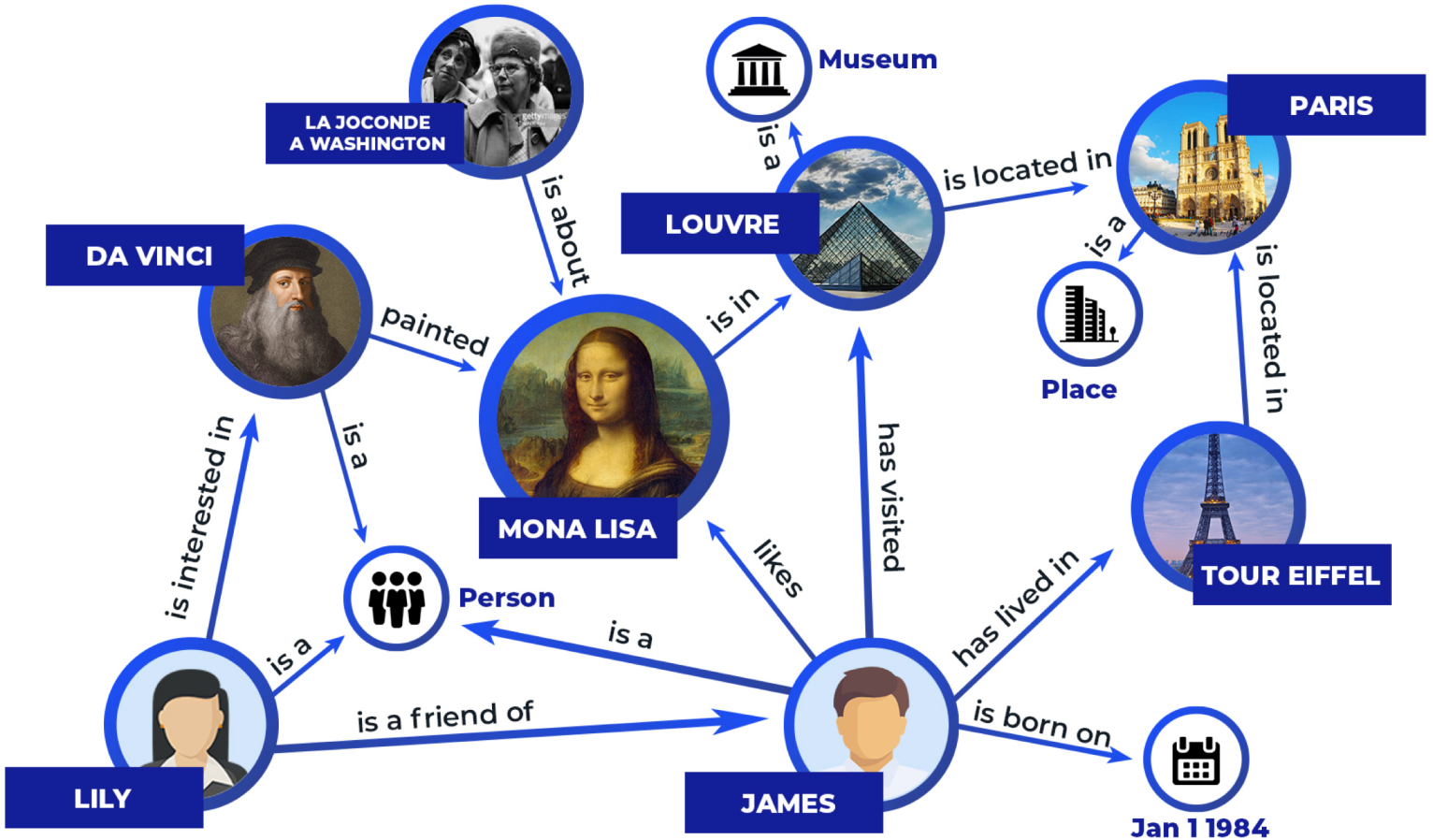
知识图谱中的关联搜索

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：

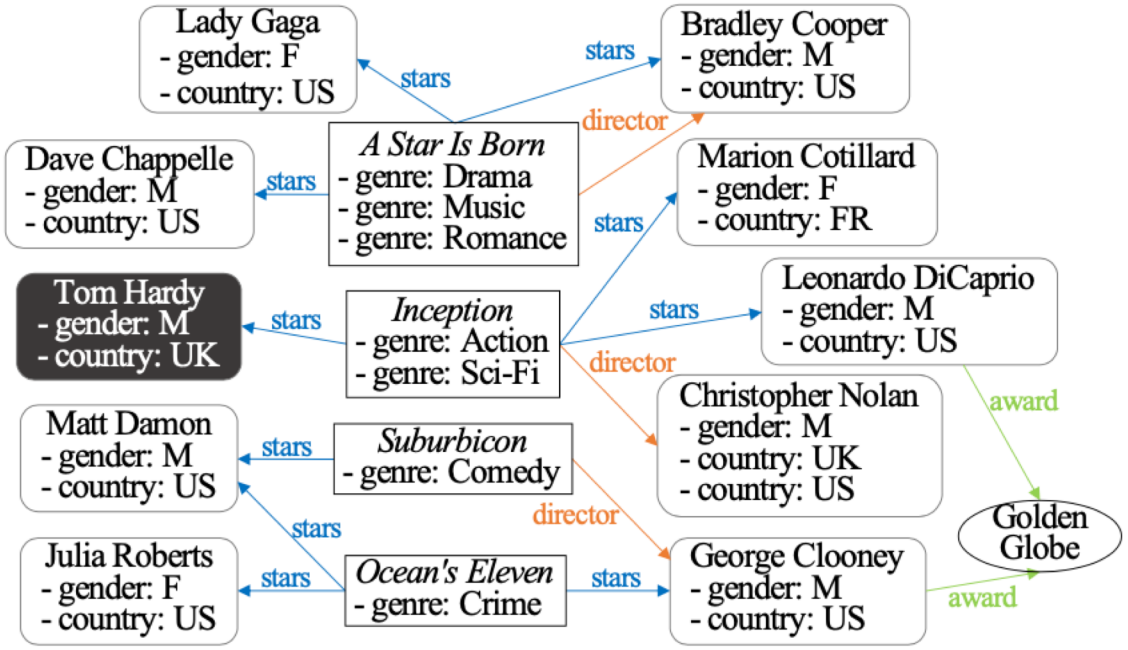
$S_1 = \{\langle$Dave Chappelle, Lady Gaga$\rangle$, $\langle$Matt Damon, Julia Roberts$\rangle\}$,

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

- 对应的元路径（meta-path）

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

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



Gu et al., WSDM 2019

实体搜索（相关实体推荐）

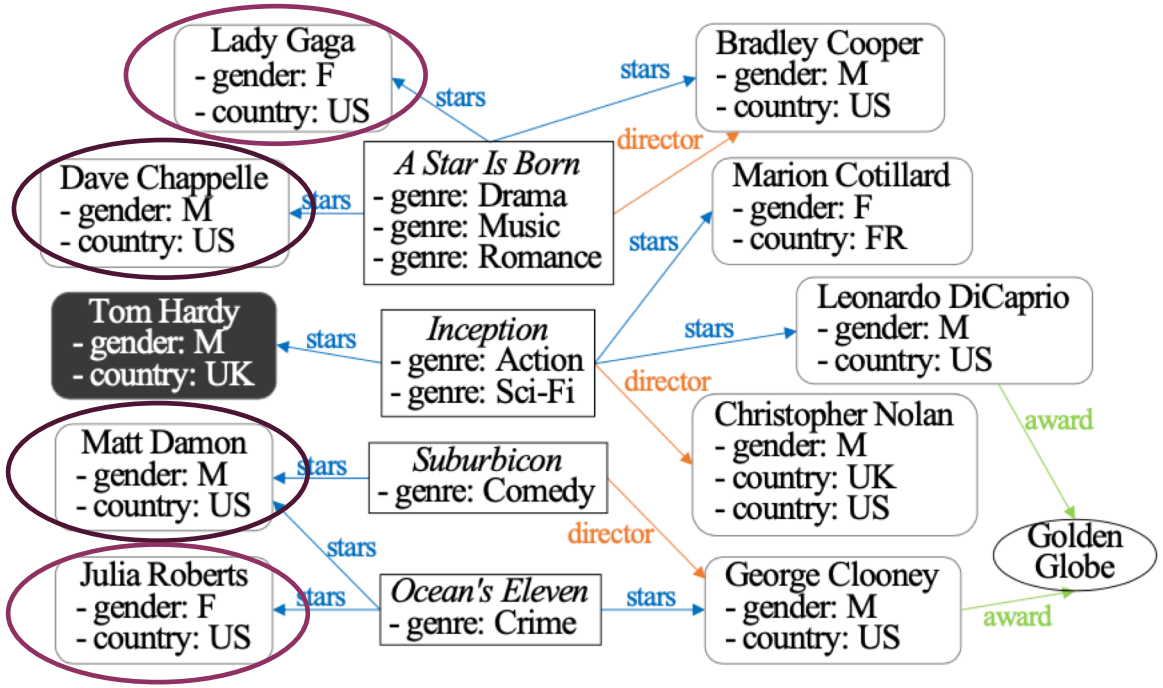
- 查询实体：Tom Hardy
- 两种相关性的样例：

$S_1 = \{\langle \texttt{Dave Chappelle, Lady Gaga} \rangle,$
$\langle \texttt{Matt Damon, Julia Roberts} \rangle\},$
$S_2 = \{\langle \texttt{Dave Chappelle, Bradley Cooper} \rangle,$
$\langle \texttt{Matt Damon, George Clooney} \rangle\}.$

