RelSim: Relation Similarity Search in Schema-Rich Heterogeneous Information Networks

Chenguang Wang, Yizhou Sun, Yanglei Song, Jiawei Han,

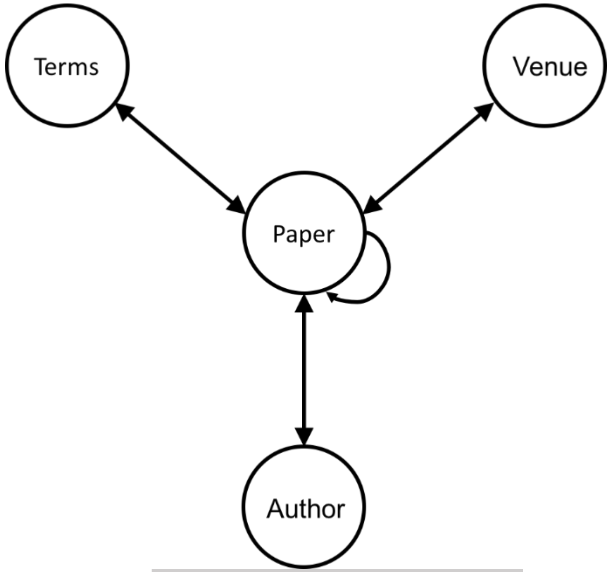
Yangqiu Song, Lidan Wang, Ming Zhang

Outline

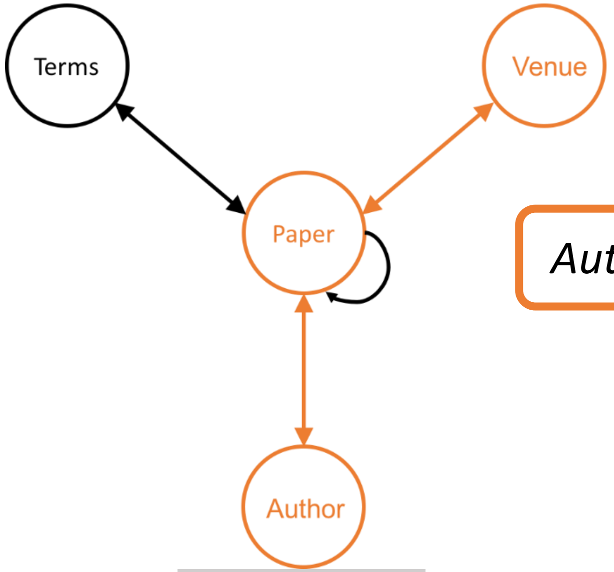
Motivation	The issues of previous HIN studies
RelSim	Compute the similarity between relation instances
Experiments	Achieve the-state-of-arts similarity search results on five datasets

Heterogeneous Information Networks

- HIN: Network with multiple object types and/or multiple link types, e.g., DBLP.
- Network schema: High-level description of a network.
- Meta-path: A **path/link** in the network schema.



