SEED：基于RDF知识库的实体扩展与知识纠错系统

陈跃国 中国人民大学

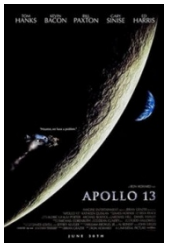
概要

- 研究背景
- 问题定义
- 关键技术
- 实验结果
- 系统展示
- 总结与展望

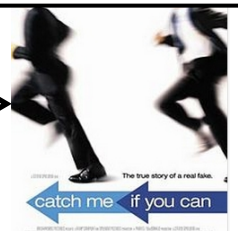
实体集合扩展-研究背景

应用场景

I. QA
II. 信息推荐
III. 知识库探索式搜索
IV. 知识库纠错

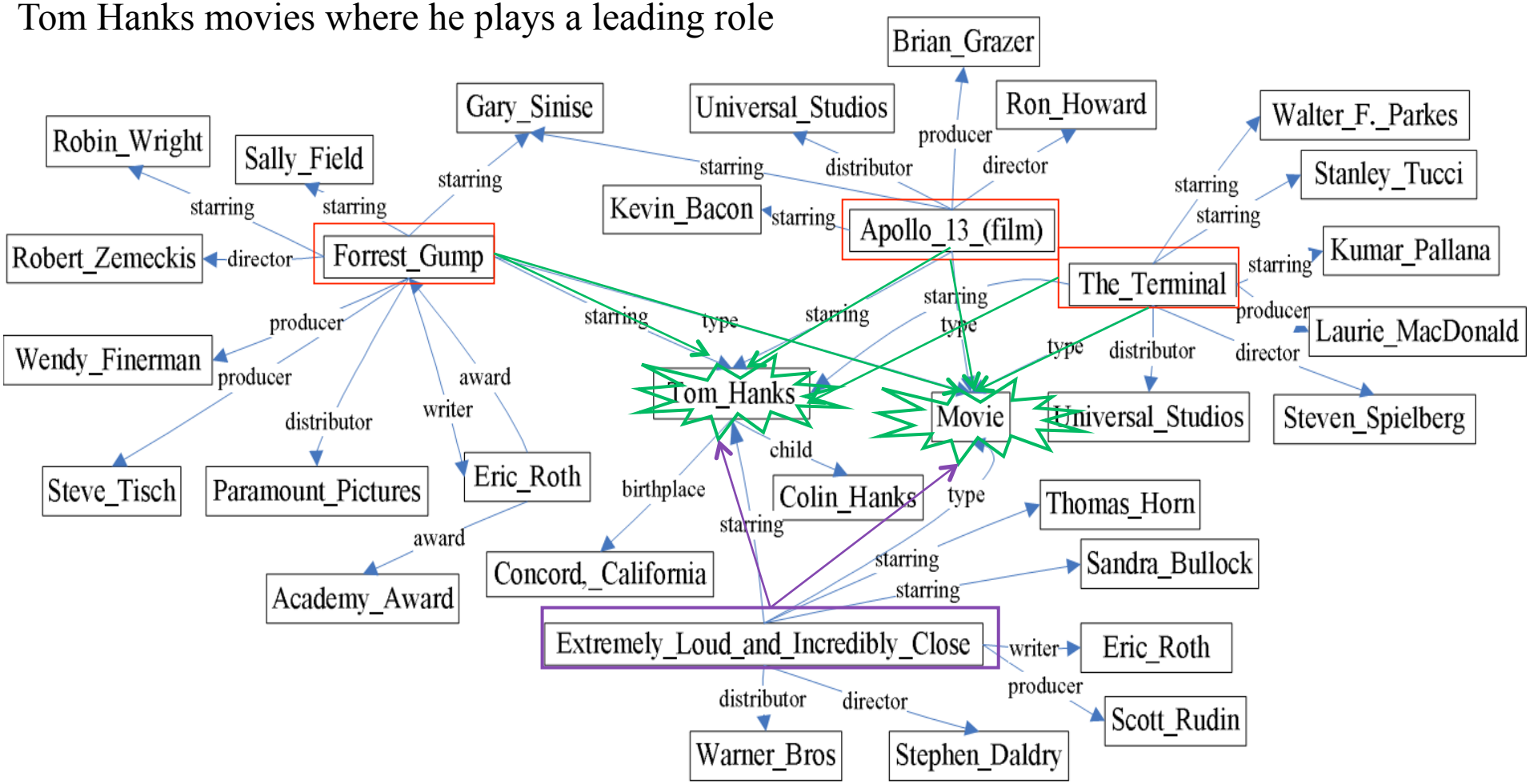


Movie



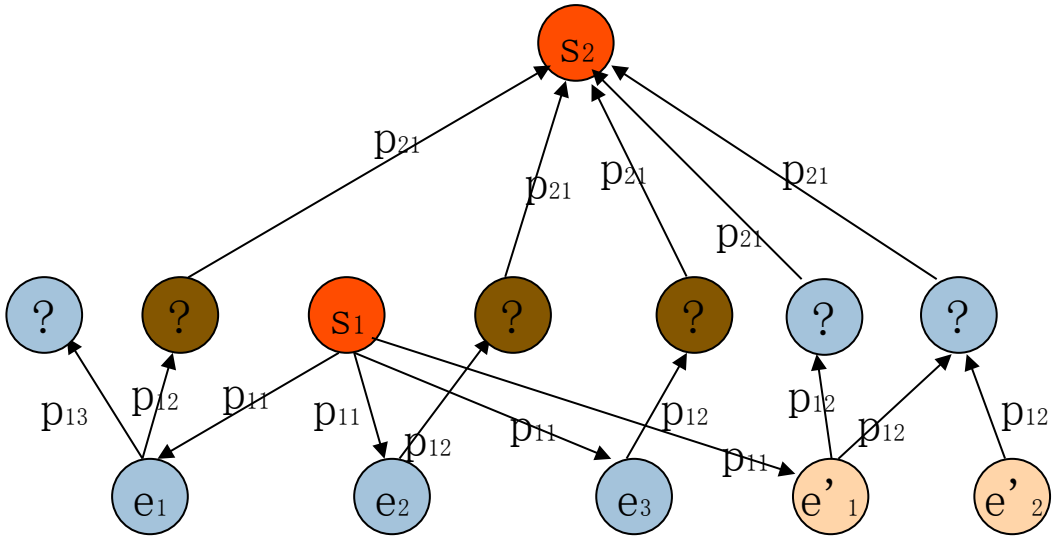
问题定义

Tom Hanks movies where he plays a leading role