- 对应的元路径（meta-path）

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

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



Gu et al., WSDM 2019

实体搜索（相关实体推荐）

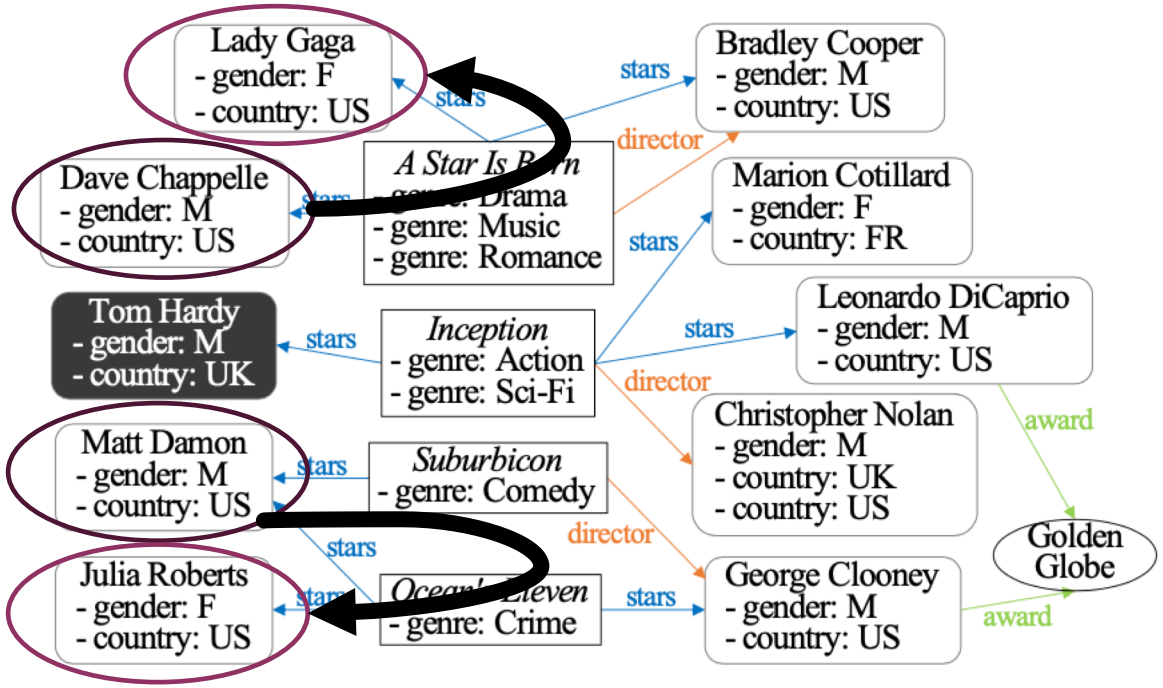
- 查询实体：Tom Hardy
- 两种相关性的样例：

$S_1 = \{\langle \texttt{Dave Chappelle}, \texttt{Lady Gaga}\rangle,$
$\langle \texttt{Matt Damon}, \texttt{Julia Roberts}\rangle\},$
$S_2 = \{\langle \texttt{Dave Chappelle}, \texttt{Bradley Cooper}\rangle,$
$\langle \texttt{Matt Damon}, \texttt{George Clooney}\rangle\}.$

- 对应的元路径（meta-path）

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

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



Gu et al., WSDM 2019

实体搜索（相关实体推荐）

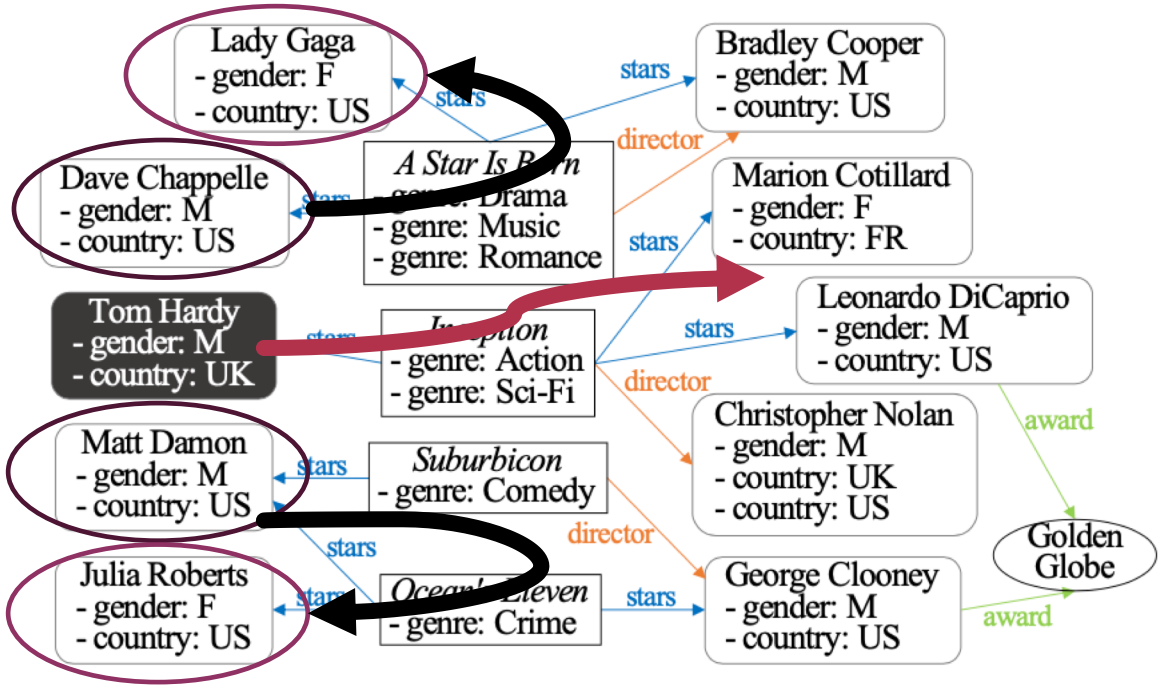
- 查询实体：Tom Hardy
- 两种相关性的样例：

$$S_1 = \{\langle \texttt{Dave Chappelle}, \texttt{Lady Gaga}\rangle, \langle \texttt{Matt Damon}, \texttt{Julia Roberts}\rangle\},$$
$$S_2 = \{\langle \texttt{Dave Chappelle}, \texttt{Bradley Cooper}\rangle, \langle \texttt{Matt Damon}, \texttt{George Clooney}\rangle\}.$$

- 对应的元路径（meta-path）

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

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



Gu et al., WSDM 2019

实体搜索（相关实体推荐）

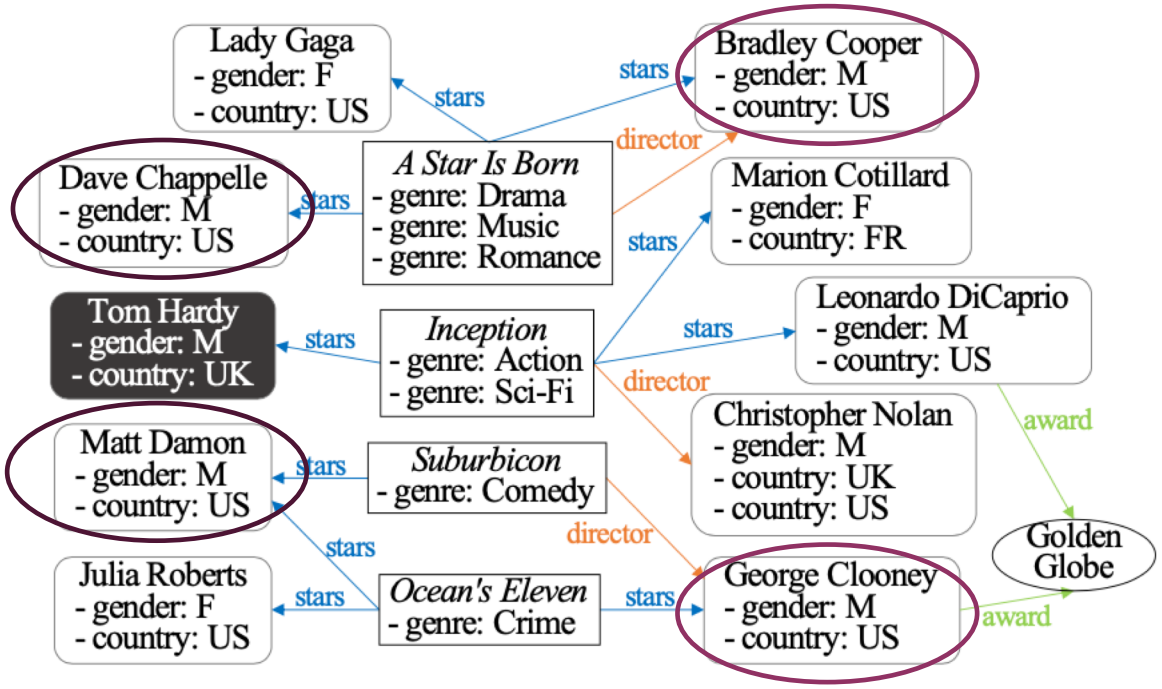
- 查询实体：Tom Hardy
- 两种相关性的样例：

$S_1 = \{\langle \texttt{Dave Chappelle, Lady Gaga} \rangle,$
$\langle \texttt{Matt Damon, Julia Roberts} \rangle\},$
$S_2 = \{\langle \texttt{Dave Chappelle, Bradley Cooper} \rangle,$
$\langle \texttt{Matt Damon, George Clooney} \rangle\}.$

