



南京大學
NANJING UNIVERSITY



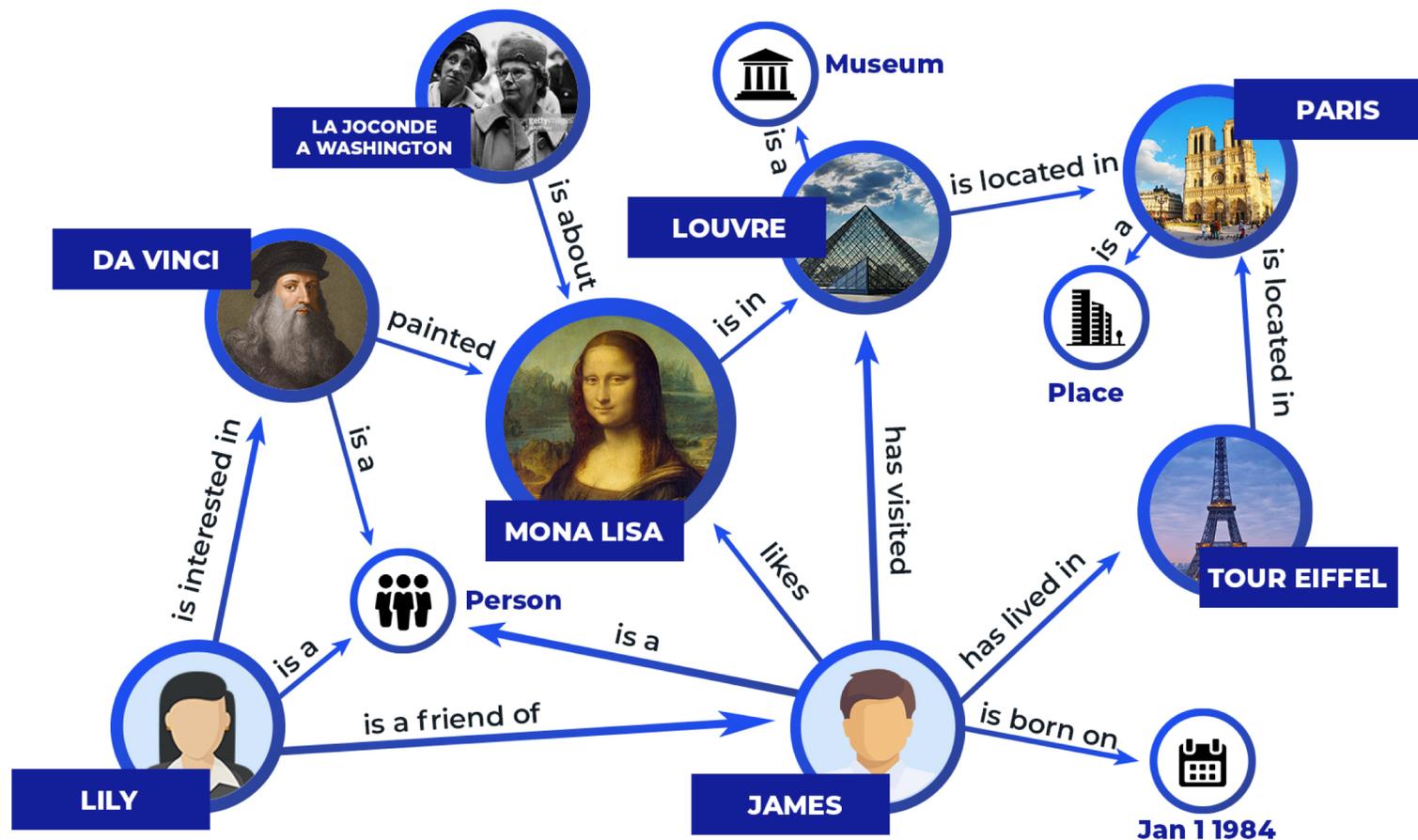
知识图谱中的关联搜索

程龚（南京大学 计算机科学与技术系 副教授）

第3届知识工程与问答技术研讨会，2019年12月14日，南京

提纲

- 关联实体搜索
 - 元路径
 - 生成模型
- 实体关联搜索
 - 子图发现
 - 子图排序



提纲

- 关联实体搜索
 - 元路径
 - 生成模型
- 实体关联搜索
 - 发现
 - 排序

实体搜索（相关实体推荐）

- 查询实体：Tom Hardy

- 两种相关性的样例：

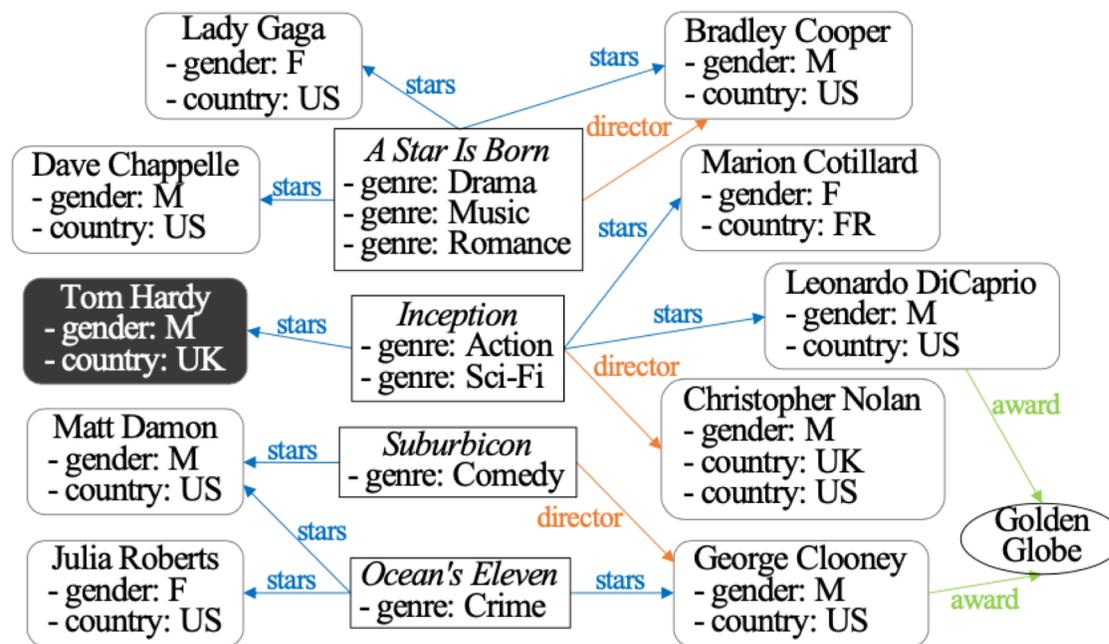
$S_1 = \{ \langle \text{Dave Chappelle, Lady Gaga} \rangle, \langle \text{Matt Damon, Julia Roberts} \rangle \},$

$S_2 = \{ \langle \text{Dave Chappelle, Bradley Cooper} \rangle, \langle \text{Matt Damon, George Clooney} \rangle \}.$

- 对应的元路径（meta-path）

$\mathcal{P}_1 : [\text{query}] \xleftarrow{\text{stars}} \cdot \xrightarrow{\text{stars}} [\text{answer}]$

$\mathcal{P}_2 : [\text{query}] \xleftarrow{\text{stars}} \cdot \xrightarrow{\text{director}} [\text{answer}]$



实体搜索（相关实体推荐）

- 查询实体：Tom Hardy

- 两种相关性的样例：

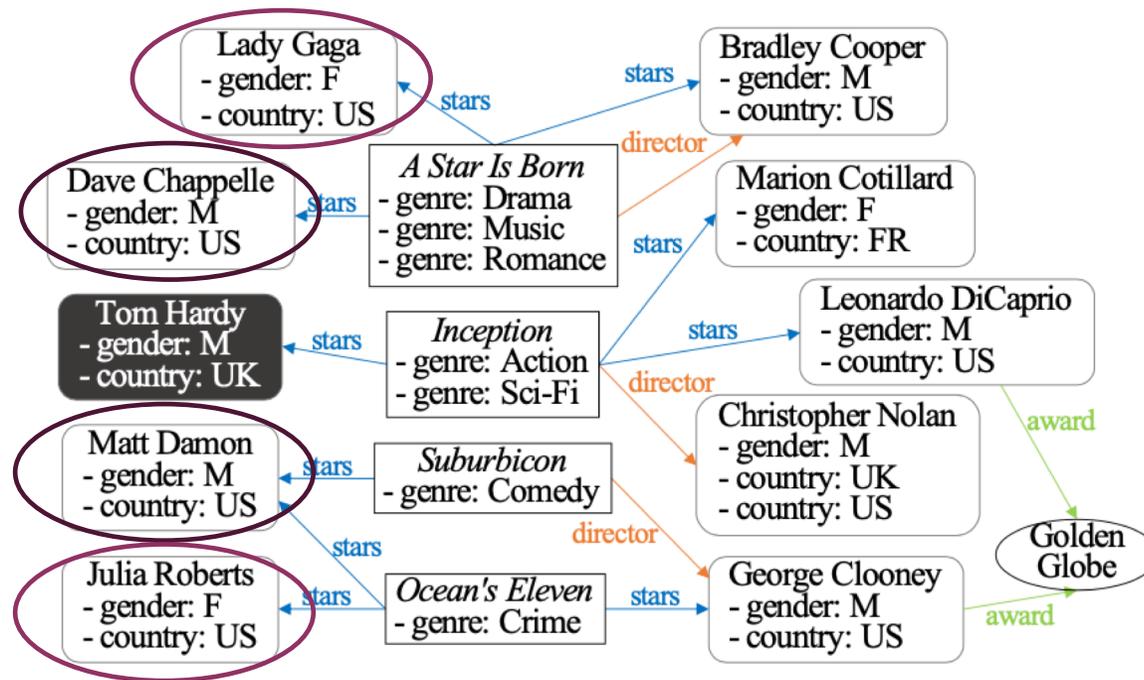
$S_1 = \{ \langle \text{Dave Chappelle}, \text{Lady Gaga} \rangle, \langle \text{Matt Damon}, \text{Julia Roberts} \rangle \},$

$S_2 = \{ \langle \text{Dave Chappelle}, \text{Bradley Cooper} \rangle, \langle \text{Matt Damon}, \text{George Clooney} \rangle \}.$