Network schema

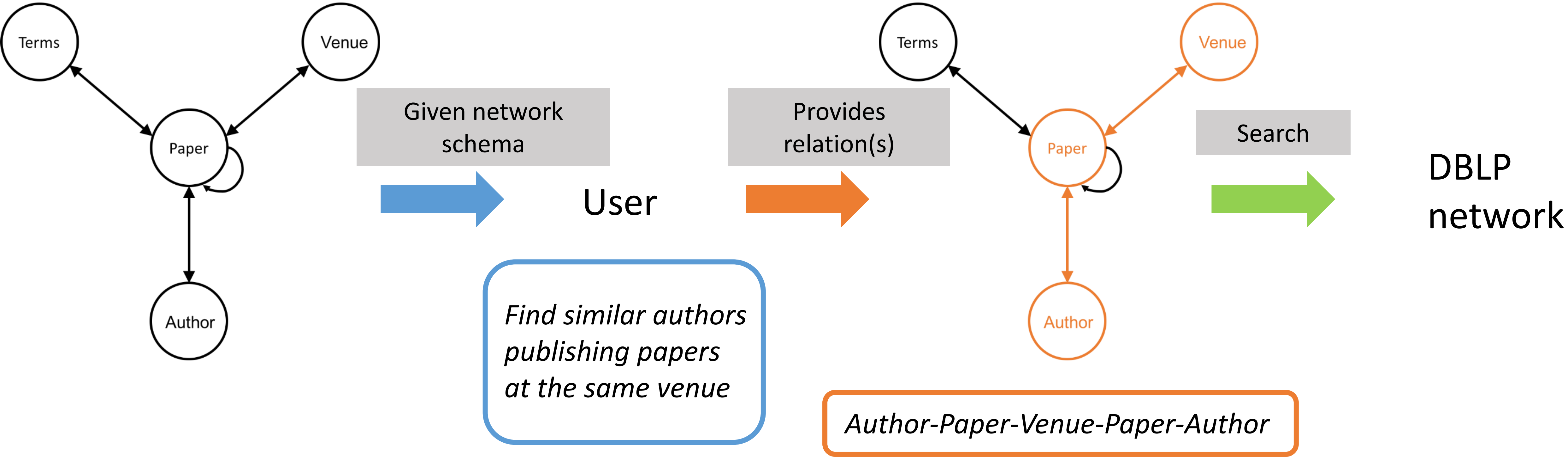


Meta-path

Author-Paper-Venue-Paper-Author

Schema-Simple vs. Schema-Rich Heterogeneous Information Networks

- Previous studies: *Schema-simple HINs*
  - Similarity search in DBLP network: <u>four entity types</u> (Paper, Author, Venue, Term), and <u>several relation types</u>; easy to search: user provide relation(s)



Schema-Simple vs. Schema-Rich Heterogeneous Information Networks

- In real world: *Schema-rich HINs*
  - Similarity search in Freebase network: <u>1,500+ entity types</u> and <u>35,000+ relation types</u>; hard to search: user CANNOT provide relation(s)

Given COMPLEX network schema → User → CANNOT provides relation(s) → ? → Search → Freebase network

Find similar person serving the same party

?

Freebase

DBpedia

Yago

Schema-Simple vs. Schema-Rich Heterogeneous Information Networks

- In real world: *Schema-rich HINs*
  - Similarity search in Freebase network: <u>1,500+ entity types</u> and <u>35,000+ relation types</u>; hard to search: user CANNOT provide relation(s)

Given COMPLEX network schema

User

CANNOT provides relation(s)

?

