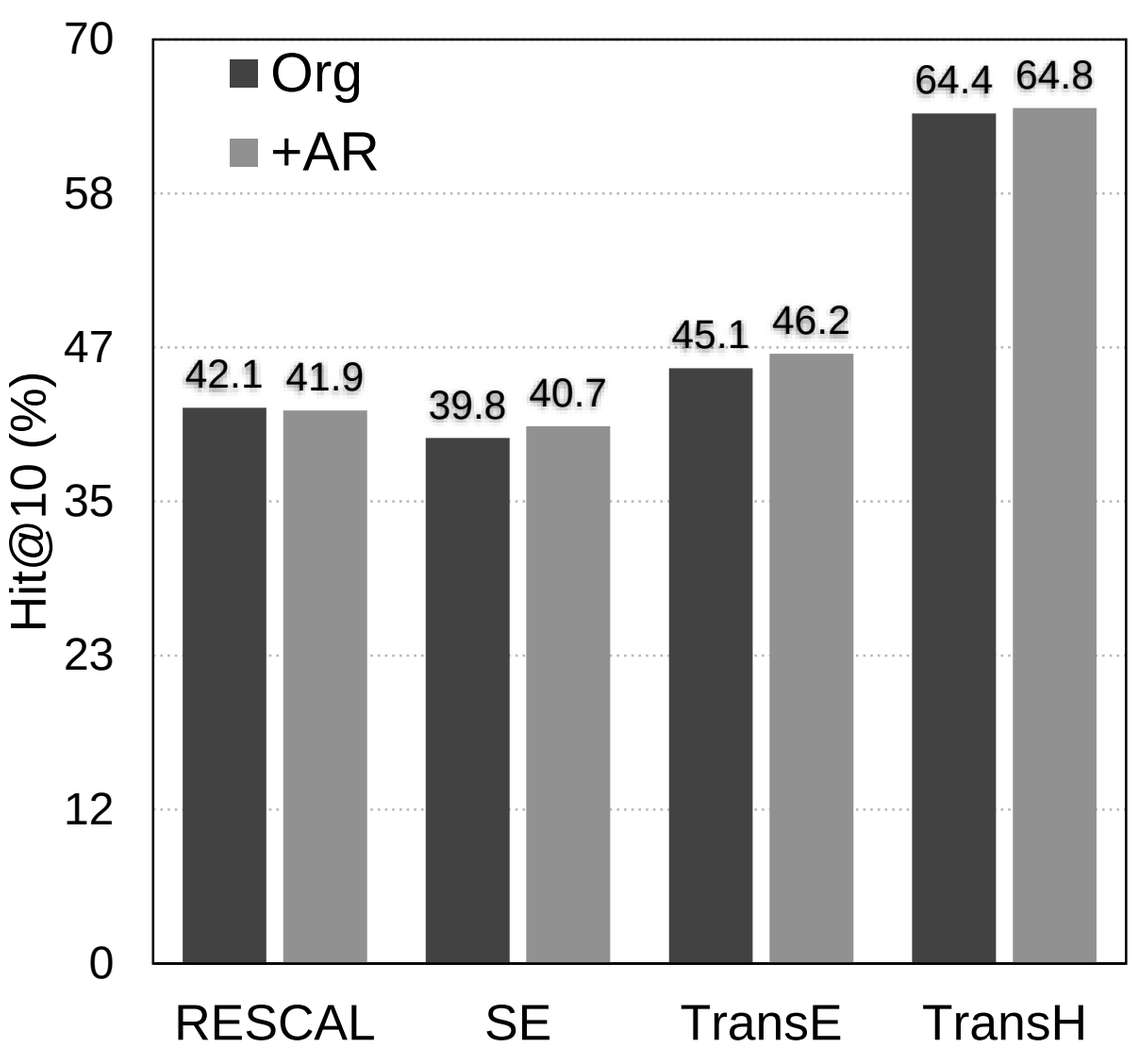
$$r_1 + r_2 \Rightarrow r'$$
$$\| e_1 + r' - e_3 \| \quad \Rightarrow \quad \| e_1 + r_1 + r_2 - e_3 \|$$