- 对应的元路径（meta-path）

$\mathcal{P}_1 : [\text{query}] \xleftarrow{\text{stars}} \cdot \xrightarrow{\text{stars}} [\text{answer}]$

$\mathcal{P}_2 : [\text{query}] \xleftarrow{\text{stars}} \cdot \xrightarrow{\text{director}} [\text{answer}]$



实体搜索（相关实体推荐）

- 查询实体：Tom Hardy

- 两种相关性的样例：

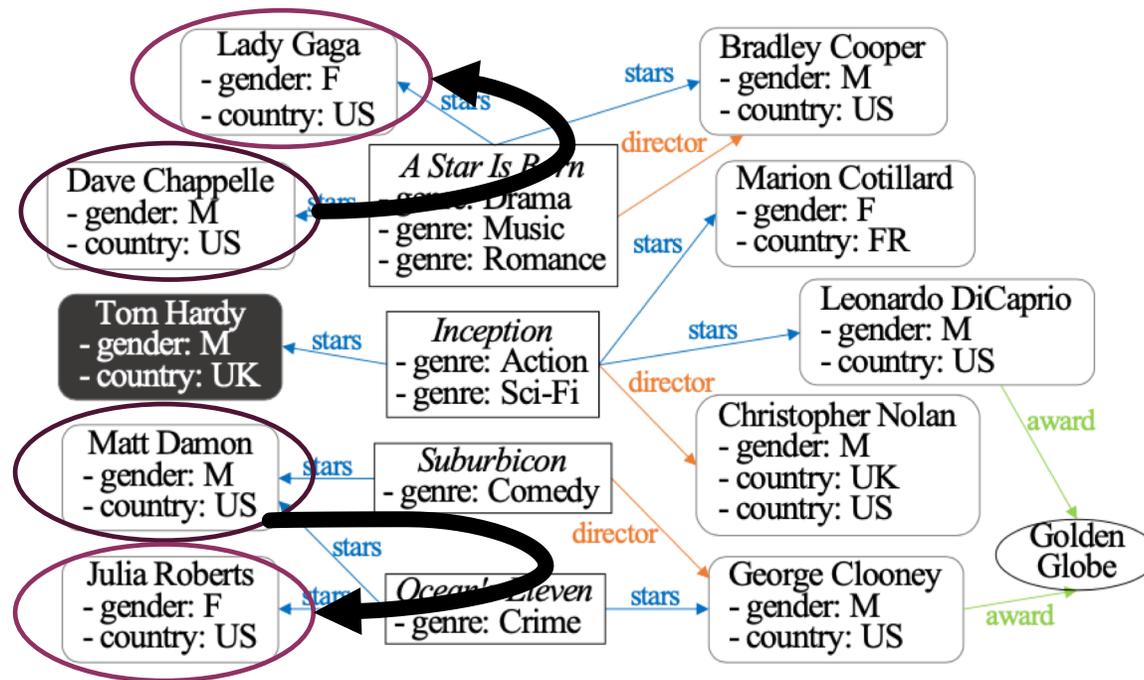
$S_1 = \{ \langle \text{Dave Chappelle, Lady Gaga} \rangle, \langle \text{Matt Damon, Julia Roberts} \rangle \},$

$S_2 = \{ \langle \text{Dave Chappelle, Bradley Cooper} \rangle, \langle \text{Matt Damon, George Clooney} \rangle \}.$

- 对应的元路径（meta-path）

$\mathcal{P}_1 : [\text{query}] \xleftarrow{\text{stars}} \cdot \xrightarrow{\text{stars}} [\text{answer}]$

$\mathcal{P}_2 : [\text{query}] \xleftarrow{\text{stars}} \cdot \xrightarrow{\text{director}} [\text{answer}]$



实体搜索（相关实体推荐）

- 查询实体：Tom Hardy

- 两种相关性的样例：

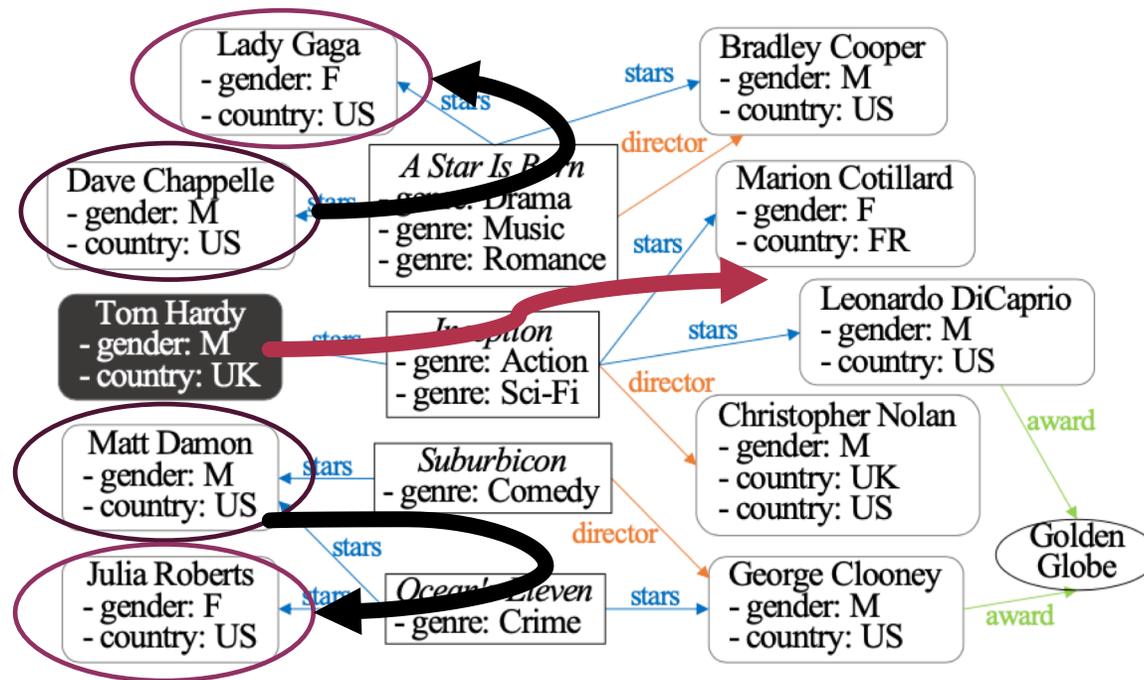
$S_1 = \{ \langle \text{Dave Chappelle, Lady Gaga} \rangle, \langle \text{Matt Damon, Julia Roberts} \rangle \},$

$S_2 = \{ \langle \text{Dave Chappelle, Bradley Cooper} \rangle, \langle \text{Matt Damon, George Clooney} \rangle \}.$

- 对应的元路径（meta-path）

$\mathcal{P}_1 : [\text{query}] \xleftarrow{\text{stars}} \cdot \xrightarrow{\text{stars}} [\text{answer}]$

$\mathcal{P}_2 : [\text{query}] \xleftarrow{\text{stars}} \cdot \xrightarrow{\text{director}} [\text{answer}]$



实体搜索（相关实体推荐）

- 查询实体：Tom Hardy

- 两种相关性的样例：

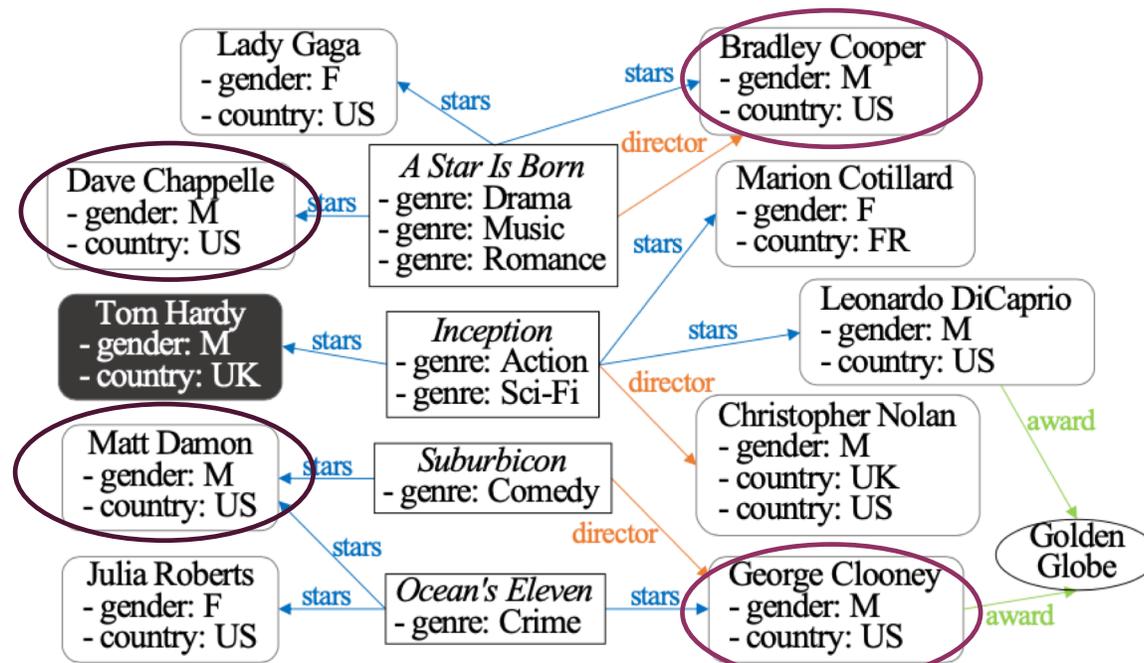
$S_1 = \{ \langle \text{Dave Chappelle, Lady Gaga} \rangle, \langle \text{Matt Damon, Julia Roberts} \rangle \},$

$S_2 = \{ \langle \text{Dave Chappelle, Bradley Cooper} \rangle, \langle \text{Matt Damon, George Clooney} \rangle \}.$

- 对应的元路径（meta-path）

$\mathcal{P}_1 : [\text{query}] \xleftarrow{\text{stars}} \cdot \xrightarrow{\text{stars}} [\text{answer}]$

$\mathcal{P}_2 : [\text{query}] \xleftarrow{\text{stars}} \cdot \xrightarrow{\text{director}} [\text{answer}]$