Search

Freebase network

Yago

Relation Similarity Search Problem

Relation Similarity Search

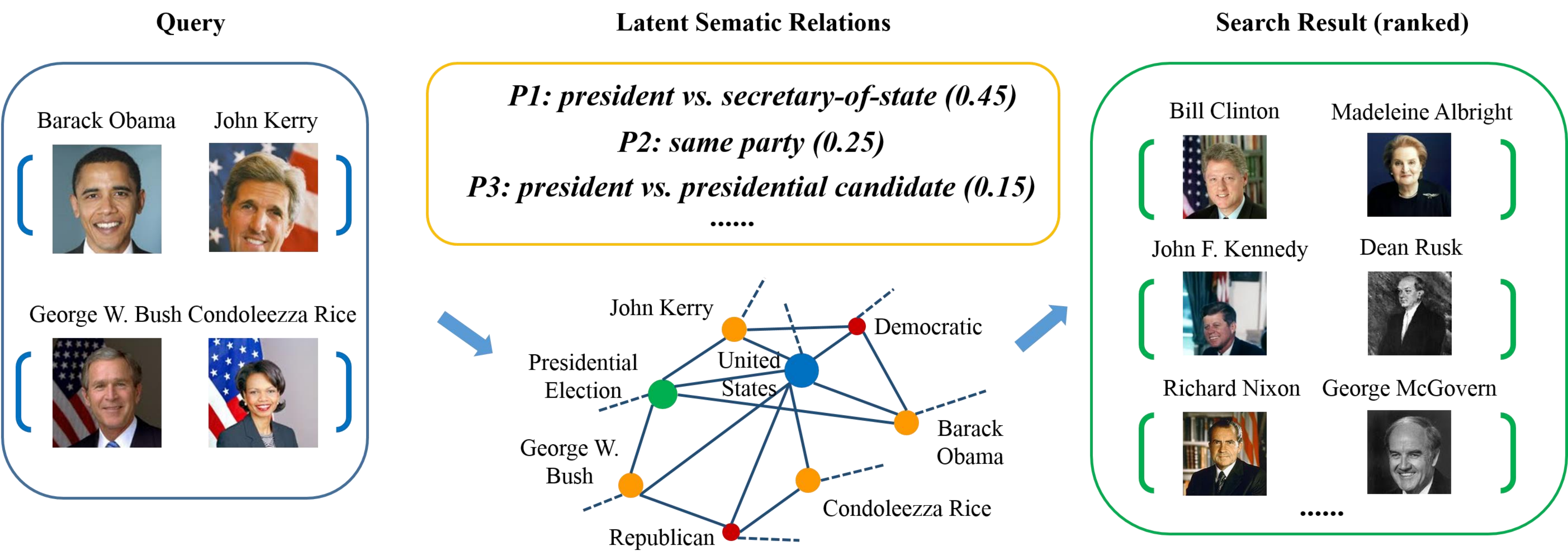
User

Freebase network

Yago

1. Users are asked to just provide a set of simple examples
2. We automatically detect the latent semantic relation (LSR) in the query for the users

Relation Similarity Search Example



Challenges

Q = *{< Barack Obama, John Kerry>, <George W. Bush, Condoleezza Rice>}*

president vs. secretary-of-state (0.45)

is president of *is secretary of state of*

President → Country ← Secretary of State

president vs. presidential candidate (0.15)

is president of *is presidential candidate of*

President → Country ← Presidential Candidate

<Bill Clinton, Madeleine Albright>

- **Q. how to measure the similarity between relation instances by distinguishing diverse latent semantic relation(s)?**

RelSim: A Relation Similarity Measure

RelSim: a meta-path-based relation similarity measure.
Given an LSR $\{w_m, P_m\}_{m=1}^{M}$, RelSim between r and r′ is defined as

Semantic overlap: the weighted number of overlapped meta-path based relations between two instances

$$RS(\mathrm{r},\mathrm{r}') = \frac{2 \times \boxed{\sum_m w_m \min(x_m, x'_m)}}{\boxed{\sum_m w_m x_m + \sum_m w_m x'_m}}$$

Semantic overlap: the weighted number of total meta-path-based relations satisfied by two instances

Intuition: <u>two relation instances are more similar when sharing more important (heavily weighted) meta-paths</u>

Properties: Range, Symmetric, Self-maximum

Latent Semantic Relation Learning

Number of meta-paths could be very large

$$RS(\mathrm{r},\mathrm{r}') = \frac{2 \times \sum_{\boxed{m}} w_m \min(x_m, \ x'_m)}{\sum_m \boxed{w_m} x_m + \sum_m w_m x'_m}$$