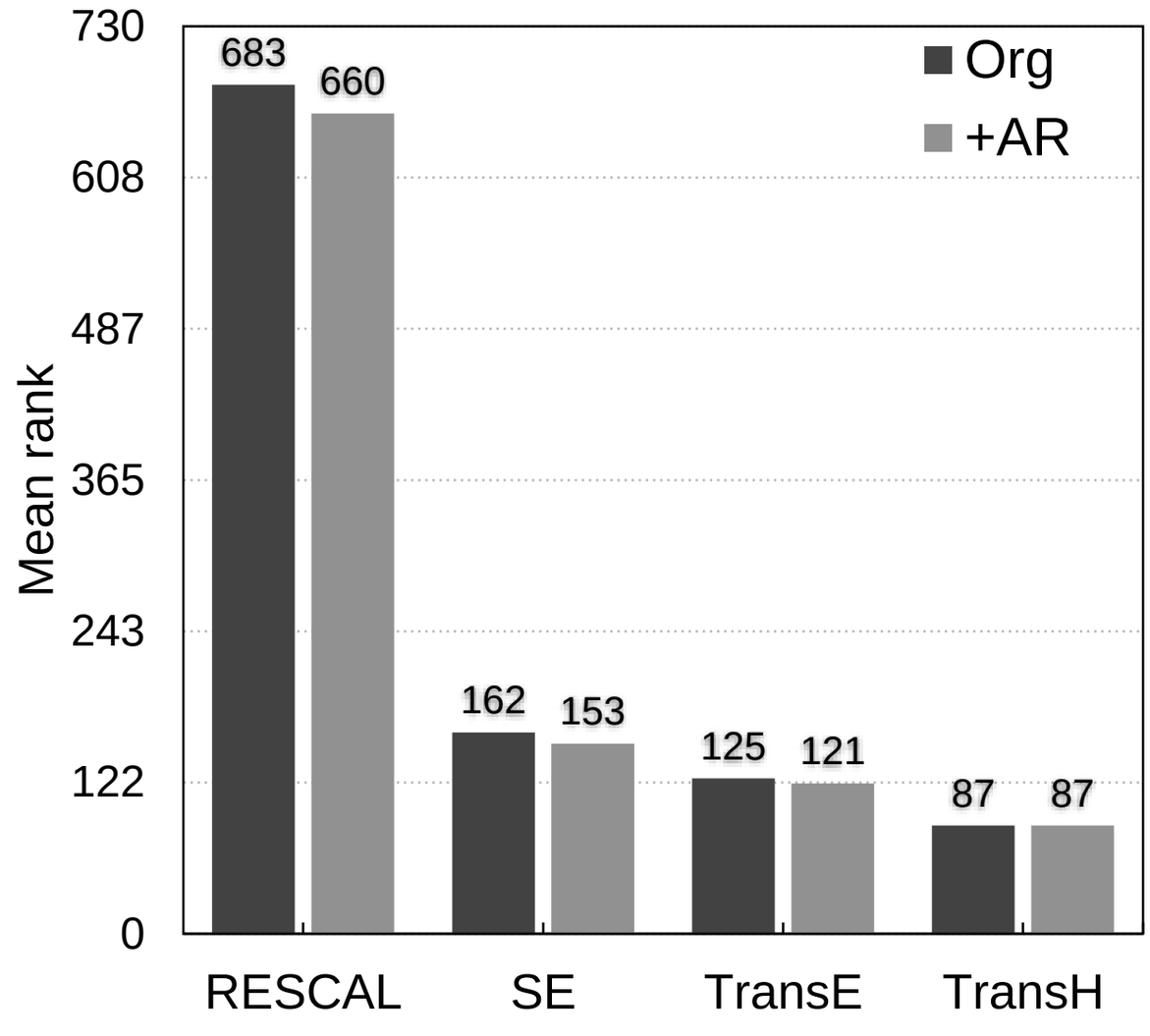
فهرست

آزمایش‌ها

اعمال قانون R-subsumption

r(x, y) => r'(x, y)