- 对应的元路径（meta-path）

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

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



Gu et al., WSDM 2019

实体搜索（相关实体推荐）

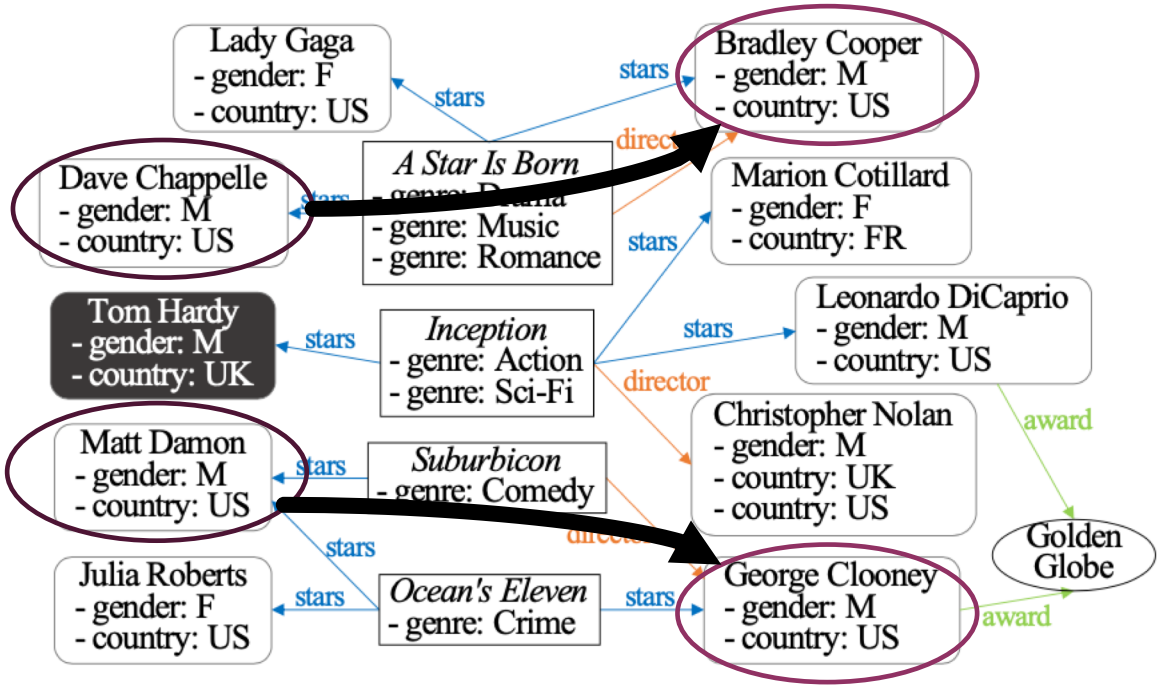
- 查询实体：Tom Hardy
- 两种相关性的样例：

$S_1 = \{\langle \texttt{Dave Chappelle, Lady Gaga} \rangle,$
$\langle \texttt{Matt Damon, Julia Roberts} \rangle\},$
$S_2 = \{\langle \texttt{Dave Chappelle, Bradley Cooper} \rangle,$
$\langle \texttt{Matt Damon, George Clooney} \rangle\}.$

- 对应的元路径（meta-path）

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

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



Gu et al., WSDM 2019

实体搜索（相关实体推荐）

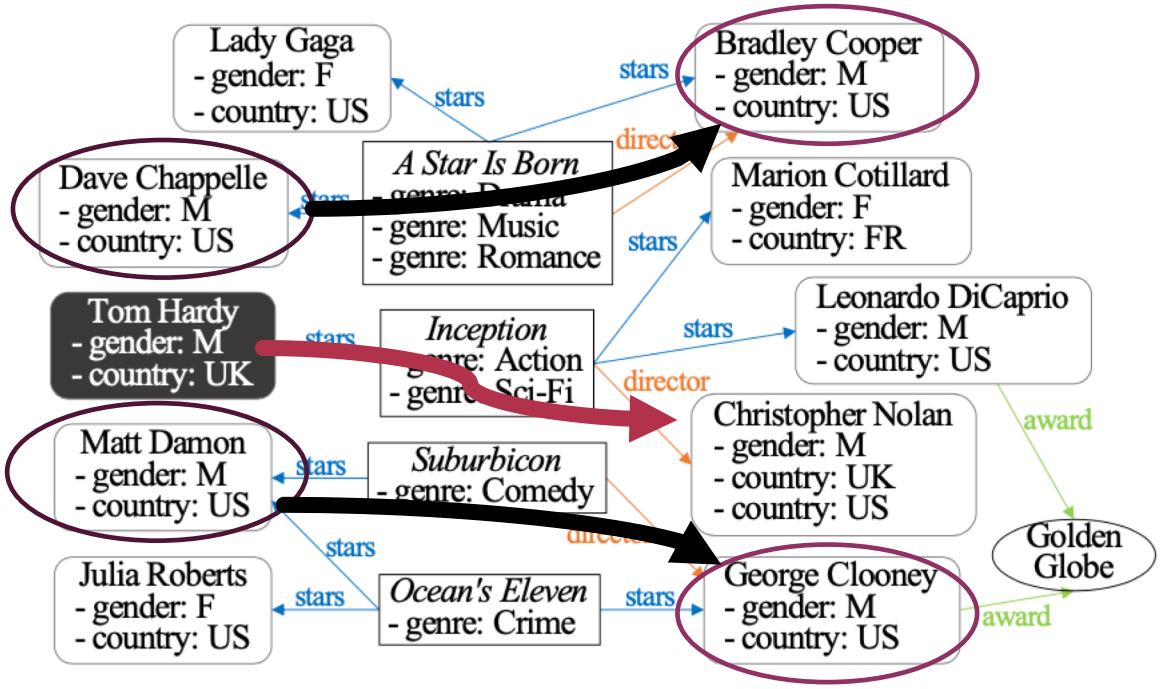
- 查询实体：Tom Hardy
- 两种相关性的样例：

$S_1 = \{\langle \texttt{Dave Chappelle}, \texttt{Lady Gaga}\rangle,$
$\langle \texttt{Matt Damon}, \texttt{Julia Roberts}\rangle\},$
$S_2 = \{\langle \texttt{Dave Chappelle}, \texttt{Bradley Cooper}\rangle,$
$\langle \texttt{Matt Damon}, \texttt{George Clooney}\rangle\}.$