实体搜索（相关实体推荐）

- 查询实体：Tom Hardy

- 两种相关性的样例：

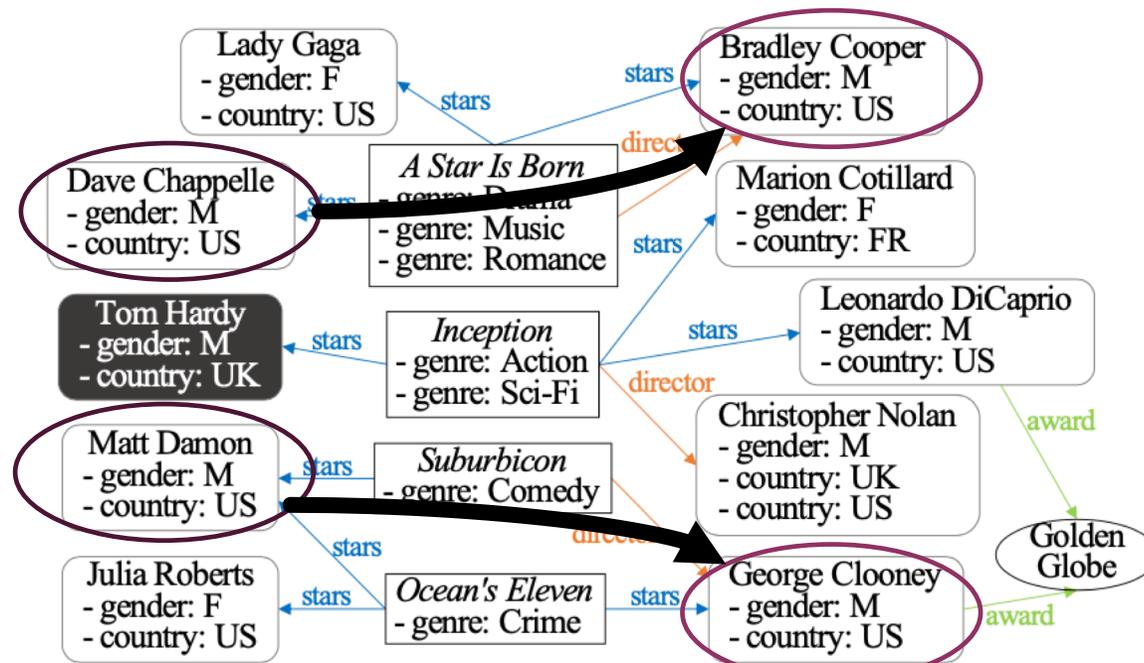
$S_1 = \{ \langle \text{Dave Chappelle, Lady Gaga} \rangle, \langle \text{Matt Damon, Julia Roberts} \rangle \},$

$S_2 = \{ \langle \text{Dave Chappelle, Bradley Cooper} \rangle, \langle \text{Matt Damon, George Clooney} \rangle \}.$

- 对应的元路径（meta-path）

$\mathcal{P}_1 : [\text{query}] \xleftarrow{\text{stars}} \cdot \xrightarrow{\text{stars}} [\text{answer}]$

$\mathcal{P}_2 : [\text{query}] \xleftarrow{\text{stars}} \cdot \xrightarrow{\text{director}} [\text{answer}]$



实体搜索（相关实体推荐）

- 查询实体：Tom Hardy

- 两种相关性的样例：

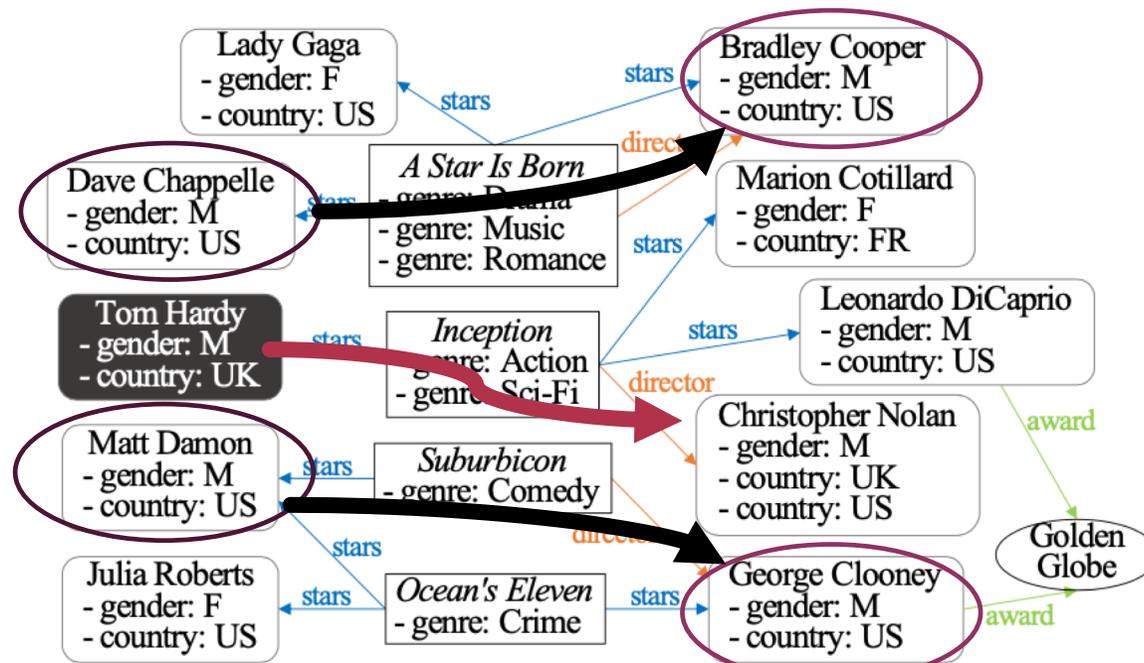
$S_1 = \{ \langle \text{Dave Chappelle, Lady Gaga} \rangle, \langle \text{Matt Damon, Julia Roberts} \rangle \},$

$S_2 = \{ \langle \text{Dave Chappelle, Bradley Cooper} \rangle, \langle \text{Matt Damon, George Clooney} \rangle \}.$

- 对应的元路径（meta-path）

$\mathcal{P}_1 : [\text{query}] \xleftarrow{\text{stars}} \cdot \xrightarrow{\text{stars}} [\text{answer}]$

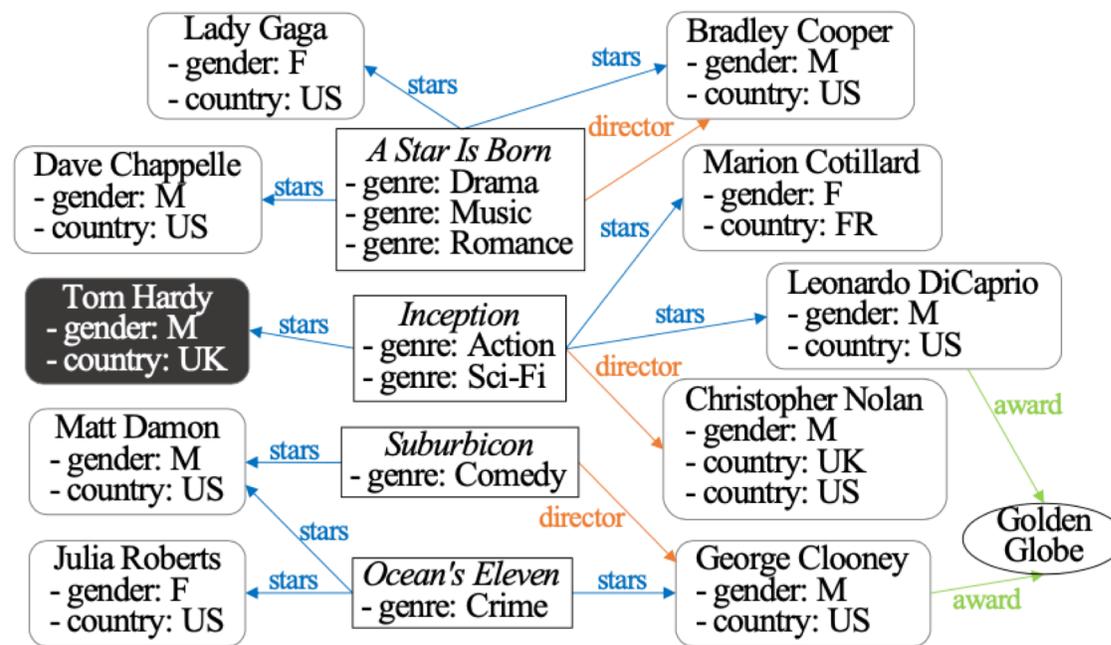
$\mathcal{P}_2 : [\text{query}] \xleftarrow{\text{stars}} \cdot \xrightarrow{\text{director}} [\text{answer}]$



基于元路径的相关度

$$\text{rel}(q, v) = \sum_{i=1}^n w_i \cdot \gamma(q, v | \mathcal{P}_i)$$

- 元路径的选取 (\mathcal{P}_i)
- 基于特定元路径的相关度计算 (γ)
 - PathCount、PathSim、PCRW.....
- 元路径的加权 (w_i)



元路径的选取

- 元路径的重要性（对样例的拟合程度）

$$\text{sig}(\mathcal{P}|q, \Lambda) = \beta^l \frac{1}{|\Lambda| \cdot |\bar{\Lambda}|} \sum_{\lambda \in \Lambda} \sum_{v \in \bar{\Lambda}} \mathbb{I}(\gamma(q, \lambda|\mathcal{P}) - \gamma(q, v|\mathcal{P}))$$

$$\mathbb{I}(\gamma(q, \lambda|\mathcal{P}) - \gamma(q, v|\mathcal{P})) = \begin{cases} 1 & \text{if } \gamma(q, \lambda|\mathcal{P}) > \gamma(q, v|\mathcal{P}) \\ 0 & \text{if } \gamma(q, \lambda|\mathcal{P}) \leq \gamma(q, v|\mathcal{P}) \end{cases}$$