Problem: given some seeds, find relevant entities having some implicit common features with those seeds

Semantic Patterns



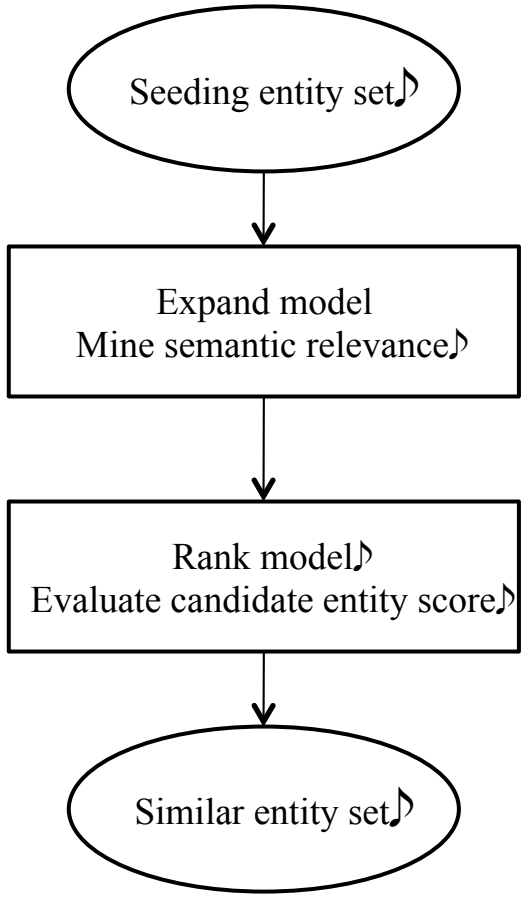
$$\pi_1 = s_1 : p_{11}$$

$$\pi_2 = s_2 : \underline{p_{21}} \circ \underline{p_{12}}$$

$$\Phi(e_1, e_2, e_3) = \{\pi_1, \pi_2\}$$

关键技术

- 引入semantic pattern的概念
  - Step 1: 根据种子节点找到相关的semantic patterns;
  - Step 2: 找到满足已找到的semantic patterns的候选实体;
  - Step 3: 根据semantic patterns对候选实体排序(ranking model)