- 对应的元路径（meta-path）

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

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

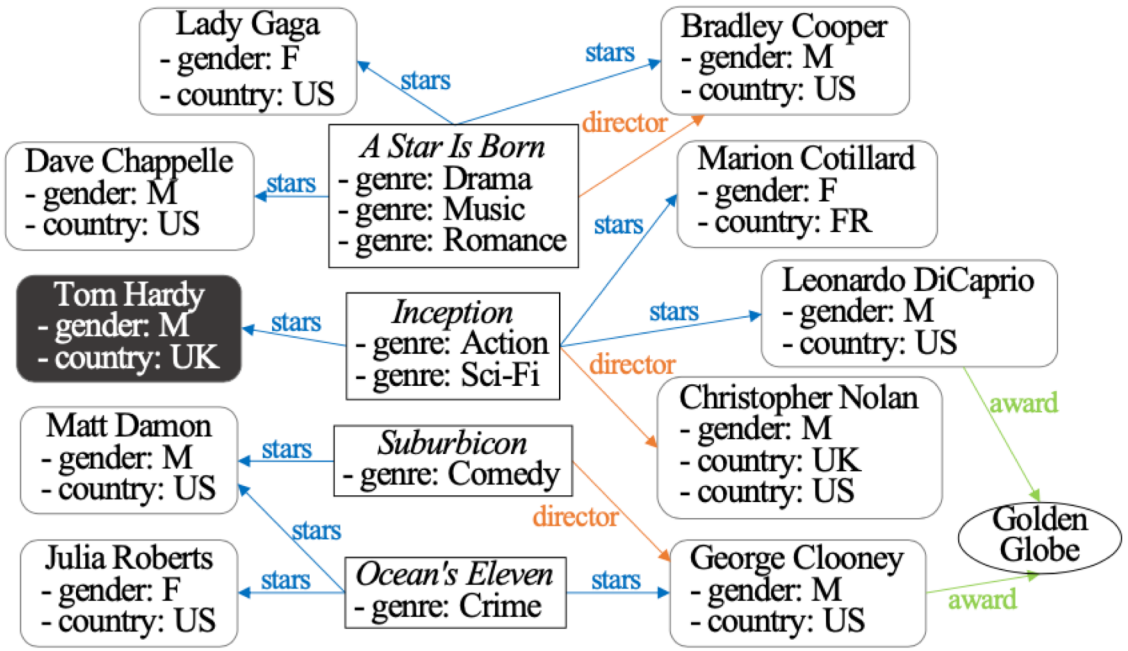


Gu et al., WSDM 2019

基于元路径的相关度

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

- 元路径的选取（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）



Gu et al., WSDM 2019

元路径的选取

- 元路径的重要性（对样例的拟合程度）

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

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

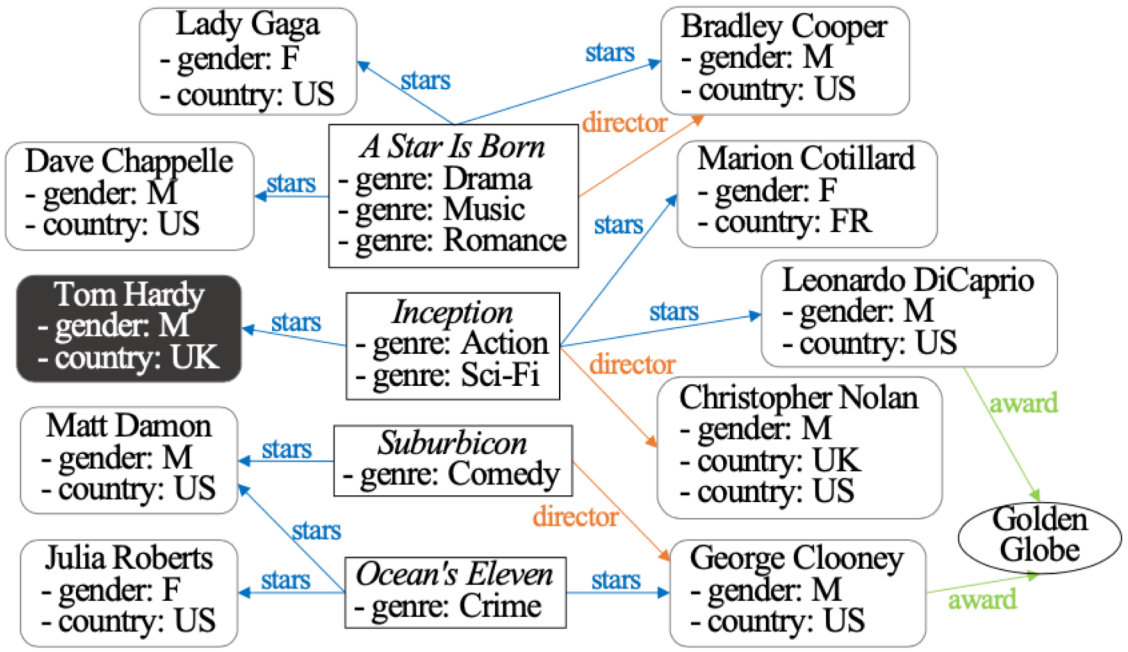
- 元路径的多样性（去除冗余的元路径）

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

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

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

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



Gu et al., WSDM 2019

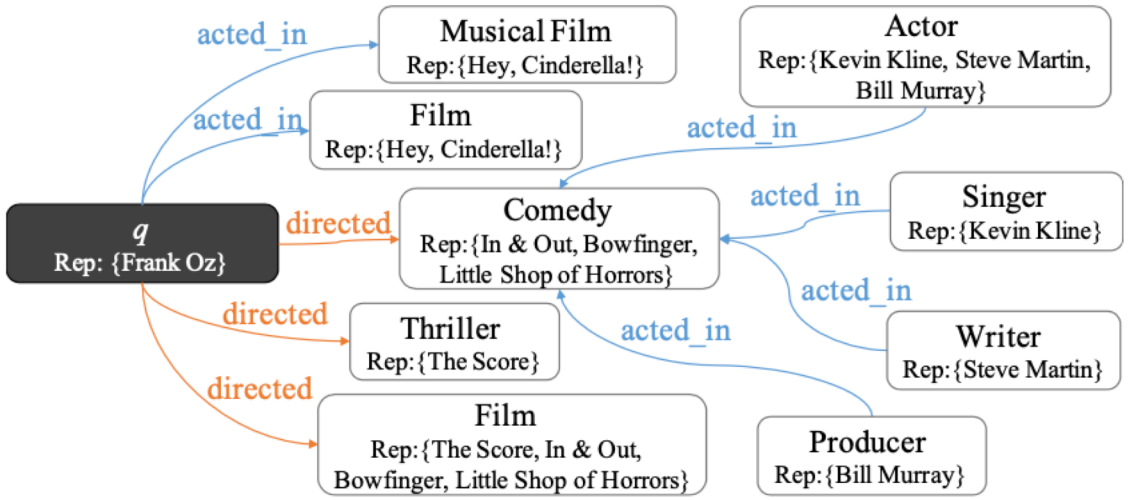
元路径的选取

- 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.

```
Initialize a search tree ST with q as the root;
MP ← ∅;
while ST can be expanded do
    stn ← SelTN(ST);
    TN ← ExpST(stn);
    foreach tn ∈ TN do
        EvalTN(tn);
        if tn contains entities in Λ then
            P_i = GetMP(tn);
            if sig(P_i|q, Λ) ≥ τ then
                if P_i ~_q P_j for any P_j ∈ MP then
                    Skip tn and continue with the next iteration;
                else
                    MP ← MP ∪ {P_i};
                    if |MP| = n then
                        Terminate the loop;
return MP;
```



启发函数：

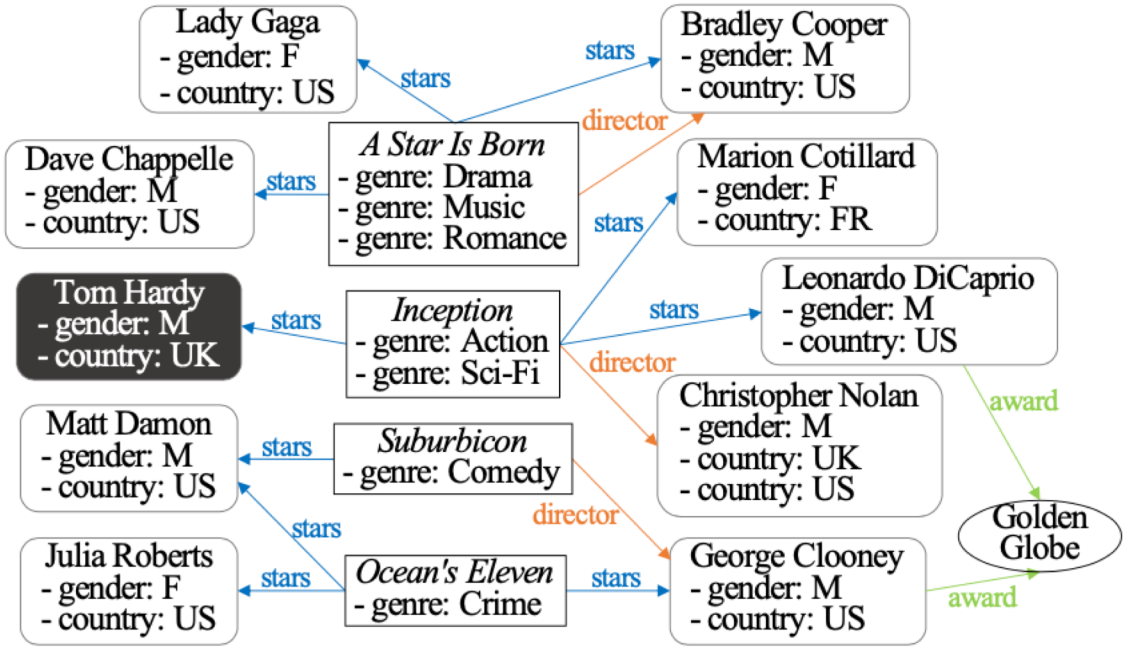
- 路径短 $\mathtt{prms}_{\mathtt{dist}}(tn) = \mathtt{dpth}(tn) + \frac{1}{|\Lambda|}\sum_{\lambda\in\Lambda}\min_{v\in\mathtt{Rep}(tn)\setminus\{\lambda\}}\mathtt{dist}(v,\lambda)$
- 度数小 $\mathtt{prms}_{\mathtt{deg}}(tn) = \sum_{v\in\mathtt{Rep}(tn)}\mathtt{deg}(v)$