The weight/importance of each meta-path is different when query is different

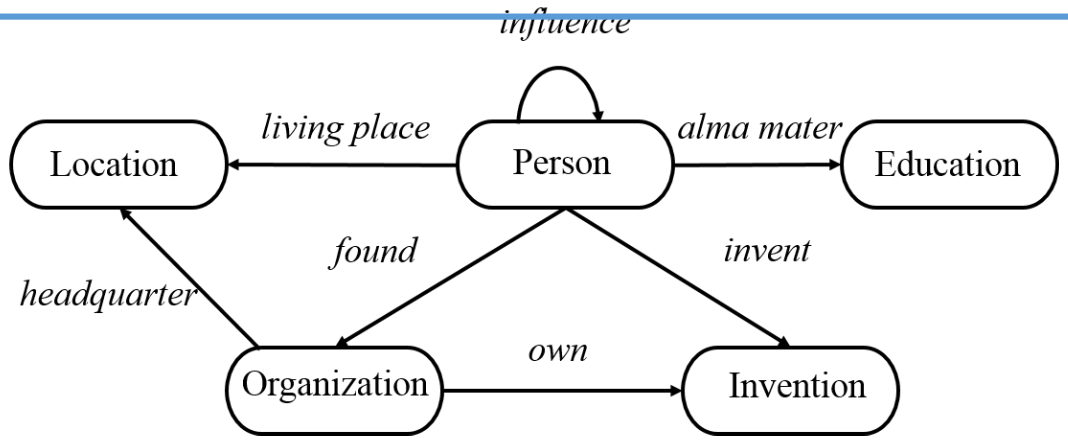
1. Meta-path candidates generation: enumerating all the possible meta-paths between entities in large-scale networks is impractical;
2. Meta-path weights optimization: the real semantic meaning in a query is specific.

Meta-Path Candidates Generation

Query based network schema: a sub-network schema of a schema-rich HIN that only contains the entity and relation types that relevant to the query.

Freebase

1,500+ entity types
35,000+ relation types



Query based meta-path generation algorithm: using binary search based on the query based network schema.

Meta-Path Weights Optimization

Intuition: Discover important query-based meta-paths by optimizing the weights.

e.g. <Larry Page, Sergey Brin> and <Jerry Yang, David Filo> share,

$\mathrm{PER} \xrightarrow{\text{alma mater}} \mathrm{EDU} \xleftarrow{\text{alma mater}} \mathrm{PER}$ $\quad$ $\mathrm{PER} \xrightarrow{\text{invest}} \mathrm{ORG} \xleftarrow{\text{employee}} \mathrm{PER}$

the later is a less important one (satisfy with randomly choosing instances).

Negative sample generation: since there is a lot of background noise. Randomly replacing the subject(object) entity of one instance by the subject(object) entity of another. e.g. <Larry Page, Paul Allen>

Meta-Path Weights Optimization

Inspired by the ranking loss, we propose the optimization model:

$$\min \sum_{k=1}^{K} max\{0, \boxed{c} - \boxed{\omega^T x_k + \omega^T \widehat{x_k}}\}$$

$$\text{s.t. } \omega_m \geq 0 \ \forall \mathrm{m} = 1, \ldots, \mathrm{M}$$

$$\sum_{m=1}^{M} \omega_m = 1$$

If c < 1 , consider the accident that positive and negative examples share the important meta-paths

maximize the weights of meta-paths that have the biggest difference between positive and negative examples

By introducing slack variables, the above optimization problem is turned into a linear programming with (M + K) variables and (M + 1 + 2K) constraints, solved by interior point method:

$$\min_{\omega,\alpha} \sum_{k=1}^{K} \alpha_k$$

$$\text{s.t.} \quad \omega_m \geq 0 \quad \forall m = 1, \ldots, M \quad \sum_{\mathrm{m}=1}^{M} \omega_m = 1$$

$$\alpha_k \geq 0 \quad \alpha_k \geq c - \omega^T x_k + \omega^T \tilde{x}_k \quad \forall k = 1, \ldots, K$$

Experiments

- Datasets: five real world datasets are constructed based on Freebase
  - The largest one is **Rel-Full** dataset: five popular relation categories in Freebase are selected,
  - For each relation category, randomly sample 5,000 entity pairs, then enumerate all the neighbor entities and relations within 2-hop of each entity.