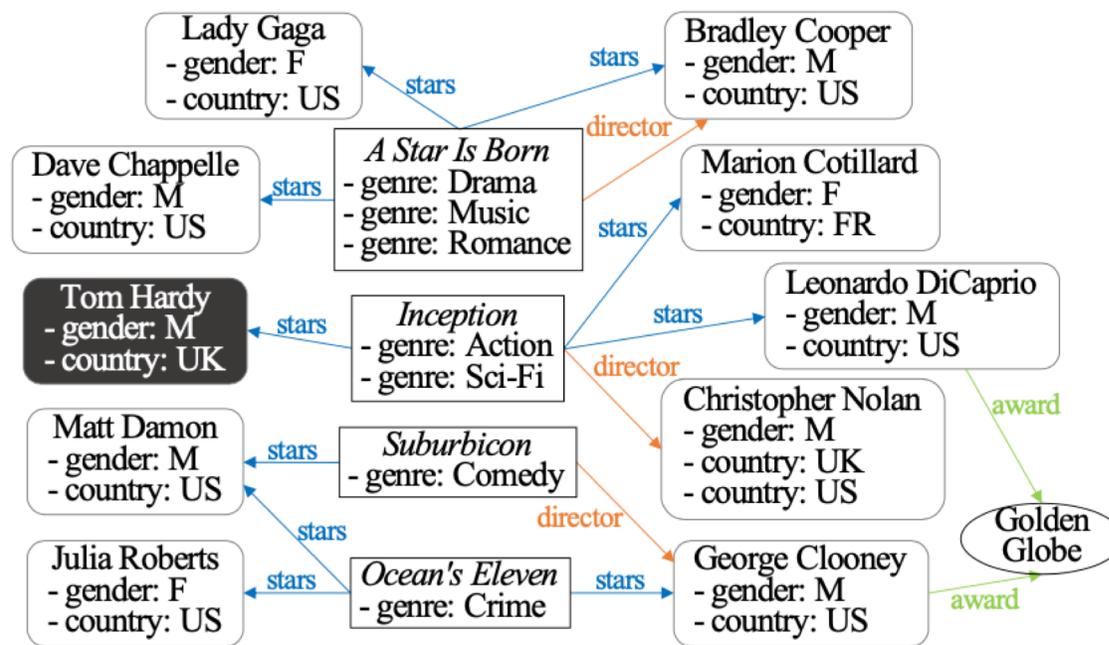
- 元路径的多样性（去除冗余的元路径）

$$\text{to maximize } \sum_{i=1}^m y_i \cdot \text{sig}(\mathcal{P}_i|q, \Lambda),$$

$$\text{subject to } \sum_{i=1}^m y_i \leq n,$$

$$y_i + y_j \leq 1 \text{ for all } i \neq j \text{ and } \mathcal{P}_i \sim_q \mathcal{P}_j,$$

$$y_i \in \{0, 1\} \text{ for all } 1 \leq i \leq m,$$



元路径的选取

元路径的启发式搜索

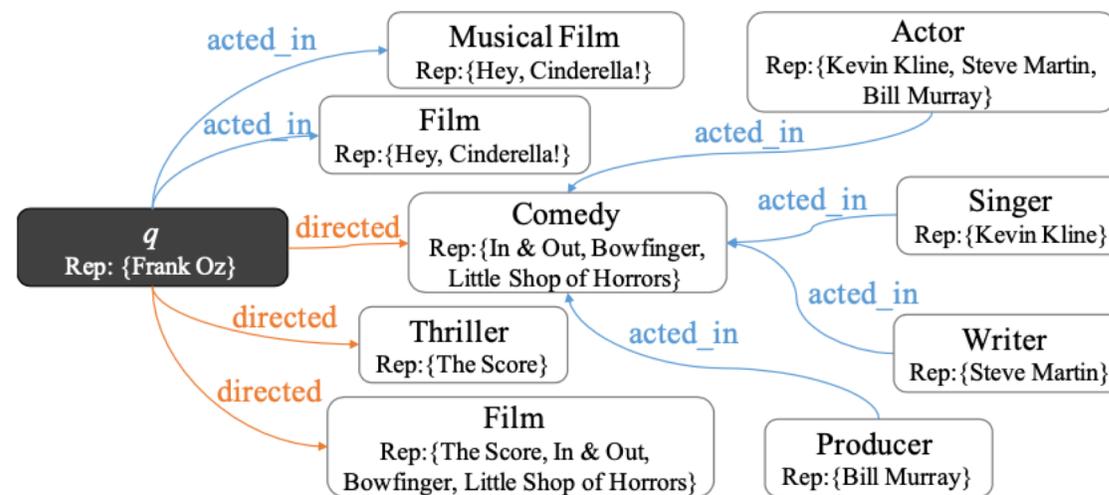
Algorithm 2: Heuristic meta-path search.

Input: A KG G , a query entity q , a set of user-provided examples Λ , a number n of meta-paths to select, and a significance threshold τ .

Output: A set of meta-paths MP .

```

1 Initialize a search tree  $\mathcal{ST}$  with  $q$  as the root;
2  $MP \leftarrow \emptyset$ ;
3 while  $\mathcal{ST}$  can be expanded do
4    $stn \leftarrow \text{SelTN}(\mathcal{ST})$ ;
5    $TN \leftarrow \text{ExpST}(stn)$ ;
6   foreach  $tn \in TN$  do
7     EvalTN( $tn$ );
8     if  $tn$  contains entities in  $\Lambda$  then
9        $\mathcal{P}_i = \text{GetMP}(tn)$ ;
10      if  $\text{sig}(\mathcal{P}_i|q, \Lambda) \geq \tau$  then
11        if  $\mathcal{P}_i \sim_q \mathcal{P}_j$  for any  $\mathcal{P}_j \in MP$  then
12          Skip  $tn$  and continue with the next iteration;
13        else
14           $MP \leftarrow MP \cup \{\mathcal{P}_i\}$ ;
15          if  $|MP| = n$  then
16            Terminate the loop;
17 return  $MP$ ;
  
```



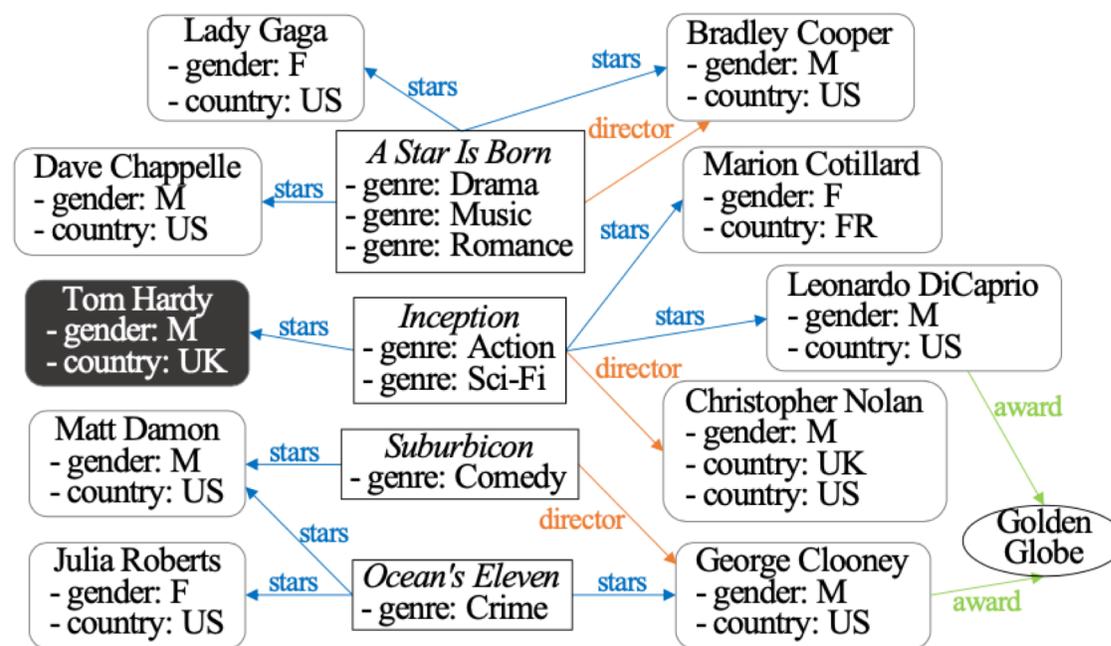
启发函数:

- 路径短
$$\text{prms}_{\text{dist}}(tn) = \text{dpth}(tn) + \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} \min_{v \in \text{Rep}(tn) \setminus \{\lambda\}} \text{dist}(v, \lambda)$$
- 度数小
$$\text{prms}_{\text{deg}}(tn) = \sum_{v \in \text{Rep}(tn)} \text{deg}(v)$$

元路径的加权

$$\text{rel}(q, v) = \sum_{i=1}^n w_i \cdot \gamma(q, v | \mathcal{P}_i)$$

- 模型选择
 - Linear soft-margin support vector machine
 - 严惩false negative (充分尊重用户给定样例)
- 候选与负例生成
 - 元路径可达的非样例实体



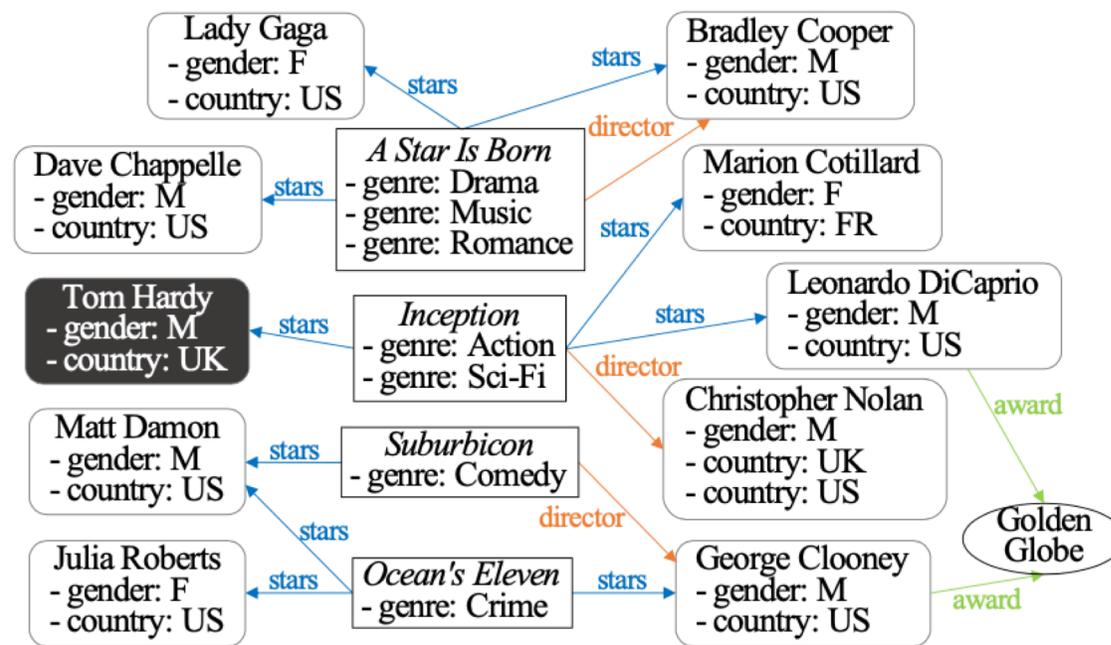
提纲

- 关联实体搜索
 - 元路径
 - 生成模型
- 实体关联搜索
 - 发现
 - 排序

基于生成模型的相关度

$$\text{rel}(q, v|S) = \sum_{\mathcal{P}_i \in \Omega_{\text{mp}}} \gamma(q, v|\mathcal{P}_i) \cdot \Pr(\mathcal{P}_i|S) \cdot J(\mathcal{P}_i)$$