排序模型

- 实体相关度 $r(e) = \sum_{\pi \in \tilde{\Phi}(Q) \wedge e \models \pi} d(\pi) * r(\pi, Q)$
- Discriminability $d(\pi) = \prod_{i=1}^{n} \min(\frac{1}{|E(\pi_i)|}, \frac{|E(\pi_i)|}{\sum_{e \in E(\pi_i)} |E(e : l_i)|})$
- Relevance $r(\pi, Q) = \prod_{e \in Q} p(e, \pi)$

实验设置

- RDF 知识库
  - DBPedia Version 3.9，4M entities
- 测试集
  - QALD-2/3/4，INEX-XER 2009
- 评价指标
  - Precision@N，R-Precision，MRR
- 对比对象
  - BBR, SEAL, LDSD, ARM and QBESS

主要结果

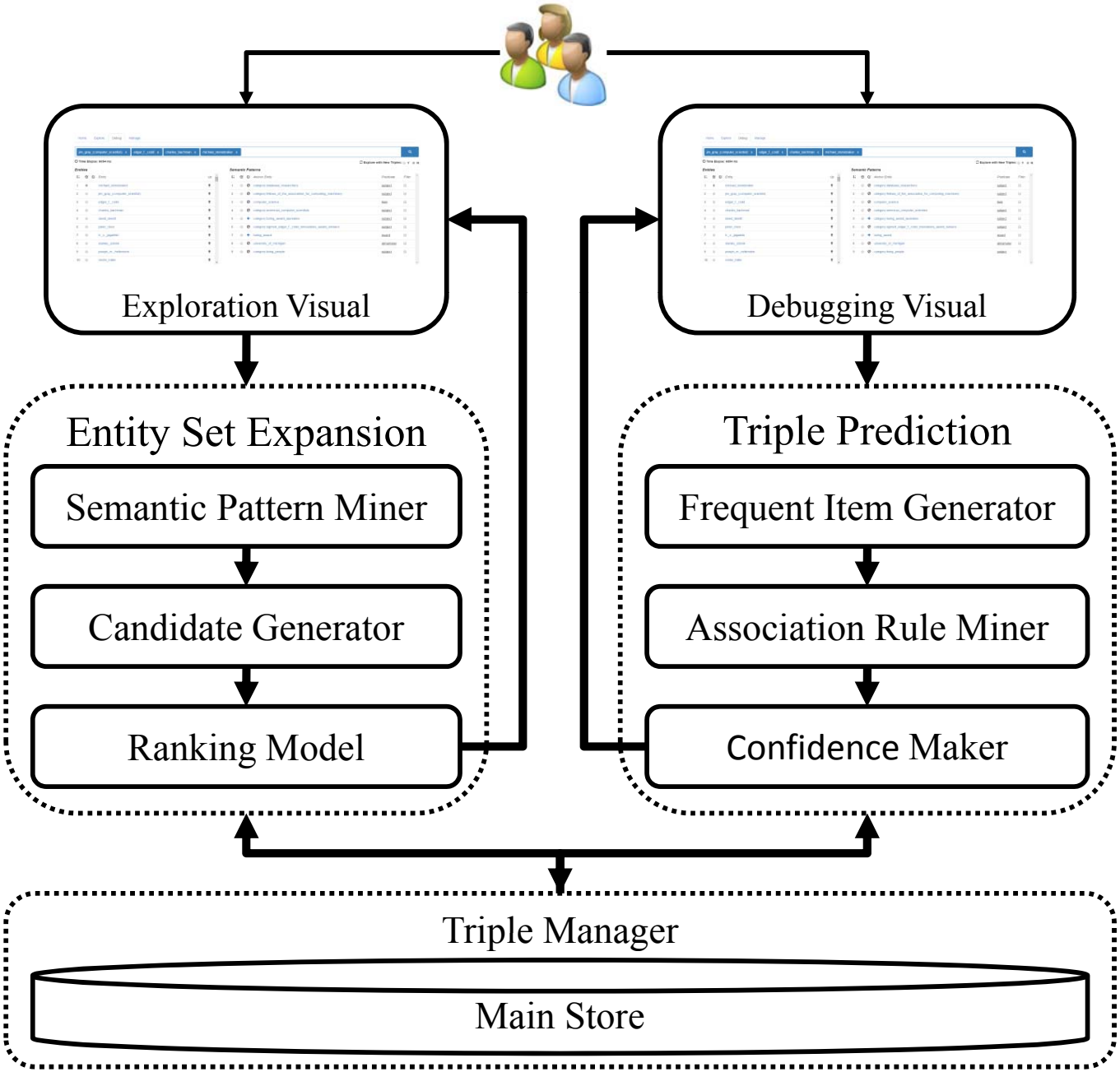
Table 1: Comparison on the QALD dataset

SLT	seeds	p@5	p@10	p@20	MRR	R-pre
SEAL	2	.377	.290	.208	.550	.269
BBR	2	.350	.307	.253	.478	.283
LDSD	2	.133	.100	.078	.273	.100
QBESS	2	.400	.353	.267	.545	.338
ARM	2	.503$^{\bullet}$	.420$^{\bullet}$	.322$^{\bullet}$	.661$^{\bullet}$	.375$^{\bullet}$
ESER	2	**.547**$^{\bullet}_{*}$	**.460**$^{\bullet}_{*}$	**.372**$^{\bullet}_{*}$	**.699**$^{\bullet}_{*}$	**.457**$^{\bullet}_{*}$
SEAL	3	.453	.363	.267	.591	.340
BBR	3	.467	.363	.287	.604	.337
LDSD	3	.103	.082	.063	.194	.084
QBESS	3	.410	.343	.266	.536	.359
ARM	3	.550$^{\bullet}$	.468$^{\bullet}$	.372$^{\bullet}$	.665$^{\bullet}$	.446$^{\bullet}$