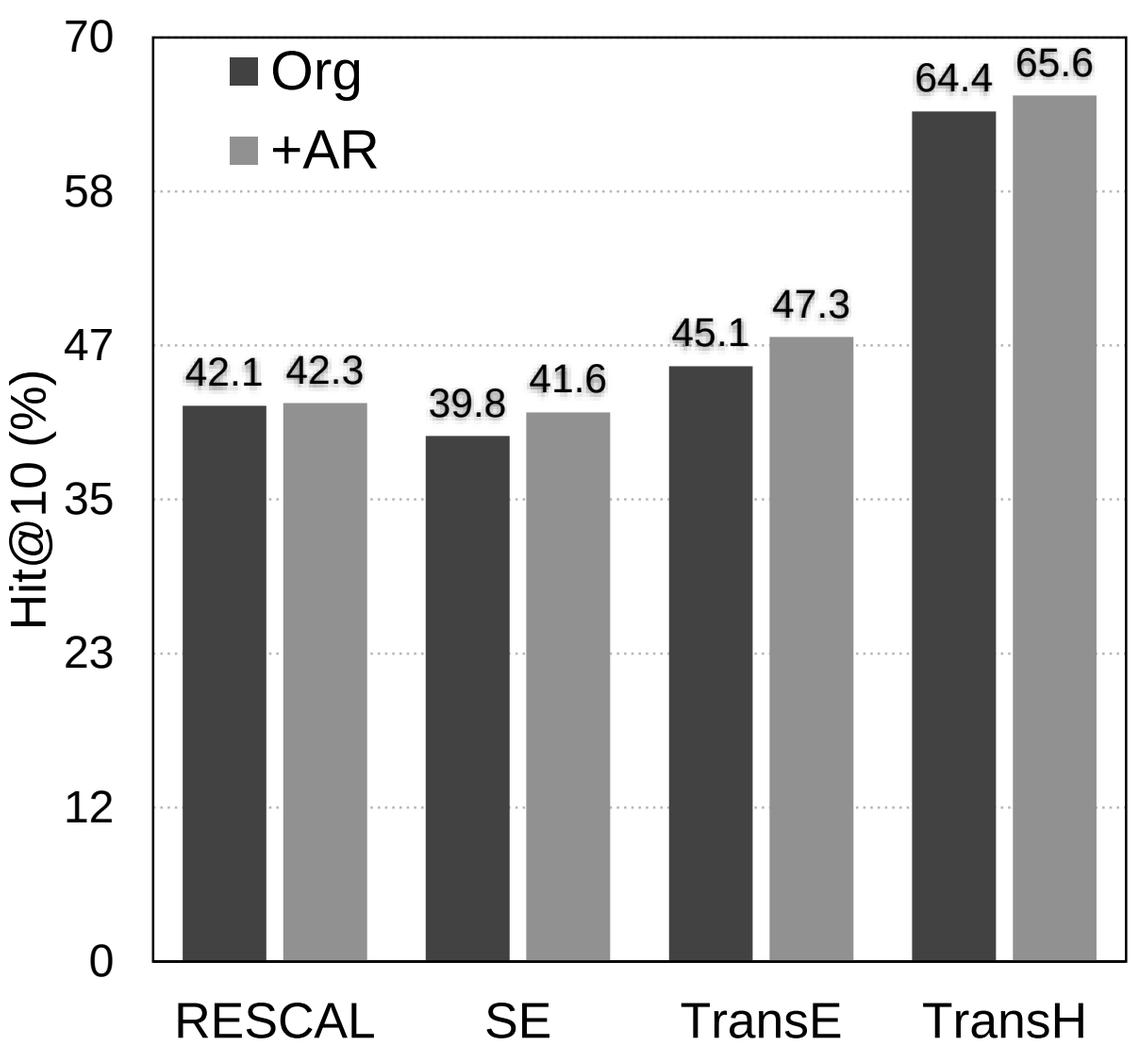