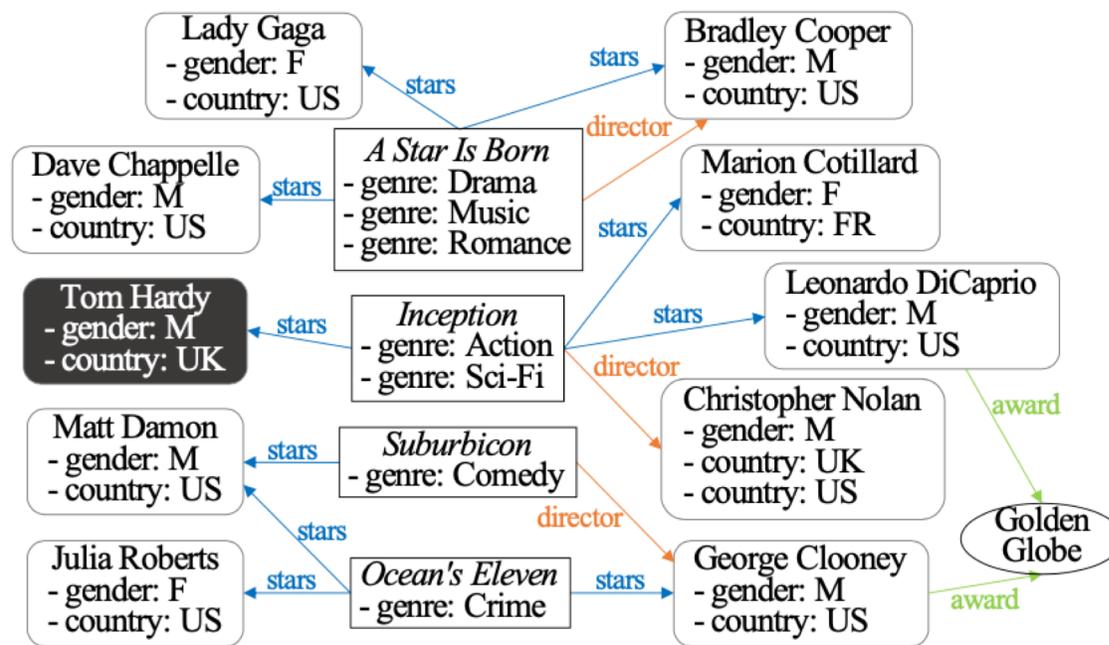
- 元路径的选取 (\mathcal{P}_i)
 - $\Omega_{\text{mp}} = \bigcup_{\langle s, t \rangle \in S} \{\mathcal{P} : \exists p \models \mathcal{P}, s \rightsquigarrow_p t\}$
- 基于特定元路径的相关度计算 (γ)
 - PathCount、PathSim、PCRW.....
- 元路径的加权: $\Pr(\mathcal{P}_i|S)$ 视作后验概率
- 正则化项: $J(\mathcal{P}_i)$ 惩罚长路径
 - $J(\mathcal{P}_i) = e^{-\beta \cdot \text{len}(\mathcal{P}_i)}$



生成模型

$$\Pr(\mathcal{P}_i|S) = \frac{\Pr(\mathcal{P}_i) \cdot \Pr(S|\mathcal{P}_i)}{\Pr(S)} \propto \Pr(\mathcal{P}_i) \cdot \Pr(S|\mathcal{P}_i)$$

- $\Pr(\mathcal{P}_i)$: 先验
- $\Pr(S|\mathcal{P}_i)$: 似然



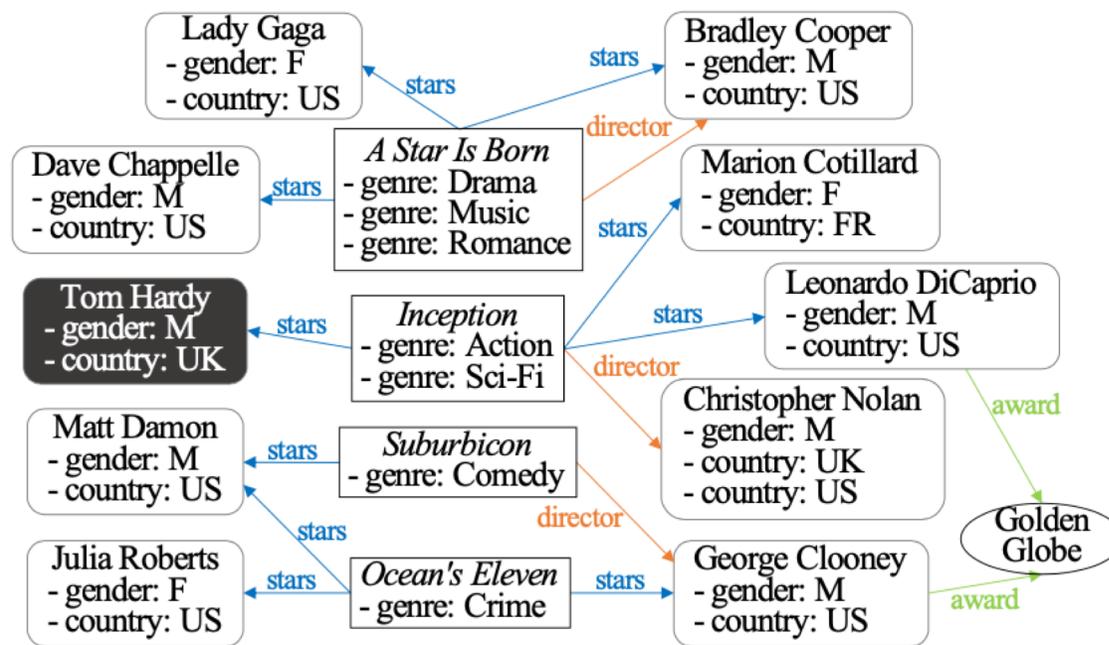
先验的计算

$$\Pr(\mathcal{P}_i) = \Pr(r_1 r_2 \cdots r_l) = \Pr(r_1) \prod_{i=2}^l \Pr(r_i | r_1 r_2 \cdots r_{i-1})$$

$$\approx \Pr(r_1) \prod_{i=2}^l \Pr(r_i | r_{i-1}),$$

$$\Pr(\mathcal{P}_i) \propto \text{pc}(r_1) \prod_{i=2}^l \frac{\text{pc}(r_{i-1} r_i)}{\text{pc}(r_{i-1})}$$

$$\Pr(\mathcal{P}_i) \propto \text{pc}(r_l) \prod_{i=1}^{l-1} \frac{\text{pc}(r_i r_{i+1})}{\text{pc}(r_{i+1})}$$



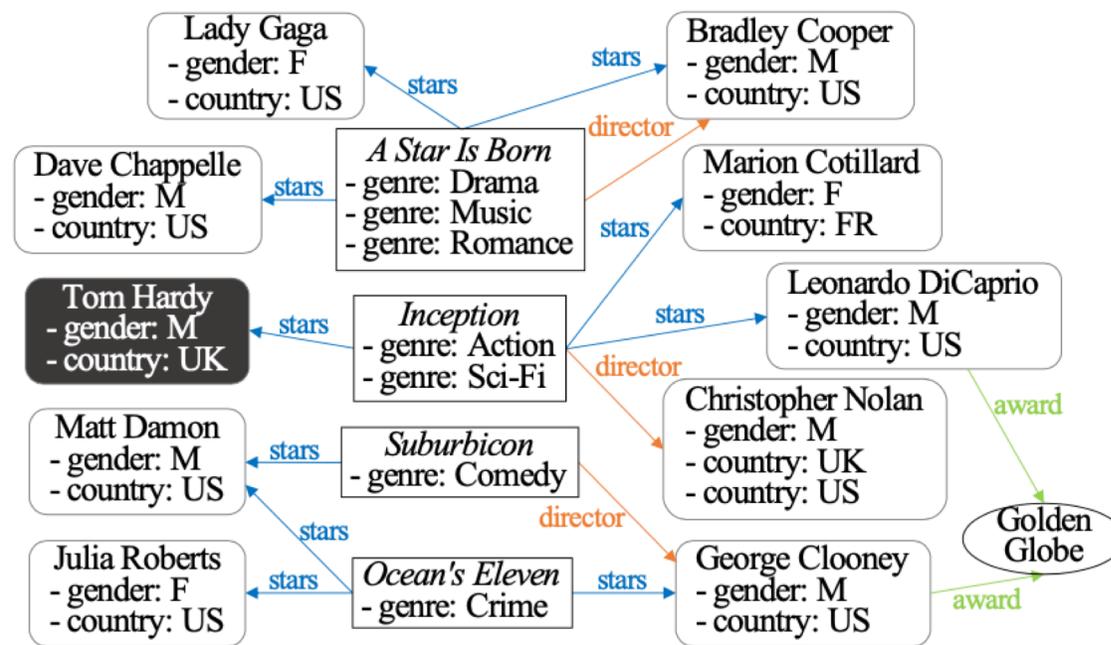
似然的计算

$$\Pr(S|\mathcal{P}_i) = \prod_{\langle s, t \rangle \in S} \Pr(\langle s, t \rangle | \mathcal{P}_i)$$

$$\Pr(\langle s, t \rangle | \mathcal{P}_i) \approx \frac{\text{pc}(s, t, \mathcal{P}_i)}{\text{apc}(\mathcal{P}_i)}$$