ESER	3	**.613**$^{\bullet}_{*}$	**.498**$^{\bullet}_{*}$	**.387**$^{\bullet}_{*}$	**.773**$^{\bullet}_{*}$	**.501**$^{\bullet}_{*}$
SEAL	4	.420	.350	.270	.539	.354
BBR	4	.467	.365	.290	.654	.345
LDSD	4	.097	.077	.055	.252	.075
QBESS	4	.373	.290	.225	.465	.340
ARM	4	.527$^{\bullet}$	.430$^{\bullet}$	.348$^{\bullet}$	.716$^{\bullet}$	.420
ESER	4	**.613**$^{\bullet}_{*}$	**.502**$^{\bullet}_{*}$	**.392**$^{\bullet}_{*}$	**.801**$^{\bullet}_{*}$	**.525**$^{\bullet}_{*}$
SEAL	5	.410	.317	.247	.535	.352
BBR	5	.453	.362	.288	.627	.336
LDSD	5	.130	.083	.060	.284	.090
QBESS	5	.370	.285	.227	.488	.353
ARM	5	.503$^{\bullet}$	.418$^{\bullet}$	.342$^{\bullet}$	.665$^{\bullet}$	.426
ESER	5	**.563**$^{\bullet}_{*}$	**.465**$^{\bullet}_{*}$	**.381**$^{\bullet}_{*}$	**.726**$^{\bullet}_{*}$	**.515**$^{\bullet}_{*}$
SEAL	mix	.447	.347	.249	.592	.335
BBR	mix	.447	.370	.292	.570	.347
LDSD	mix	.130	.100	.069	.268	.100
QBESS	mix	.510	.417	.332	.630	.420
ARM	mix	.537$^{\bullet}$	.443$^{\bullet}$	.337$^{\bullet}$	.646$^{\bullet}$	.433$^{\bullet}$
ESER	mix	**.633**$^{\bullet}_{*}$	**.510**$^{\bullet}_{*}$	**.403**$^{\bullet}_{*}$	**.799**$^{\bullet}_{*}$	**.559**$^{\bullet}_{*}$