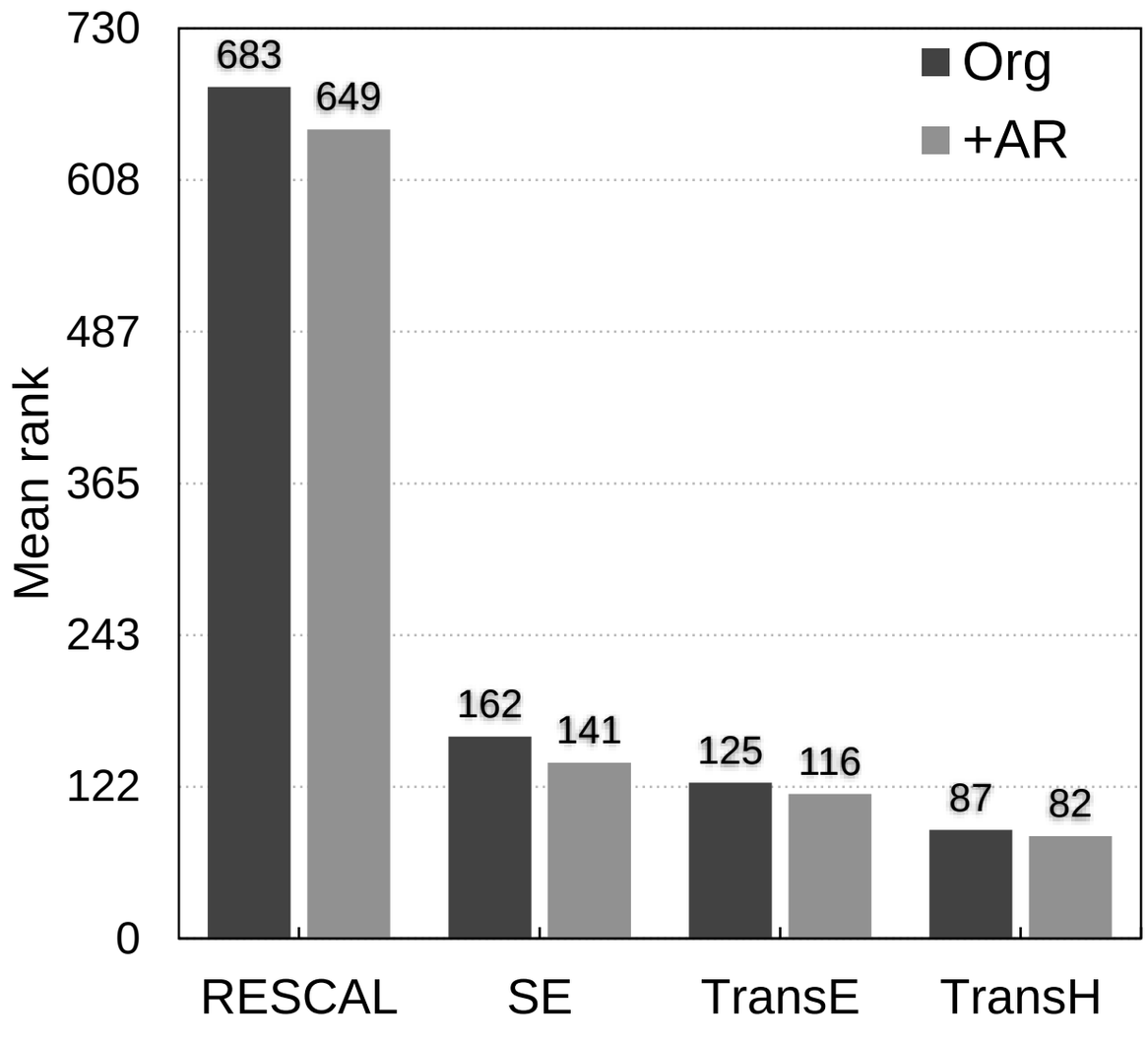


آزمایش‌ها

اعمال قانون R-Equivalence