- $\text{pc}(s, t, \mathcal{P}_i)$ 的平滑项: $\frac{\text{apc}(\mathcal{P}_i)}{|\text{ST}(s)| \cdot |\text{ST}(t)|}$
 - ST: 同类型的所有实体
- $\text{apc}(\mathcal{P}_i)$: $\text{pc}(\mathcal{P}_i)$ 的近似 (当 \mathcal{P}_i 是长路径时)

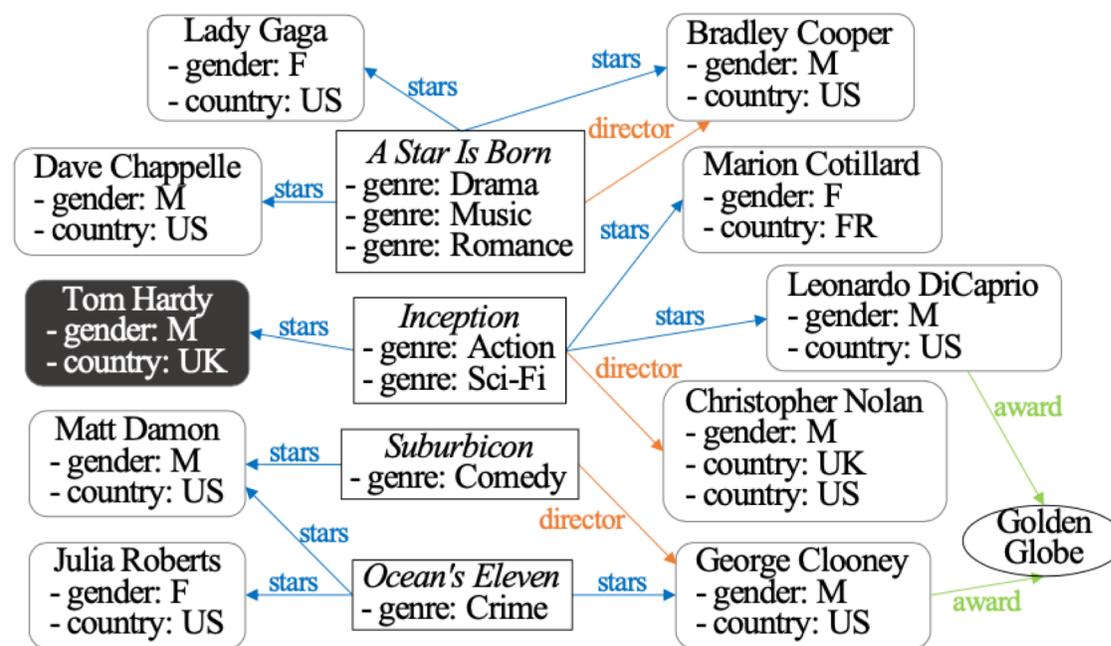
$$\text{pc}(r_l) \prod_{i=2}^l \frac{\text{pc}(r_{i-1}r_i)}{\text{pc}(r_{i-1})} \quad \overline{\text{pc}(r_l) \prod_{i=1}^{l-1} \frac{\text{pc}(r_i r_{i+1})}{\text{pc}(r_{i+1})}}$$



扩展的相关度

$$\text{rel}(q, v|S) = \sum_{\mathcal{P}_i \in \Omega_{\text{mp}}} \gamma(q, v|\mathcal{P}_i) \cdot \Pr(\mathcal{P}_i|S) \cdot J(\mathcal{P}_i) \\ + \sum_{\langle a_i, l_i \rangle \in \Omega_{\text{prop}}} \gamma(q, v|\langle a_i, l_i \rangle) \cdot \Pr(\langle a_i, l_i \rangle|S)$$

- 元路径关联 + 属性约束

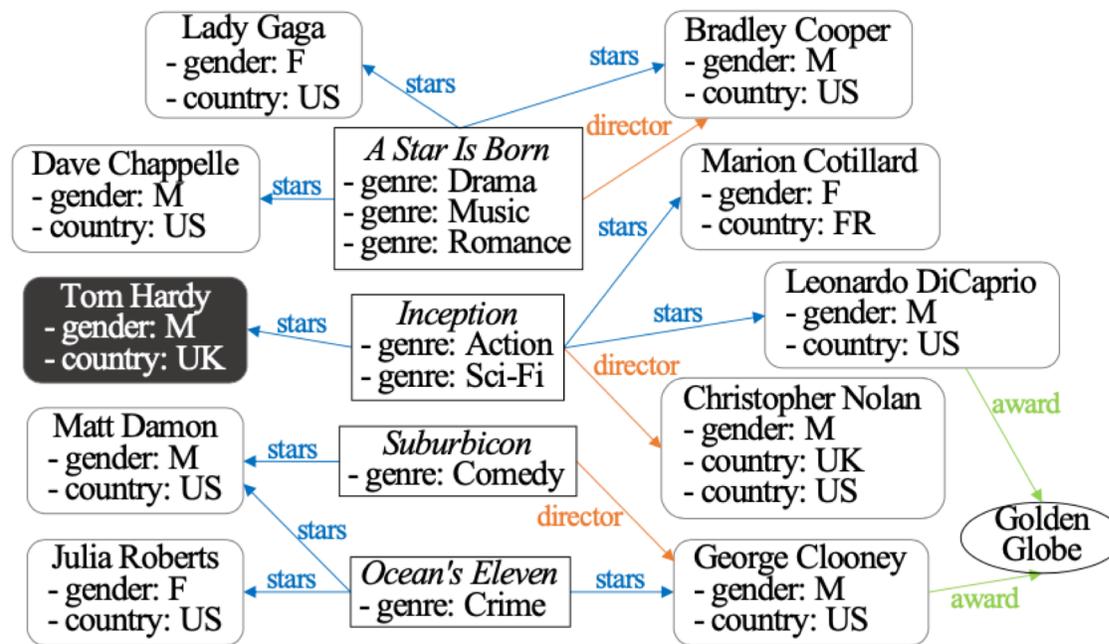


搜索算法

Input: A KG $G = \langle V, E, \Psi \rangle$, a query entity q , a set of user-provided examples S , an upper bound L on the length of allowable meta-paths, and a positive integer m .

Output: k top-ranked entities that are relevant to q .

- 1: $\Omega_{mp} \leftarrow \text{MPSearch}(G, S, L)$;
- 2: $\Omega_{prop} \leftarrow \bigcup_{\langle s, t \rangle \in S} \Phi(t)$;
- 3: $\Omega \leftarrow \Omega_{mp} \cup \Omega_{prop}$;
- 4: **for all** $\Omega_i \in \Omega$ **do**
- 5: Compute $\text{Pr}(\Omega_i | S)$;
- 6: **end for**
- 7: $\Omega_{top} \leftarrow m$ meta-paths in Ω_{mp} with the largest weights;
- 8: $C \leftarrow \bigcup_{\mathcal{P}_i \in \Omega_{top}} \{v \in V : \exists p \models \mathcal{P}_i, q \rightsquigarrow_p v\}$;
- 9: **for all** $v \in C$ **do**
- 10: Compute $\text{rel}(q, v | S)$;
- 11: **end for**
- 12: **return** k top-ranked entities in C

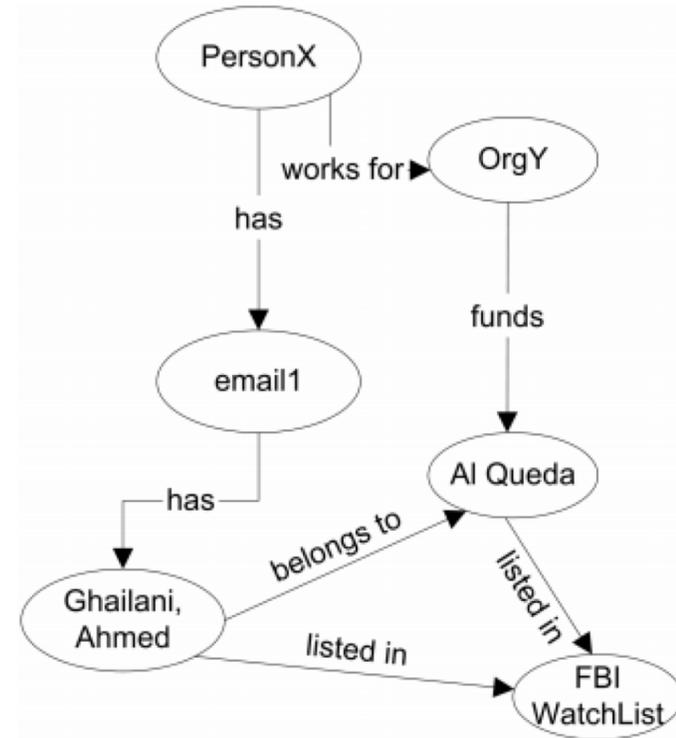


提纲

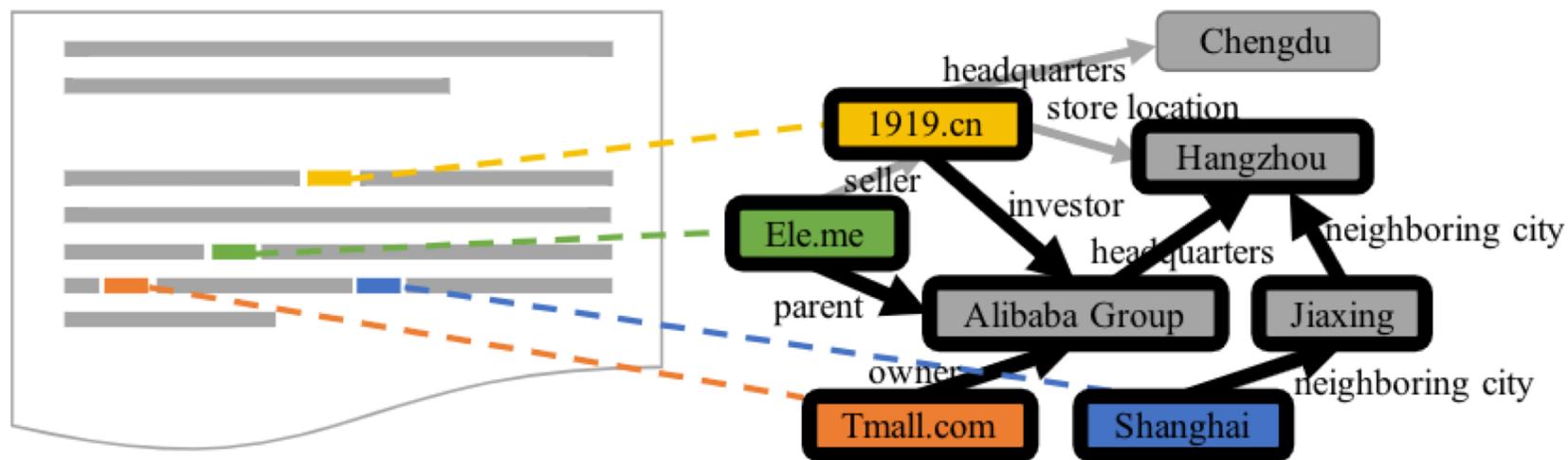
- 关联实体搜索
 - 元路径
 - 生成模型
- 实体关联搜索
 - 发现
 - 排序

实体关联搜索的应用：国家安全

1. Is the passenger known to be associated with an organization on the watch list?
2. Does the passenger work for an organization that is known to sponsor an organization on a watch-list?
3. Is there a connection between the passenger and one or more passengers on the same flight or different flights? Is such connection in the context of aviation safety?