Relation Categories	#Entities	#Relations	Examples
⟨*Organization, Founder*⟩	9,836,649	560,688,893	⟨Google, Larry Page⟩, ⟨Microsoft, Bill Gates⟩, ⟨Facebook, Mark Zuckerberg⟩
⟨*Book, Author*⟩	16,640,478	981,788,232	⟨Gone with the Wind, Margaret Mitchell⟩, ⟨The Kite Runner, Khaled Hosseini⟩
⟨*Actor, Film*⟩	4,340,986	182,121,412	⟨Leonardo DiCaprio, Inception⟩, ⟨Daniel Radcliffe, Harry Potter⟩, ⟨Jack Nicholson, Head⟩
⟨*Location, Contains*⟩	1,037,791	62,229,669	⟨United States of America, New York⟩, ⟨Victoria, Chillingollah⟩, ⟨New Mexico, Davis House⟩
⟨*Music, Track*⟩	1,653,931	86,658,343	⟨My Worlds, Baby⟩, ⟨21, Someone Like You⟩, ⟨Thriller, Beat It⟩
Total	26,841,657	1,483,834,223	⟨Google, Larry Page⟩, ⟨Leonardo DiCaprio, Inception⟩, ⟨Thriller, Beat It⟩

Similarity Search Performance

Performance (NDCG@K) of relation similarity search on Rel-Full.

	NDCG@5	NDCG@10	NDCG@20
VSM-S	0.5389	0.6296	0.7225
LRA-S	0.5880	0.6848	0.7814
IW-S	0.5210	0.6095	0.7010
RelSim-S	0.6395	0.7427	0.8432
RelSim-WS	**0.6651**	**0.7716**	**0.9559**

Finding #1: Our methods outperform the other methods in a significant way using t-test with p-value < 0.001;
Finding #2: RelSim-WS can better use the semantics in schema-rich HINs because it automatically learns the weights of different meta-paths;
Finding #3: Both RelSim-WS and RelSim-S consider more subtle semantics by incorporating the number of shared meta-paths of two relation instances.

Case Study of Meta-Paths

Example query-based meta-paths on Rel-Full. We show the most important four query-based meta-paths of different queries.

Query: $\{\langle$Google, Larry Page$\rangle$, $\langle$Microsoft, Bill Gates$\rangle$, *etc.*$\}$	ω
$\boxed{Organization \xrightarrow{\text{is founded by}} Founder}$	0.384
$Organization \xrightarrow{\text{run business in}} Industry \xrightarrow{\text{win award in}^{-1}} Founder$	0.274
$Organization \xrightarrow{\text{is founded by}} Person \xrightarrow{\text{is influence peer}^{-1}} Founder$	0.174
$Organization \xrightarrow{\text{'s leadership}} Person \xrightarrow{\text{mailing address}} Location \xrightarrow{\text{mailing address}^{-1}} Founder$	0.115
Query: $\{\langle$Google, Larry Page$\rangle$, $\langle$Yahoo!, Marissa Mayer$\rangle$, *etc.*$\}$	ω
$\boxed{Organization \xrightarrow{\text{run by}} CEO \xrightarrow{\text{job title}} Founder}$	0.32
$Organization \xrightarrow{\text{founded date}} Date \xrightarrow{\text{graduation date}^{-1}} Founder$	0.229
$Organization \xrightarrow{\text{headquarter}} Location \xrightarrow{\text{education institute}} Founder$	0.207
$Organization \xrightarrow{\text{run business in}} Industry \xrightarrow{\text{win award in}^{-1}} Founder$	0.113

Finding: Optimization model is able to distinguish the diverse LSRs.

Conclusion

Problem	Relation similarity search in schema-rich heterogeneous information networks.
Approach	RelSim, to compute the semantic similarity between relation instances.
Results	Our method performs the best on all the datasets.

Thank You! ☺