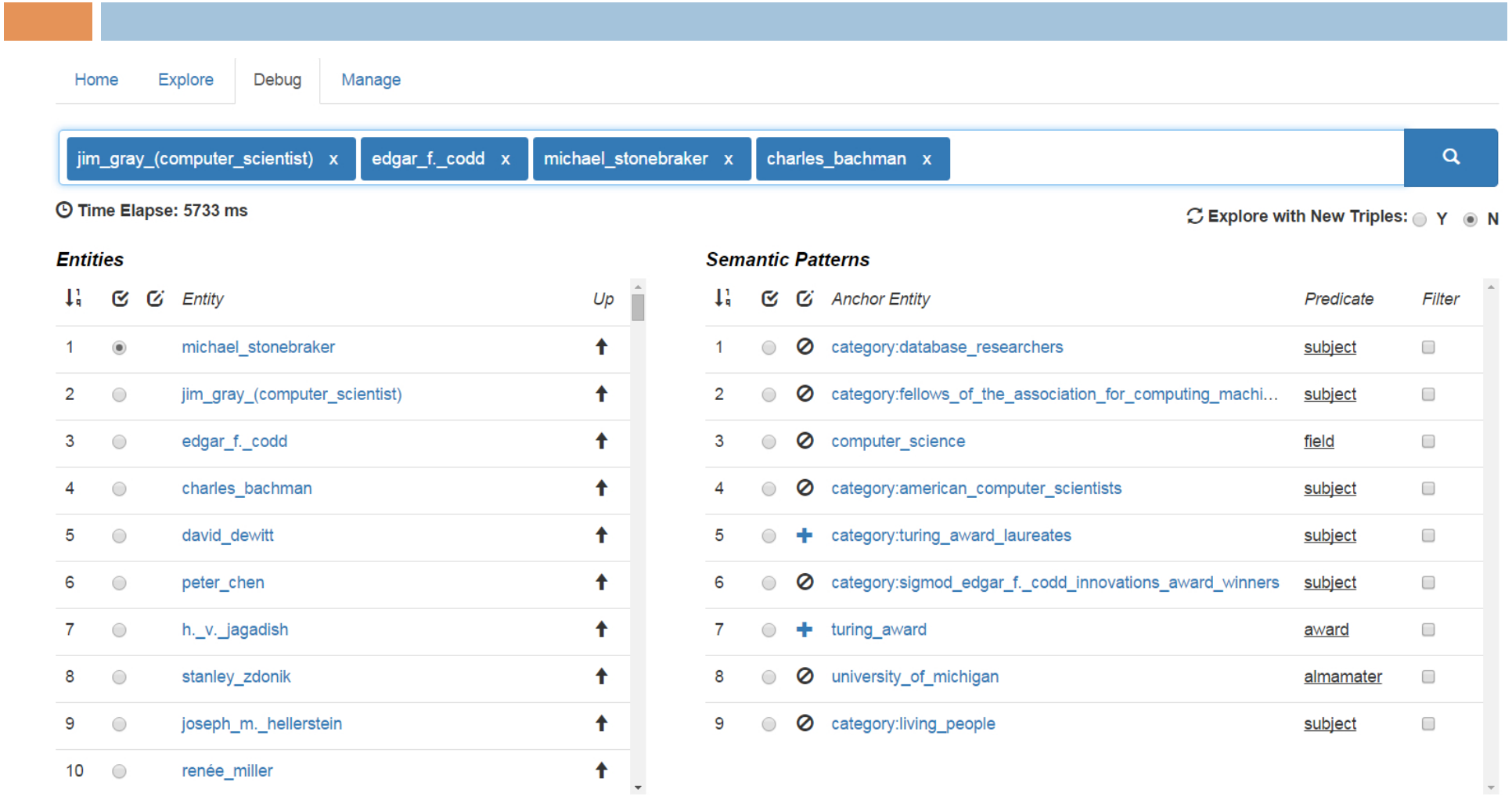
Table 2: Comparison on the INEX dataset

SLT	seeds	p@5	p@10	p@20	MRR	R-pre
SEAL	2	**.412**$_{\star}$	**.388**$_{\star}$	**.331**$_{\star}$	.542$_{\star}$	**.327**$_{\star}$
BBR	2	.312	.265	.200	.501	.208
LDSD	2	.092	.081	.063	.237	.072
QBESS	2	.246	.223	.186	.330	.202
ARM	2	.223	.217	.188	.367	.199
ESER	2	.400$_{\star}$	.383$_{\star}$	.287$_{\star}$	**.551**$_{\star}$	.304$_{\star}$
SEAL	3	.462$_{\star}$	**.433**$_{\star}$	**.354**$_{\star}$	.547$_{\star}$	**.377**$_{\star}$
BBR	3	.312	.288	.217	.506	.242
LDSD	3	.092	.065	.047	.247	.055
QBESS	3	.219	.194	.169	.310	.190
ARM	3	.188	.185	.168	.262	.187
ESER	3	**.500**$_{\star}$	.415$_{\star}$	.311$_{\star}$	**.684**$_{\star}$	.340$_{\star}$
SEAL	4	.423$_{\star}$	.383$_{\star}$	.319$_{\star}$	.530$_{\star}$	.339$_{\star}$
BBR	4	.312	.290	.239	.527	.252
LDSD	4	.088	.069	.057	.193	.067
QBESS	4	.181	.163	.144	.222	.161
ARM	4	.254	.233	.186	.345	.208
ESER	4	**.504**$_{\star}$	**.446**$_{\star}$	**.341**$_{\star}$	**.633**$_{\star}$	**.376**$_{\star}$
SEAL	5	.377	.340	.284	.418	.311
BBR	5	.346	.327	.257	.593	.282
LDSD	5	.092	.075	.059	.179	.069
QBESS	5	.173	.154	.135	.205	.153
ARM	5	.227	.227	.187	.322	.219