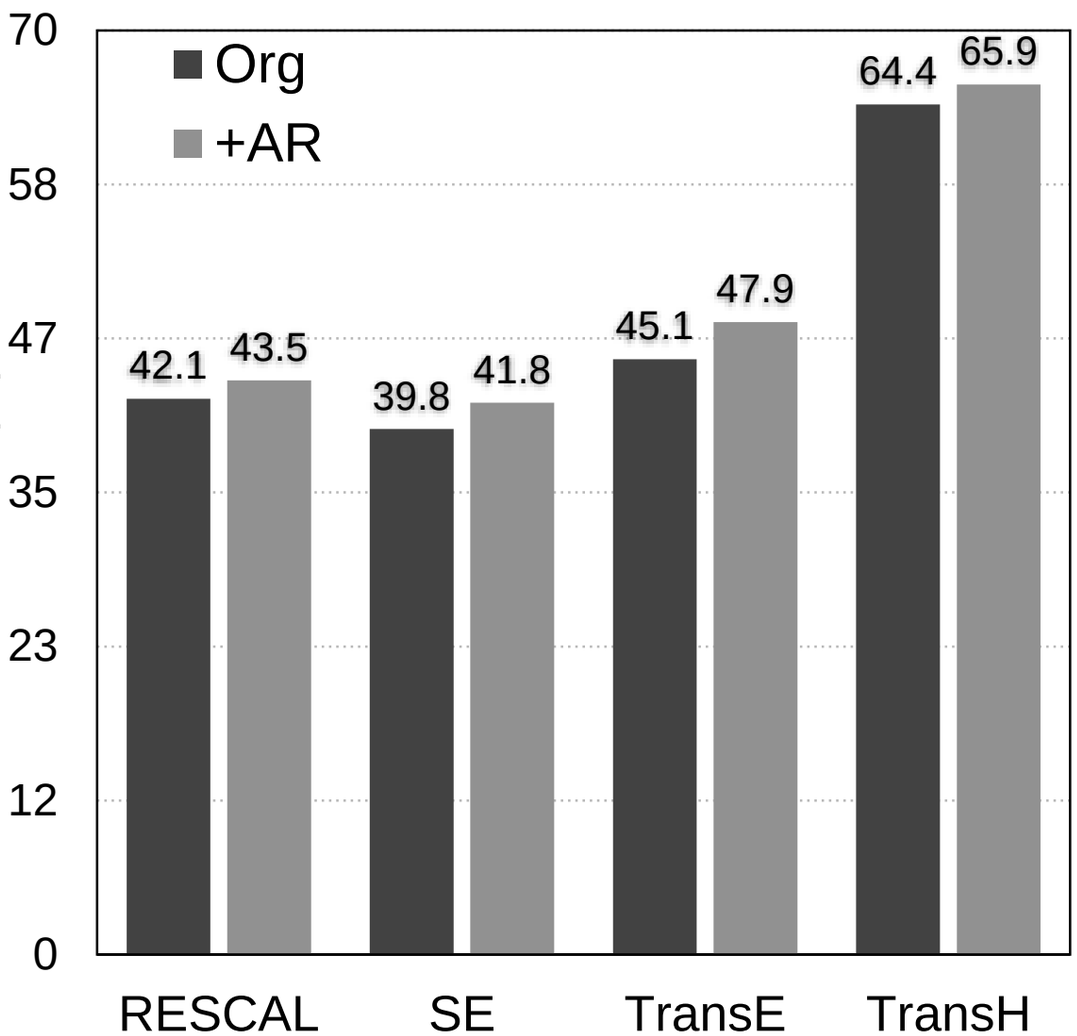
r(x, y) <=> r'(x, y)