Gu et al., WSDM 2019

元路径的加权

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

- 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



Gu et al., WSDM 2019

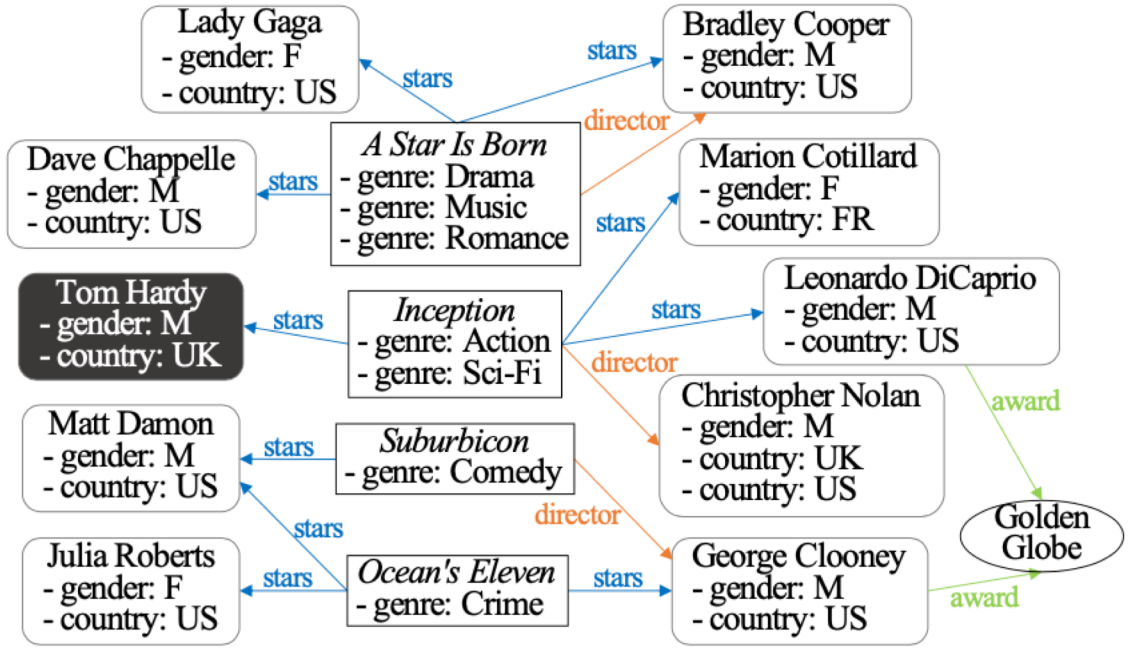
提纲

- 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

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

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

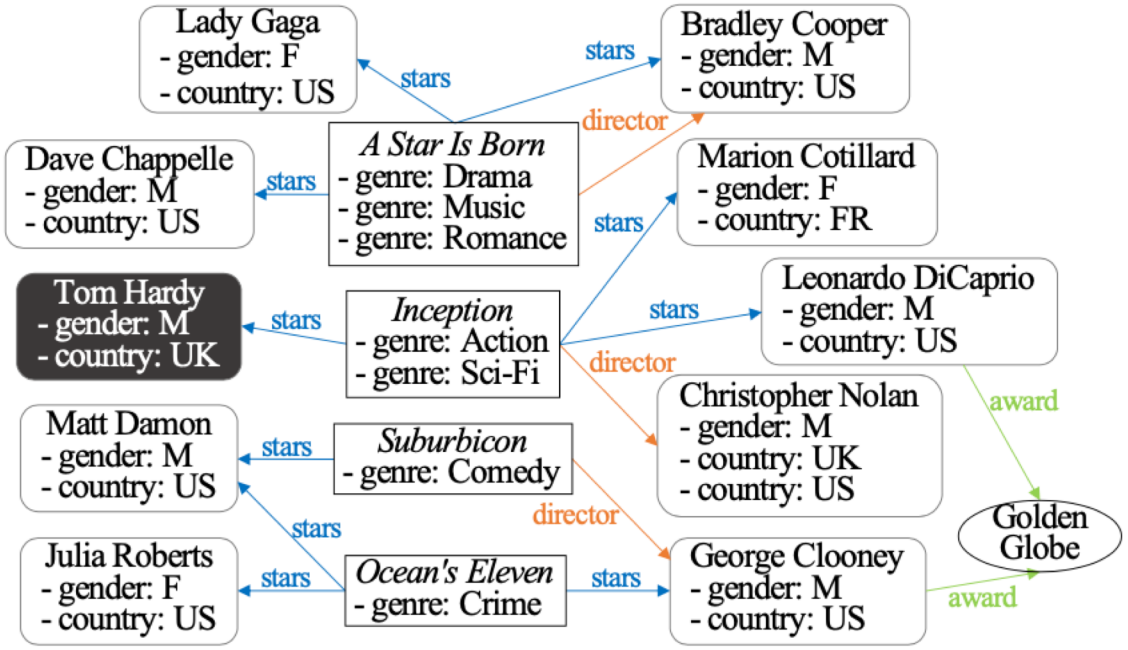


Zhou et al., WSDM 2020

生成模型

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

- $\Pr(P_i)$：先验
- $\Pr(S|P_i)$：似然



Zhou et al., WSDM 2020

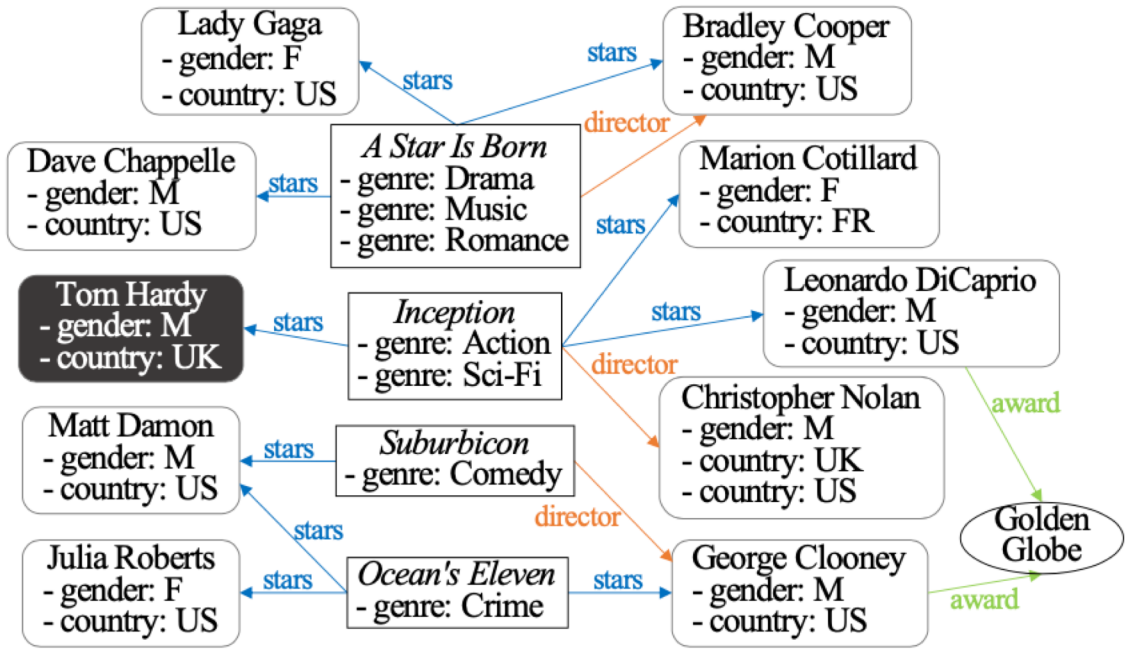
先验的计算

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

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

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

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



Zhou et al., WSDM 2020

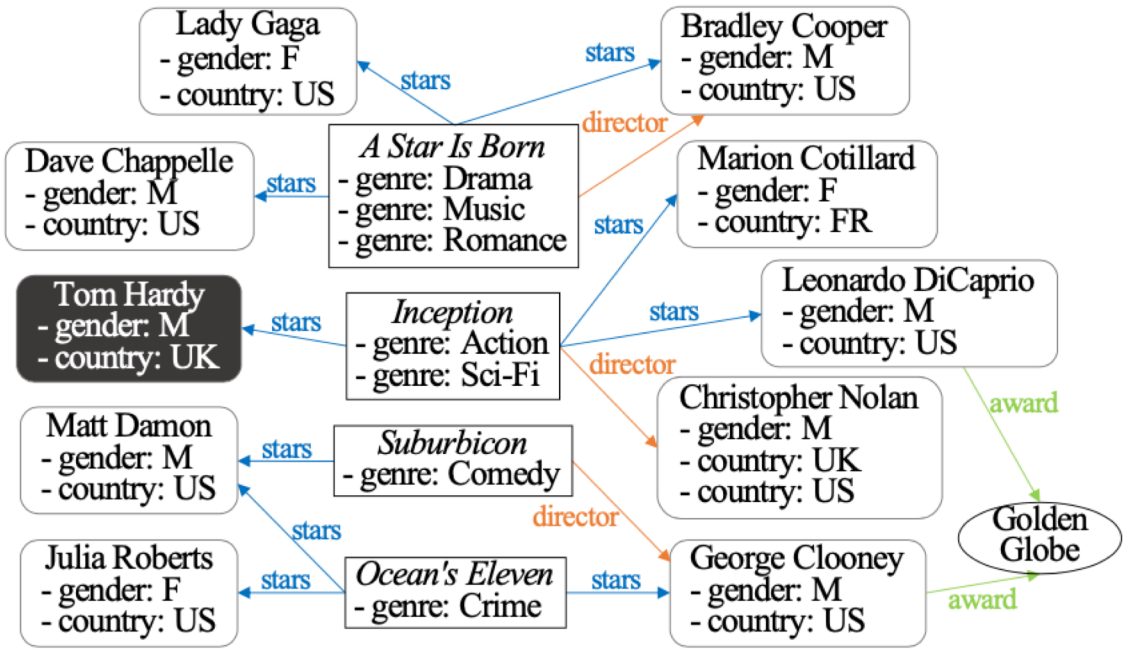
似然的计算

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

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

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

$$\text{pc}(r_1)\prod_{i=2}^{l}\frac{\text{pc}(r_{i-1}r_i)}{\text{pc}(r_{i-1})} \qquad \text{pc}(r_l)\prod_{i=1}^{l-1}\frac{\text{pc}(r_i r_{i+1})}{\text{pc}(r_{i+1})}$$