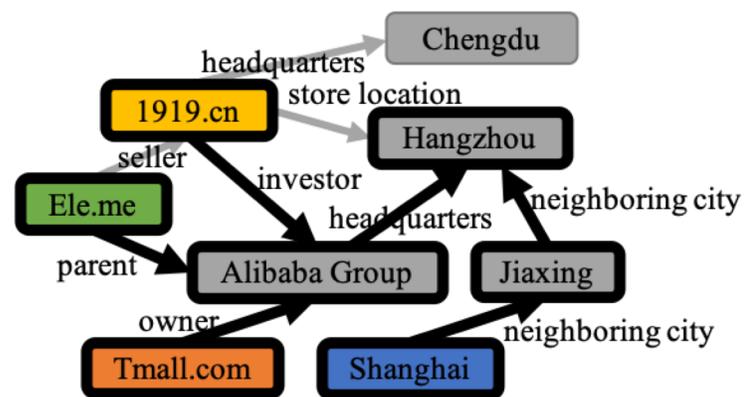


实体关联搜索的应用：新闻阅读

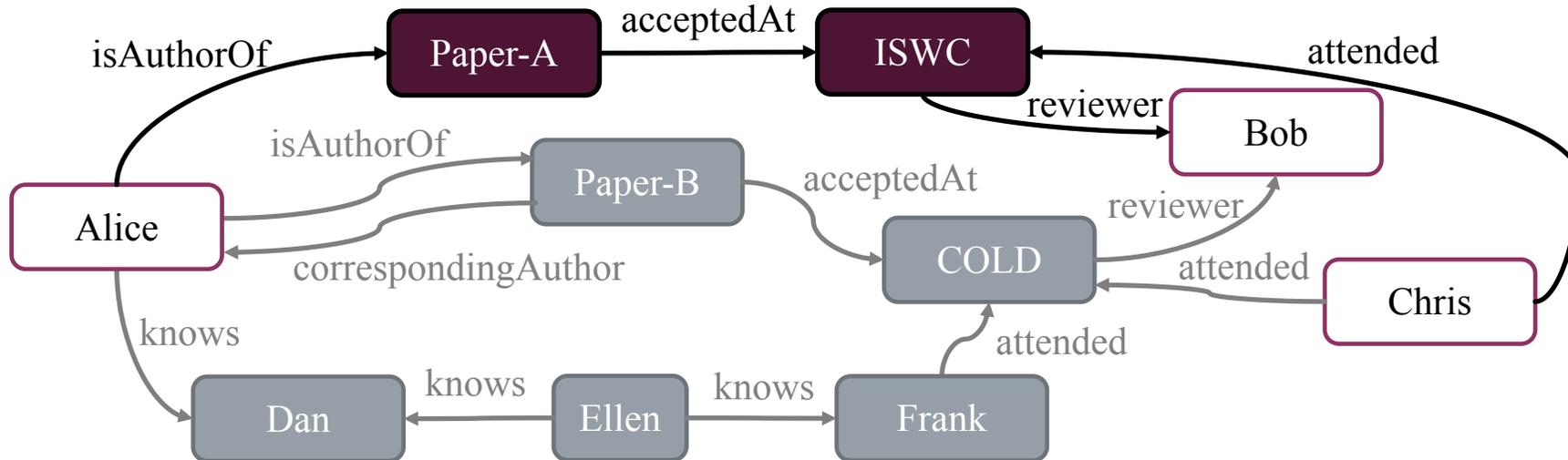


实体关联的定义

- 包含所有查询实体的极小连通子图
- 直径受限



实体关联的发现



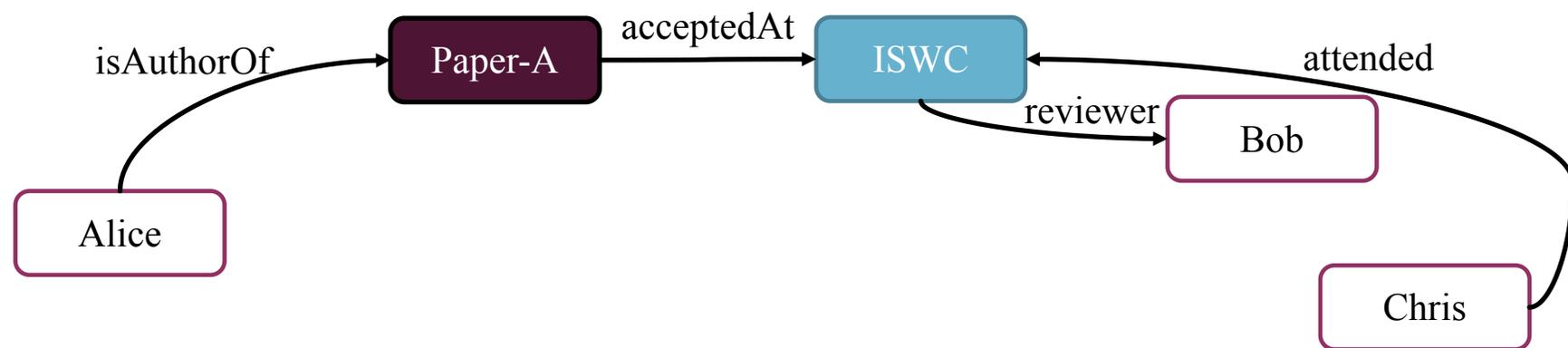
实体关联的发现：基本算法

■ 路径搜索 + 合并

Input: an entity-relation graph G , a query Q , and a diameter constraint D .

Output: all the SAs that are results of Q and have a diameter of D or less.

```
1:  $X \leftarrow$  empty set of SAs
2: for  $i \leftarrow 1$  to  $n$  do //  $n$ : number of query vertices
3:    $P_i \leftarrow$  BLPathEnum( $G, v_i^Q, D$ ) //  $v_i^Q$ :  $i$ -th query vertex
4: end for
5: for all  $\langle p_1, \dots, p_n \rangle \in (P_1 \times \dots \times P_n)$  do
6:   if  $p_1, \dots, p_n$  have a common end vertex then
7:      $x \leftarrow$  subgraph formed by joining  $p_1, \dots, p_n$ 
8:     if  $x$  is minimal and  $diam(x) \leq D$  then
9:        $X \leftarrow X \cup \{x\}$ 
10:    end if
11:  end if
12: end for
13: return  $X$ 
```



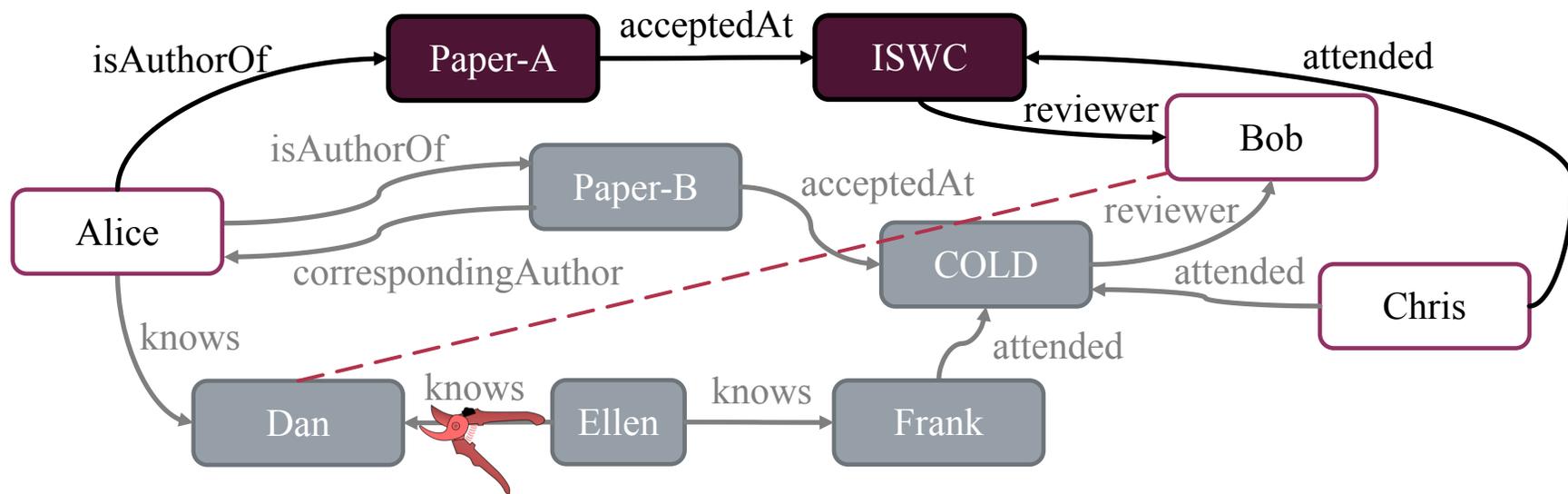
实体关联的发现：改进算法

■ 基于距离剪枝

Input: an entity-relation graph G , a query vertex $v_i^Q \in V_Q$, and a diameter constraint D .