ESER	5	**.492**$^{\bullet}_{\star}$	**.433**$^{\bullet}_{\star}$	**.336**$^{\bullet}_{\star}$	**.629**$^{\bullet}_{\star}$	**.381**$^{\bullet}_{\star}$
SEAL	mix	.473$_{*}$	.398$_{*}$	.305	.644	.330
BBR	mix	.358	.315	.251	.597	.279
LDSD	mix	.112	.087	.062	.271	.063
QBESS	mix	.427	.367	.273	.579	.300
ARM	mix	.350	.304	.239	.535	.273
ESER	mix	**.515**$^{\bullet}_{\star}$	**.440**$^{\bullet}_{\star}$	**.350**$\star$	**.701**$\star$	**.409**$^{\bullet}_{\star}$

SEED


Exploration Visual
Debugging Visual
Entity Set Expansion
Semantic Pattern Miner
Candidate Generator
Ranking Model
Triple Prediction
Frequent Item Generator
Association Rule Miner
Confidence Maker
Triple Manager
Main Store

用户界面



ICDE'16 Demo

知识纠错

- Semantic patterns能够帮助我们预测知识缺陷
- Association rule mining方法的引入
  - Top相关实体作为transactions
  - 每个实体对应谓词集合作为一个transaction的items
  - 找到所有的频繁项X
  - 如果存在一个频繁项X能够imply一个谓词p, 则预测p可能是e的一个谓词
  - 同理，可以预测另一个实体e'是不是的e关联实体
  - 进而预测e是否满足一个SP：e':p，即<e, p, e'>是否成立

系统展示

- 前端：
  - Html5+CSS3+Javascript+Ajax
    - Bootsrap
    - Echarts
- 后端：
  - 类消息传递的模型
    - 可扩展
  - Mysql
- URL：
  - http://202.112.114.205:8080/SEED

总结与展望

- 总结
  - 老问题上的新方法
  - 性能优良
  - 扩展性强

- 展望
  - 加强纠错功能
    - 多版本控制
    - 内部错误挖掘
  - 提升查询性能
    - 速度
    - 准确度

- 谢谢关注！

- 陈跃国
- chenyueguo@ruc.edu.cn