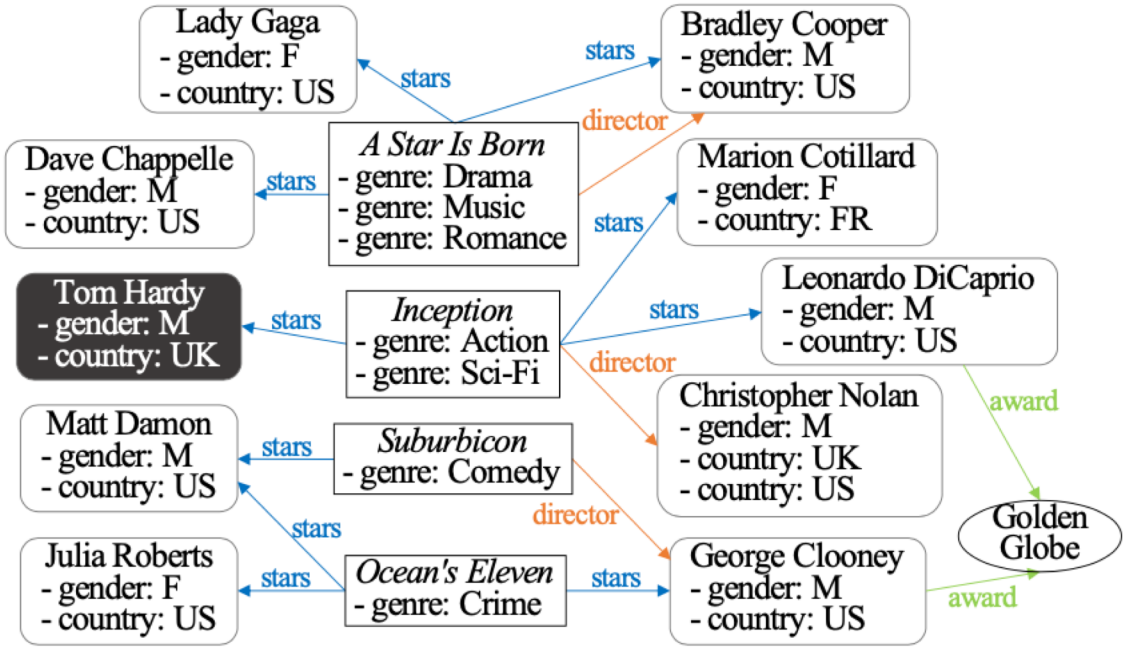


Zhou et al., WSDM 2020

扩展的相关度

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

- 元路径关联 + 属性约束



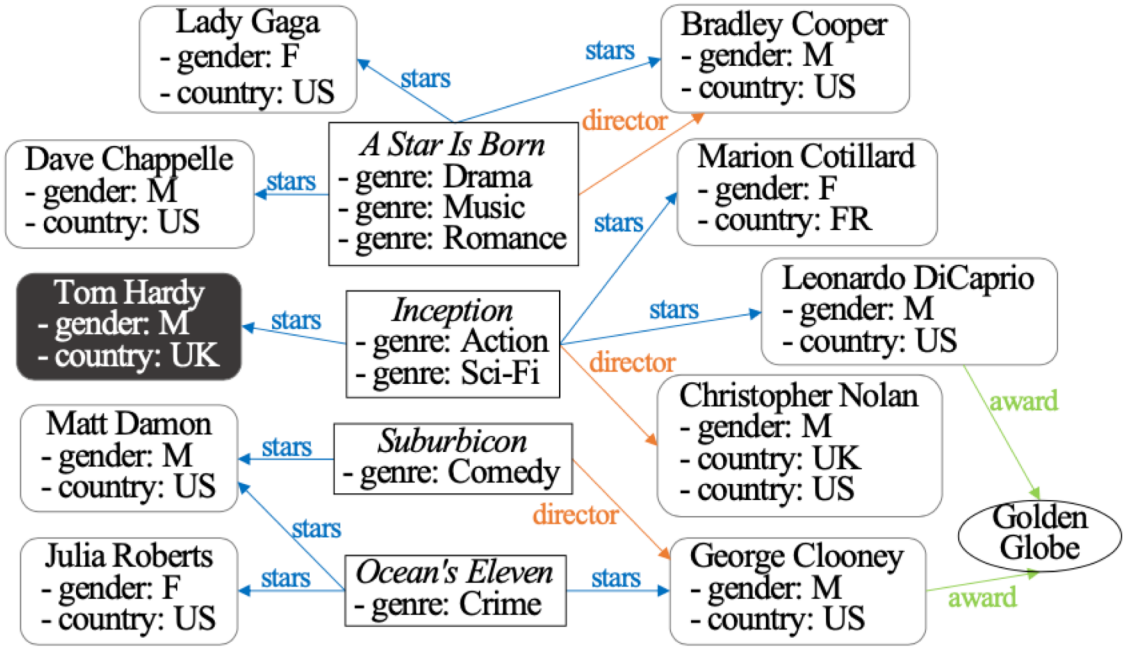
Zhou et al., WSDM 2020

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.

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



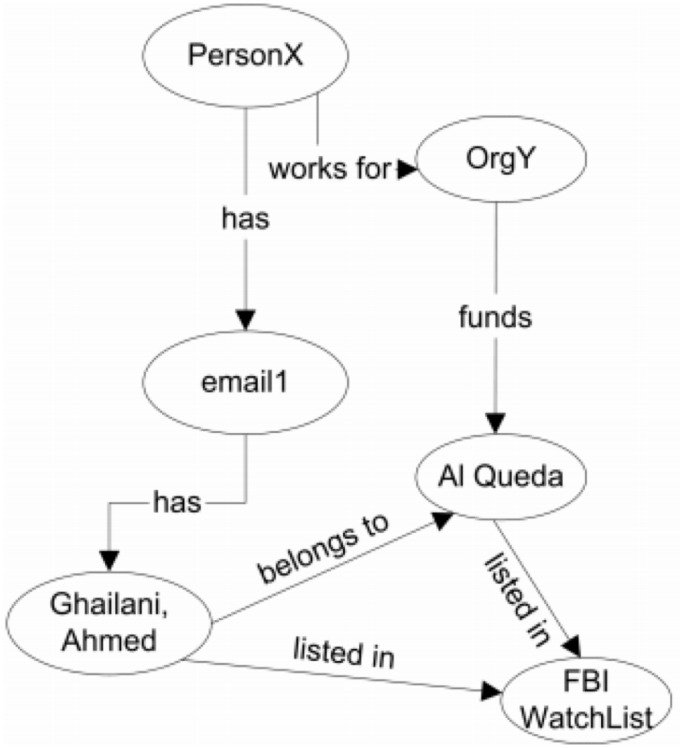
Zhou et al., WSDM 2020

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

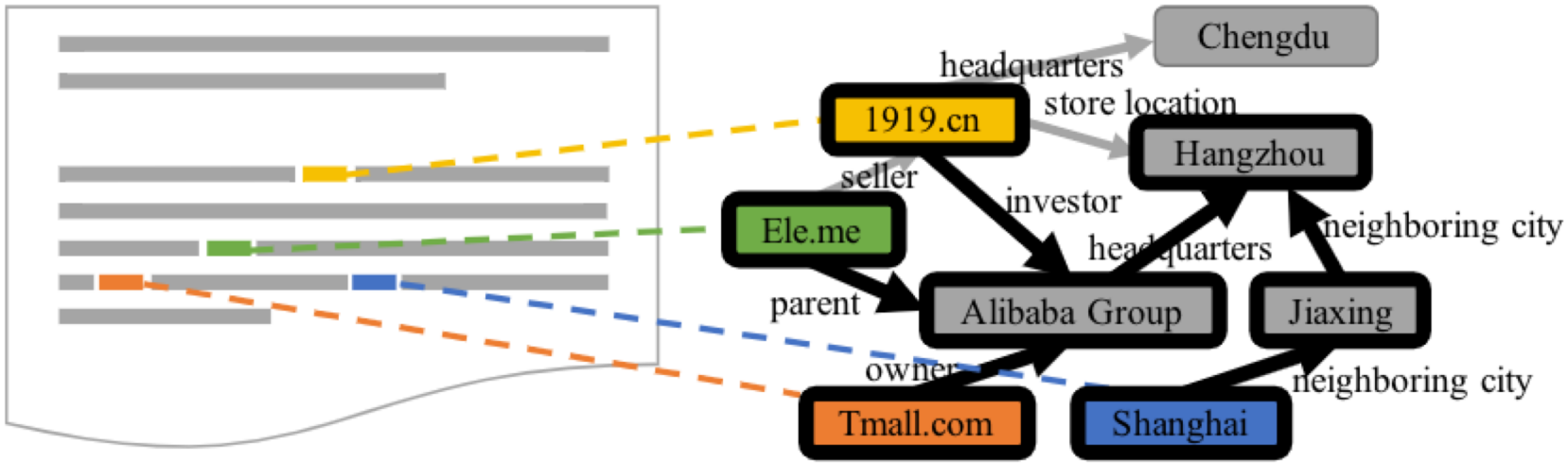
实体关联搜索的应用：国家安全

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?