Output: all the paths not longer than $\lfloor \frac{D+1}{2} \rfloor$ starting from v_i^Q in G .

```
1:  $P^0 \leftarrow \{v_i^Q\}$  //  $v_i^Q$ : treated as a path of length 0
2: for  $j \leftarrow 1$  to  $\lfloor \frac{D+1}{2} \rfloor$  do
3:    $P^j \leftarrow$  empty set of paths //  $P^j$ : paths of length  $j$ 
4:   for all  $p \in P^{j-1}$  do
5:      $pev \leftarrow$  end vertex of  $p$ 
6:     if  $\ln(p) + \text{dist}(pev, v_j^Q) \leq D$  for all  $v_j^Q \in (V_Q \setminus \{v_i^Q\})$ 
7:       then
8:          $P^j \leftarrow P^j \cup$  paths formed by joining  $p$  and each
9:         arc incident from/to  $pev$ 
10:      end if
11:   end for
12:    $P \leftarrow P^0 \cup \dots \cup P^{\lfloor \frac{D+1}{2} \rfloor}$ 
13: for all  $p \in P$  do
14:    $pev \leftarrow$  end vertex of  $p$ 
15:   if  $\text{dist}(pev, v_j^Q) > \lfloor \frac{D+1}{2} \rfloor$  for any  $v_j^Q \in (V_Q \setminus \{v_i^Q\})$ 
16:     then
17:        $P \leftarrow P \setminus \{p\}$ 
18:     end if
19: end for
20: return  $P$ 
```



• 长度(Alice→Dan) + 距离(Dan, Bob) > 直径约束

提纲

- 关联实体搜索
 - 元路径
 - 生成模型
- 实体关联搜索
 - 发现
 - 排序

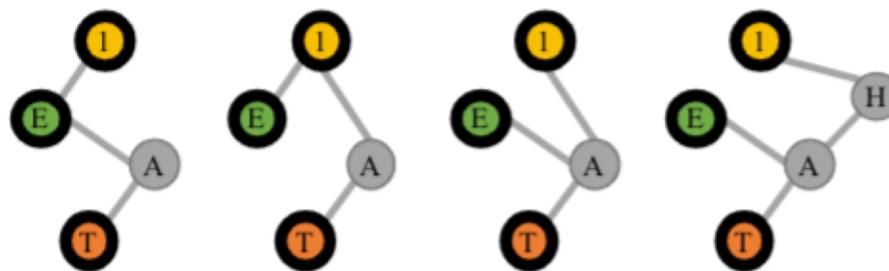
实体关联的排序：排序方法

■ 现有的方法

- 子图的规模 (Size)
- 关系的局部频率 (Freq)
- 实体的中心度 (Centr)
- 关系的信息量 (RInf)
- 实体的信息量 (EInf)
- 实体的具体性 (Spec)

■ 新提出的方法

- 关系的多样性 (RHet)
- 实体的同类性 (EHom)



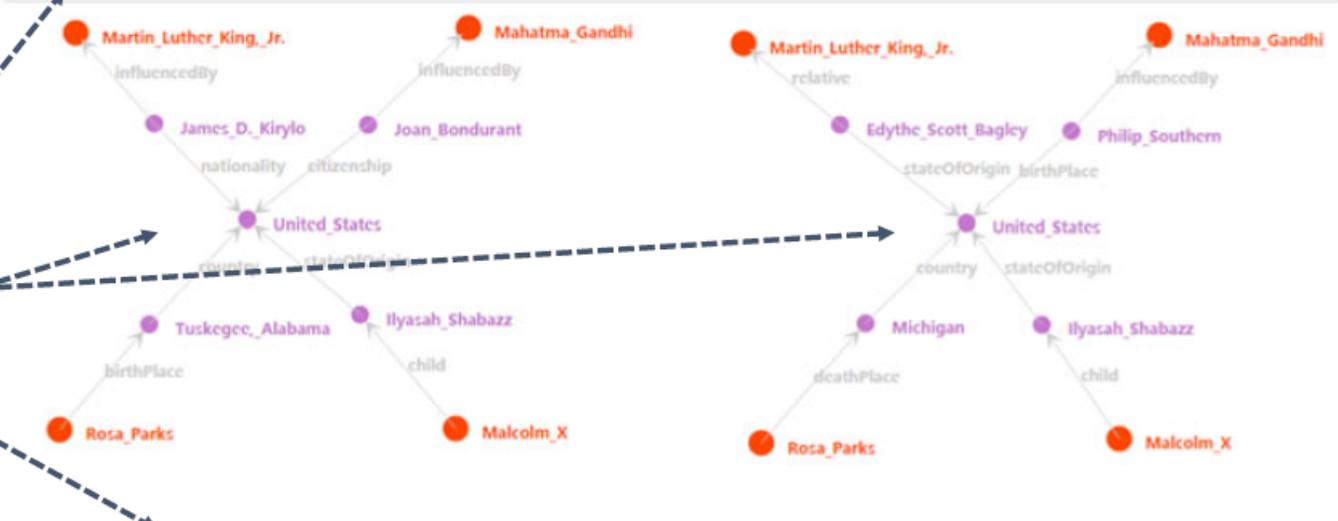
实体关联的排序：用户实验

Short abstracts of Wikipedia articles corresponding to the query entities

Martin Luther King, Jr. (January 15, 1929 – April 4, 1968), was an American Baptist minister, activist, humanitarian, and leader in the African-American Civil Rights Movement. He is best known for his...	Rosa Louise McCauley Parks (February 4, 1913 – October 24, 2005) was an African-American Civil Rights activist, who was the first lady of civil rights and "the mother ...	Mohandas Karamchand Gandhi (/ˈɡɑːndi, ˈɡænz-/; Hindustani: [ˈmoɦɪsəˈdɑːs ˈkərəmtʃənd ˈɡɑːndʒi]; 2 October 1869 – 30 January 1948) was the preeminent leader of the Indian independence movement in Brit...	Malcolm X (/ˈmælkəm ˈɛks/; May 19, 1925 – February 21, 1965), born Malcolm Little and also known as el-Hajj Malik el-Shabazz (Arabic: الحاج ملايكة شابazz), was an American Muslim minister and a human ...
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Ilyasah Shabazz (born July 22, 1962) is the third daughter of Malcolm X and Betty Shabazz. She is an author, most notably of a memoir, *Growing Up X*, community organizer, social activist, and motivatio...

Short abstract of the Wikipedia article corresponding to the entity on mouse over



A pair of semantic associations being compared

Five possible results of comparison

- Significantly more important
- Slightly more important
- Equally important
- Slightly more important
- Significantly more important

An optional explanation for the result of comparison

On the right-hand side there are more common entities.

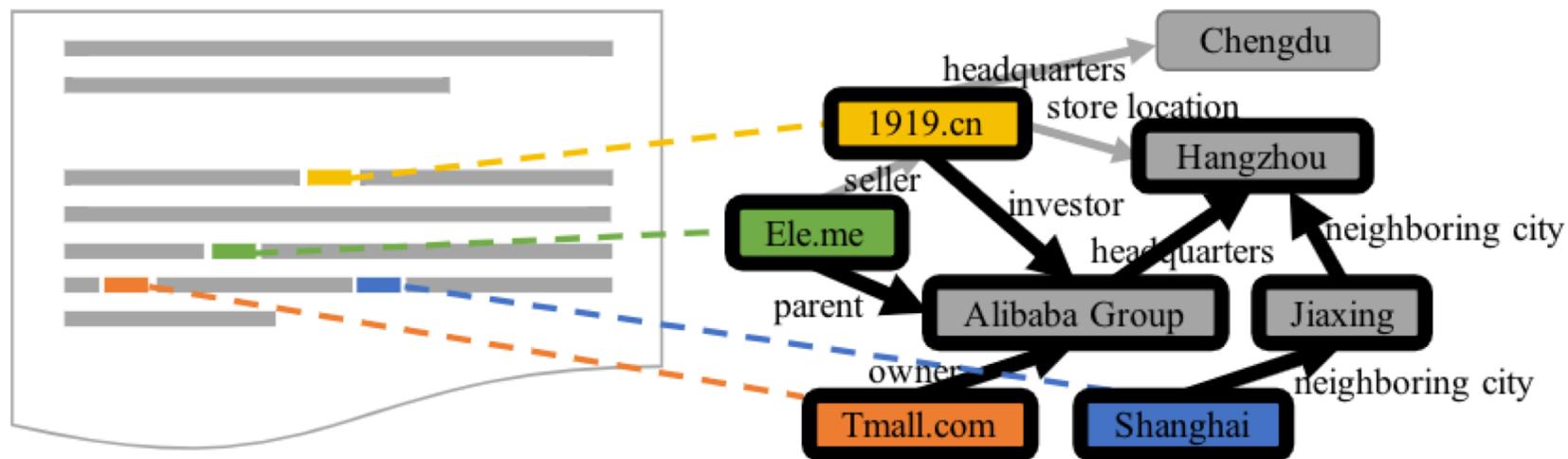
实体关联的排序：实验结论

- 用户喜欢的实体关联：子图规模小、实体同类性强
- 用户对其它指标不具有显著偏好

Centric technique	Mean	Standard deviation	<i>t</i> -test (<i>p</i> -value)
<i>Size</i>	-0.267	0.953	0.001
<i>Freq</i>	0.040	0.904	0.589
<i>Centr</i>	0.067	0.902	0.367
<i>RInf</i>	-0.080	0.959	0.309
<i>EInf</i>	-0.087	0.874	0.227
<i>Spec</i>	0.007	0.966	0.933
<i>RHet</i>	0.020	0.952	0.797
<i>EHom</i>	0.180	0.956	0.022

实体关联的排序：上下文相关

- 上下文：新闻中出现的其它实体
- 相关性：实体类型的相似性



相关论文

- Yu Gu, Tianshuo Zhou, Gong Cheng, Ziyang Li, Jeff Z. Pan, Yuzhong Qu.
Relevance Search over Schema-Rich Knowledge Graphs.
WSDM 2019
- Tianshuo Zhou, Ziyang Li, Gong Cheng, Jun Wang, Yu'Ang Wei.
GREASE: A Generative Model for Relevance Search over Knowledge Graphs.
WSDM 2020
- Zixian Huang, Shuxin Li, Gong Cheng, Evgeny Kharlamov, Yuzhong Qu.
MiCRon: Making Sense of News via Relationship Subgraphs.
CIKM 2019 (Demo)
- Gong Cheng, Daxin Liu, Yuzhong Qu.
Fast Algorithms for Semantic Association Search and Pattern Mining.
TKDE 2019
- Gong Cheng, Fei Shao, Yuzhong Qu.
An Empirical Evaluation of Techniques for Ranking Semantic Associations.
TKDE 2017



- 谢谢 & 欢迎讨论