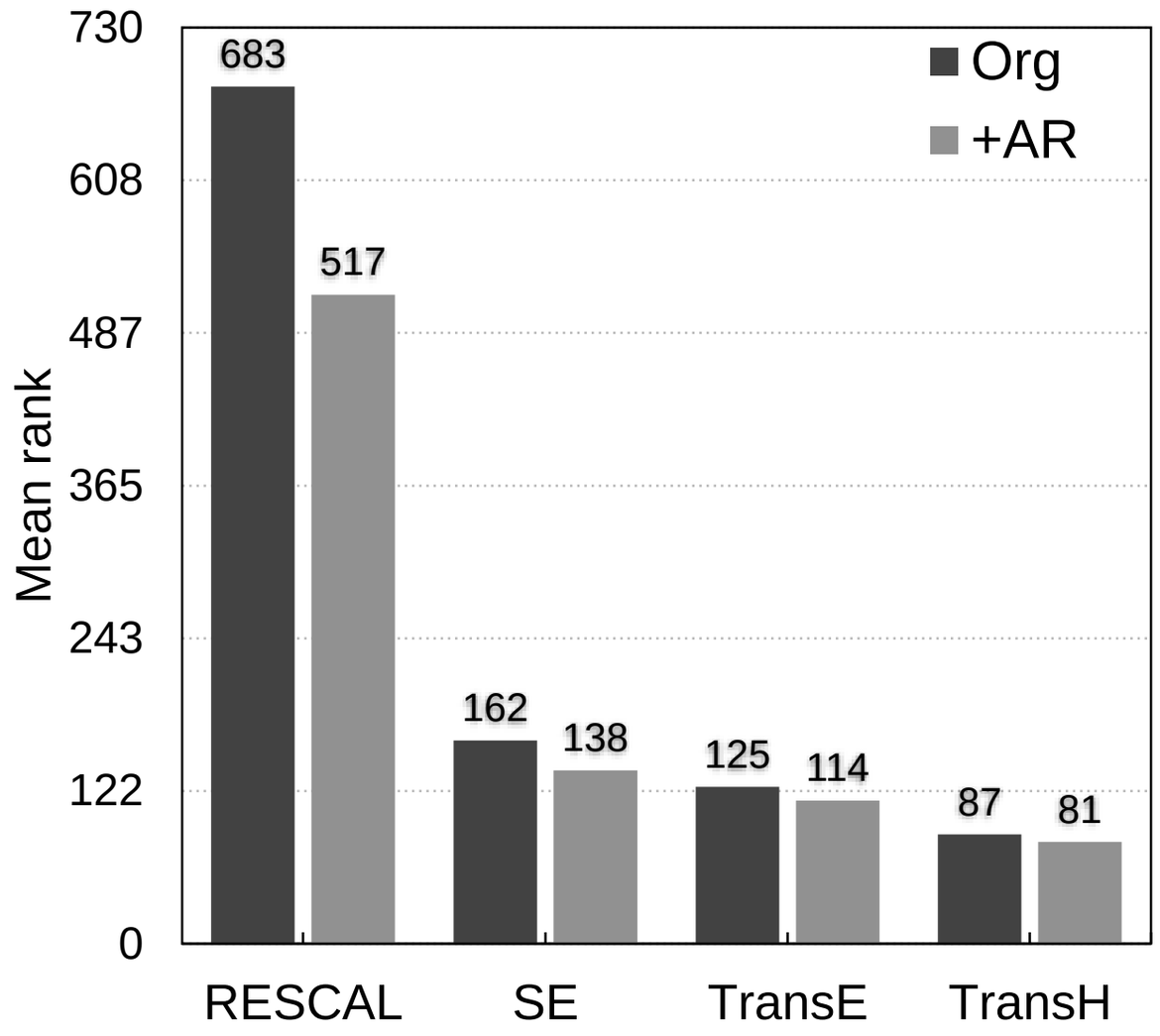




آزمایش‌ها

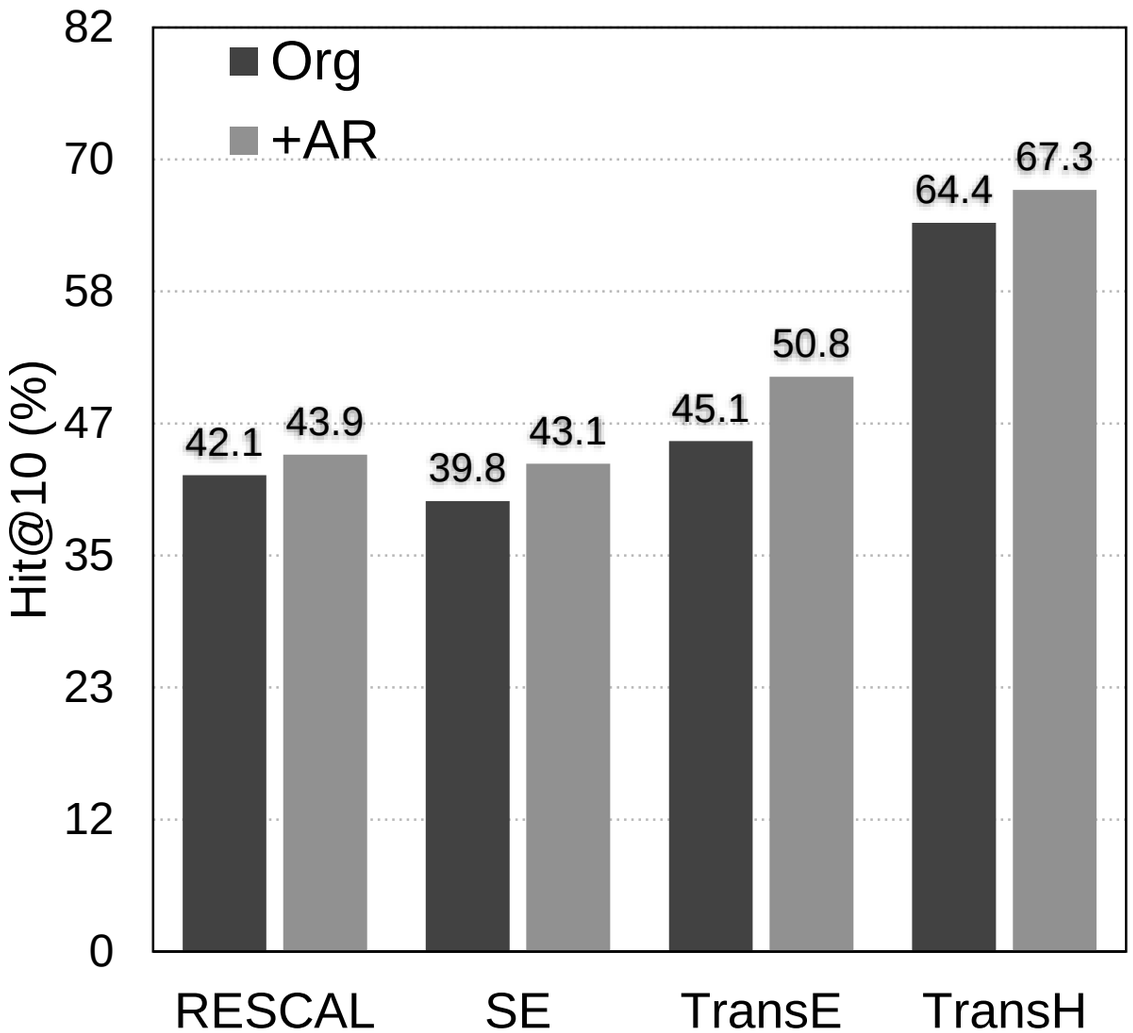
اعمال قانون 2-Hops Translation r1(x, y), r2(y, z) => r'(x, z)