Sheth et al., JDM 2005

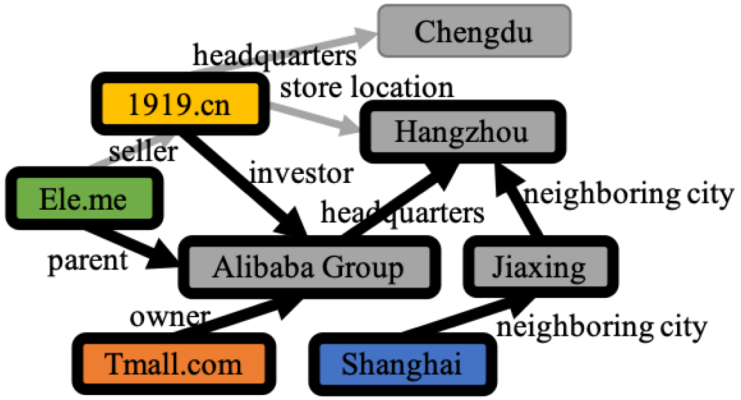
实体关联搜索的应用：新闻阅读



Huang et al., CIKM 2019

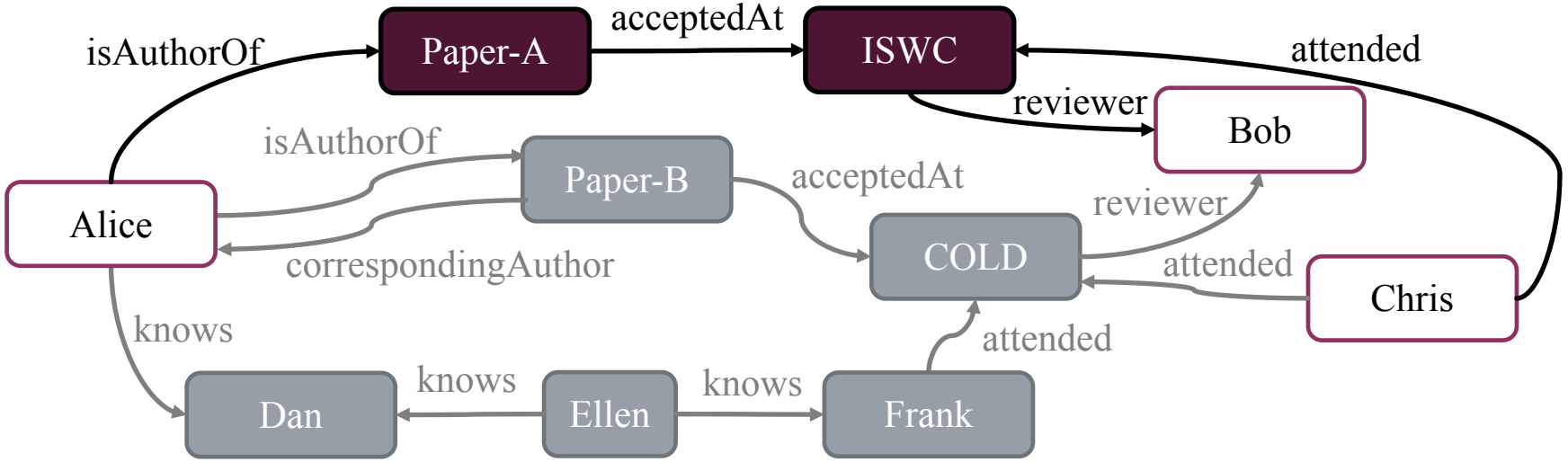
实体关联的定义

- 包含所有查询实体的极小连通子图
- 直径受限



Cheng et al., TKDE 2019

实体关联的发现



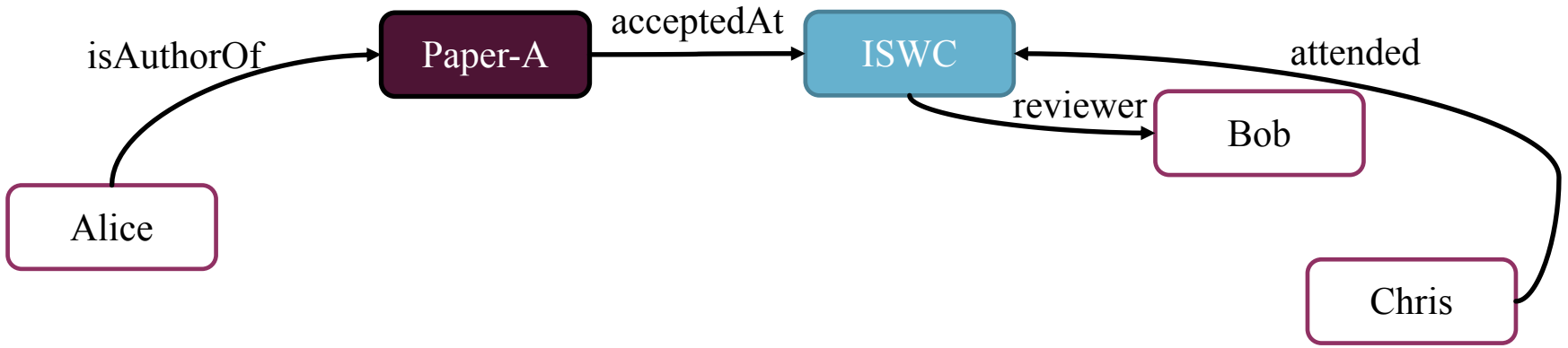
Cheng et al., TKDE 2019

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.

$X \leftarrow$ empty set of SAs
for $i \leftarrow 1$ to n **do** // n: number of query vertices
  $P_i \leftarrow$ BLPathEnum(G, v_i^Q, D) // v_i^Q: i-th query vertex
end for
for all $\langle p_1, \ldots, p_n \rangle \in (P_1 \times \cdots \times P_n)$ **do**
  if $p_1, \ldots, p_n$ have a common end vertex **then**
    $x \leftarrow$ subgraph formed by joining $p_1, \ldots, p_n$
    if x is minimal and $diam(x) \leq D$ **then**
      $X \leftarrow X \cup \{x\}$
    end if
  end if
end for
return X



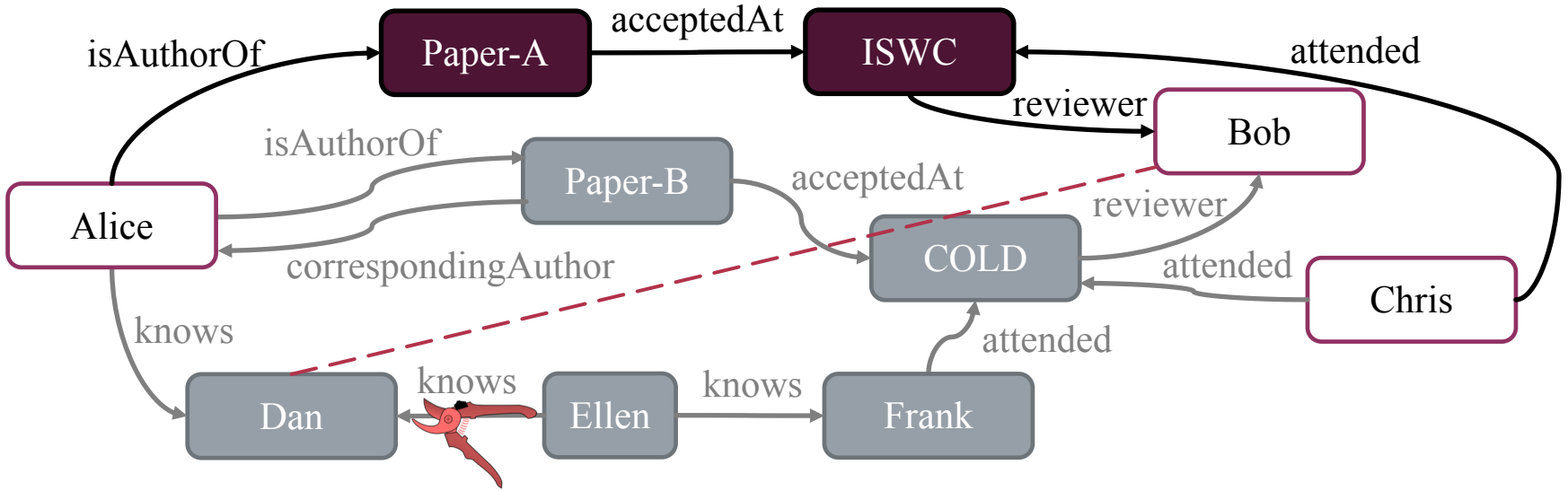
Cheng et al., TKDE 2019

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.

$P^0 \leftarrow \{v_i^Q\}$ // v_i^Q: treated as a path of length 0
for $j \leftarrow 1$ to $\lfloor \frac{D+1}{2} \rfloor$ **do**
  $P^j \leftarrow$ empty set of paths // P^j: paths of length j
  for all $p \in P^{j-1}$ **do**
    $pev \leftarrow$ end vertex of p
    if $ln(p)+dist(pev, v_j^Q) \leq D$ for all $v_j^Q \in (V_Q \backslash \{v_i^Q\})$ **then**
      $P^j \leftarrow P^j \cup$ paths formed by joining p and each arc incident from/to pev
    end if
  end for
end for
$P \leftarrow P^0 \cup \cdots \cup P^{\lfloor \frac{D+1}{2} \rfloor}$
for all $p \in P$ **do**
  $pev \leftarrow$ end vertex of p
  if $dist(pev, v_j^Q) > \lfloor \frac{D+1}{2} \rfloor$ for any $v_j^Q \in (V_Q \setminus \{v_i^Q\})$ **then**
    $P \leftarrow P \setminus \{p\}$
  end if
end for
return P



- 长度(Alice→Dan) + 距离(Dan, Bob) > 直径约束

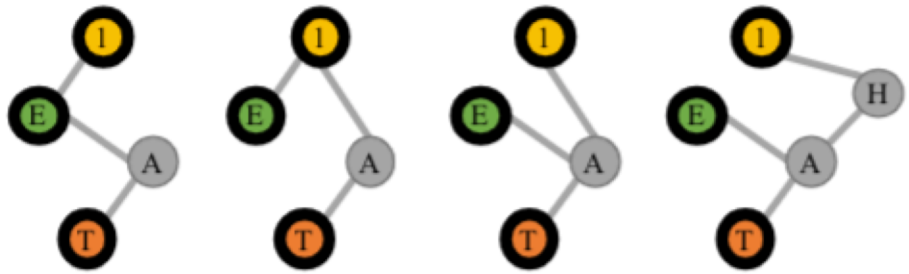
Cheng et al., TKDE 2019

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

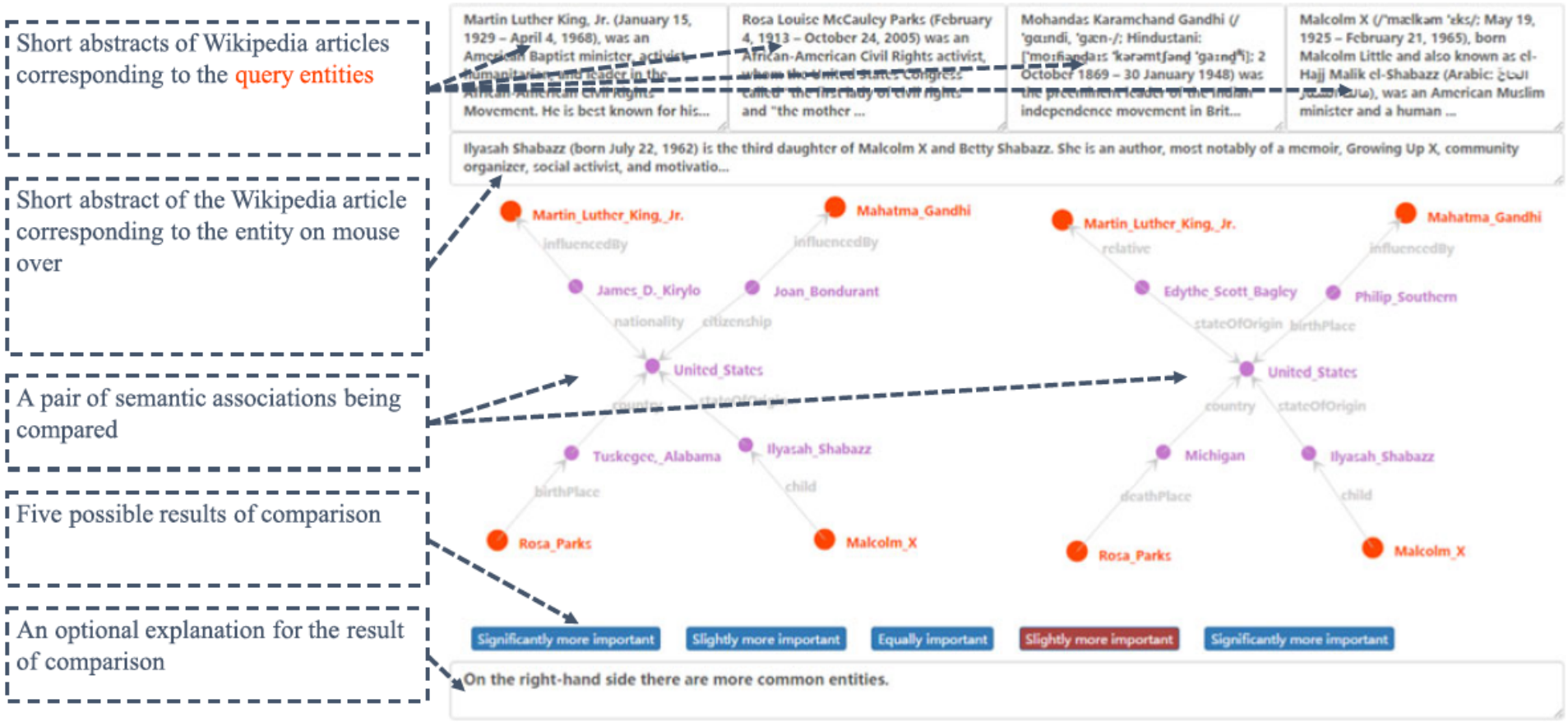
实体关联的排序：排序方法

- 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）



Cheng et al., TKDE 2017

实体关联的排序：用户实验



Cheng et al., TKDE 2017

实体关联的排序：实验结论

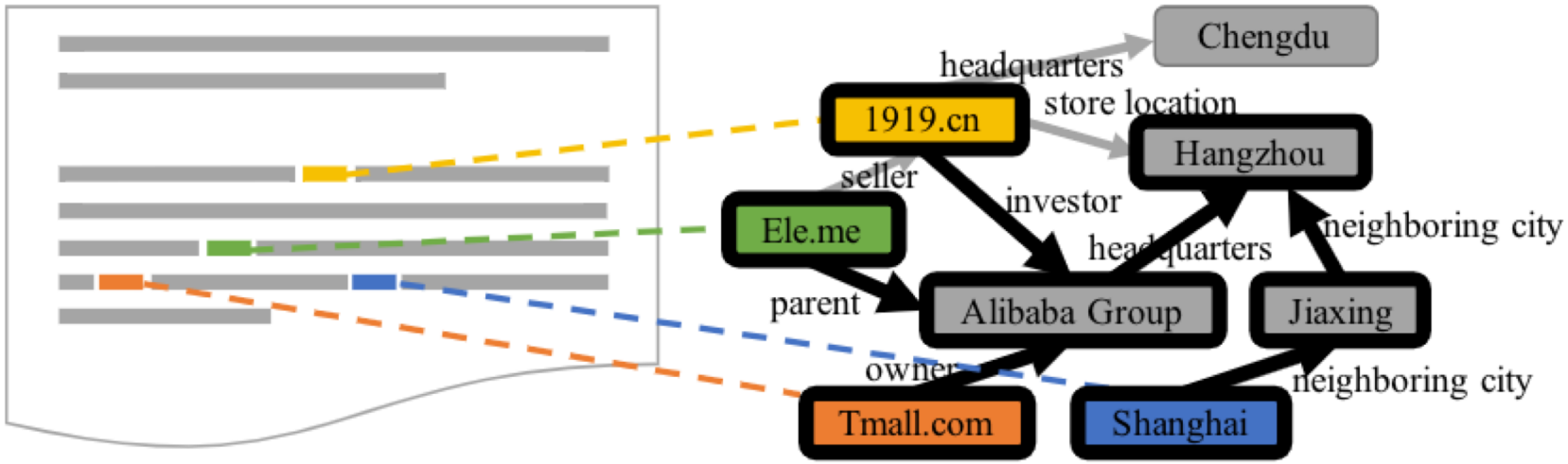
- 用户喜欢的实体关联：子图规模小、实体同类性强
- 用户对其它指标不具有显著偏好

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

Cheng et al., TKDE 2017

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



Huang et al., CIKM 2019

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