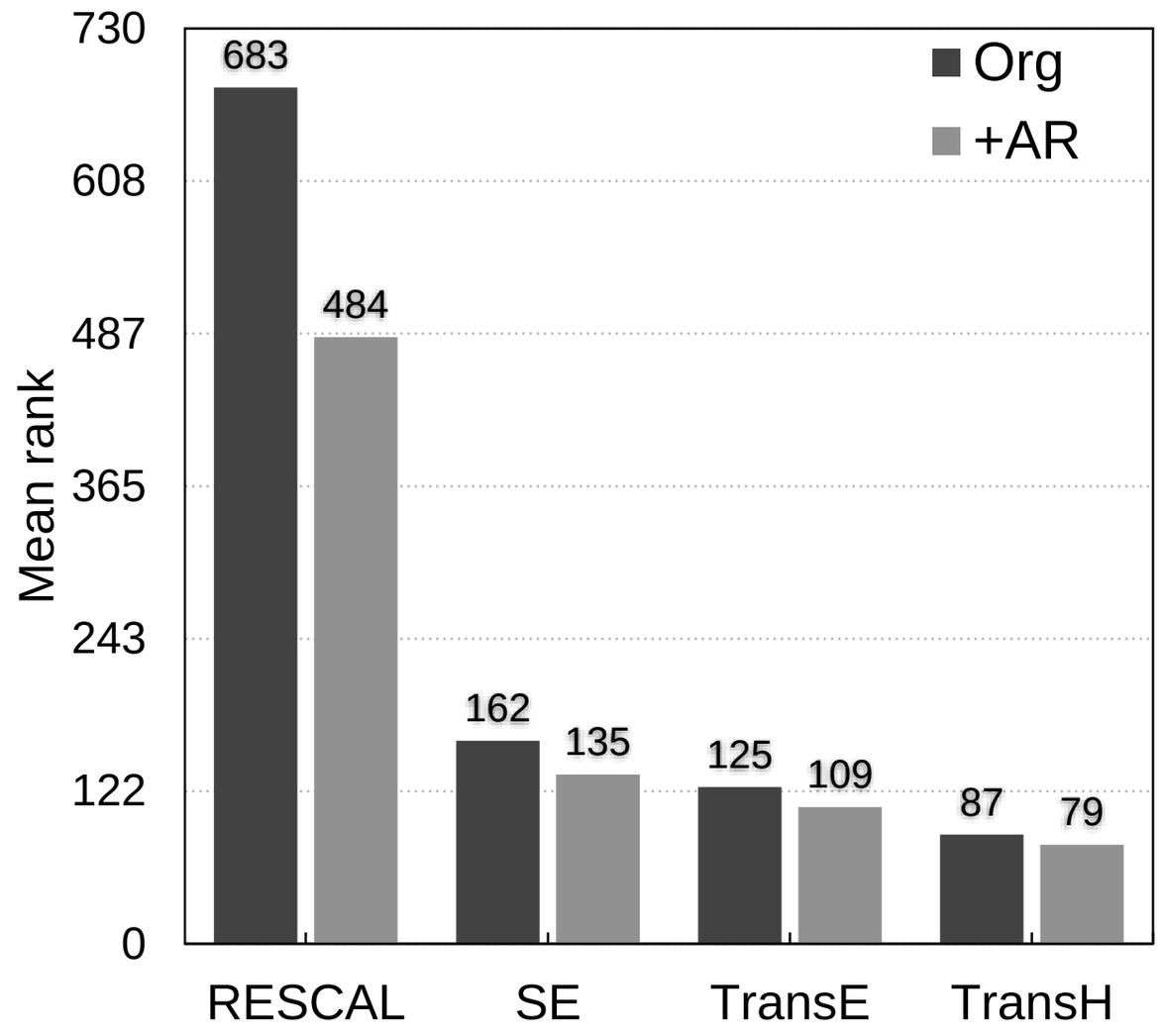




آزمایش‌ها

اعمال همه‌ی قانون‌ها





Hit@10

Hit@10 on Freebase 15K (%)							
Method	**orginal**	**R-Subsumption**	**R-Equivalence**	**2-hope**	**Triangle alignment**	**SR-Subsumption**	**all rules (~improve)**
RESCAL	42.1	41.9	42.3	43.5	42.9	42.0	**43.9 (1.8%)**
NTN	27	-	-	-	-	-	**-**
SE	39.8	40.7	41.6	41.8	41.5	40.5	**43.1 (3.3%)**
TransE	45.1	46.2	47.3	47.9	46.9	46.1	**50.8 (5.7%)**
TransH	64.4	64.8	65.6	65.9	65.2	64.9	**67.3 (2.9%)**

Mean rank

Mean rank on Freebase 15K							
Method	**orginal**	**R-Subsumption**	**R-Equivalence**	**2-hope**	**Triangle alignment**	**SR-Subsumption**	**all rules (~improve)**
RESCAL	683	660	649	517	572	644	**484 (30%)**
NTN	164	-	-	-	-	-	-
SE	162	153	141	138	144	156	**135 (17%)**
TransE	125	121	116	114	116	122	**109 (13%)**
TransH	87	87	82	81	83	87	**79 (10%)**

نوآوری

- بهبود مدل‌های پیش‌بینی لینک مبتنی بر ویژگی‌های پنهان با استفاده از قوانین انجمنی
- ارائه‌ی چارچوبی برای سنجش نقاط قوت و ضعف روش‌های موجود مبتنی بر ویژگی‌های پنهان

کارهای آینده

- استفاده از کشف جامعه (community detection) برای فیلتر کردن نتایج و حذف جواب‌های نامربوط به سوال در روش‌های مبتنی بر ویژگی‌های پنهان
- کشف و استفاده‌ی قوانین با پیچیدگی بیشتر از ۲ مرحله

سوال؟

سپاس.