基于众包的非确定知识图谱清洗

林欣
华东师范大学计算机系

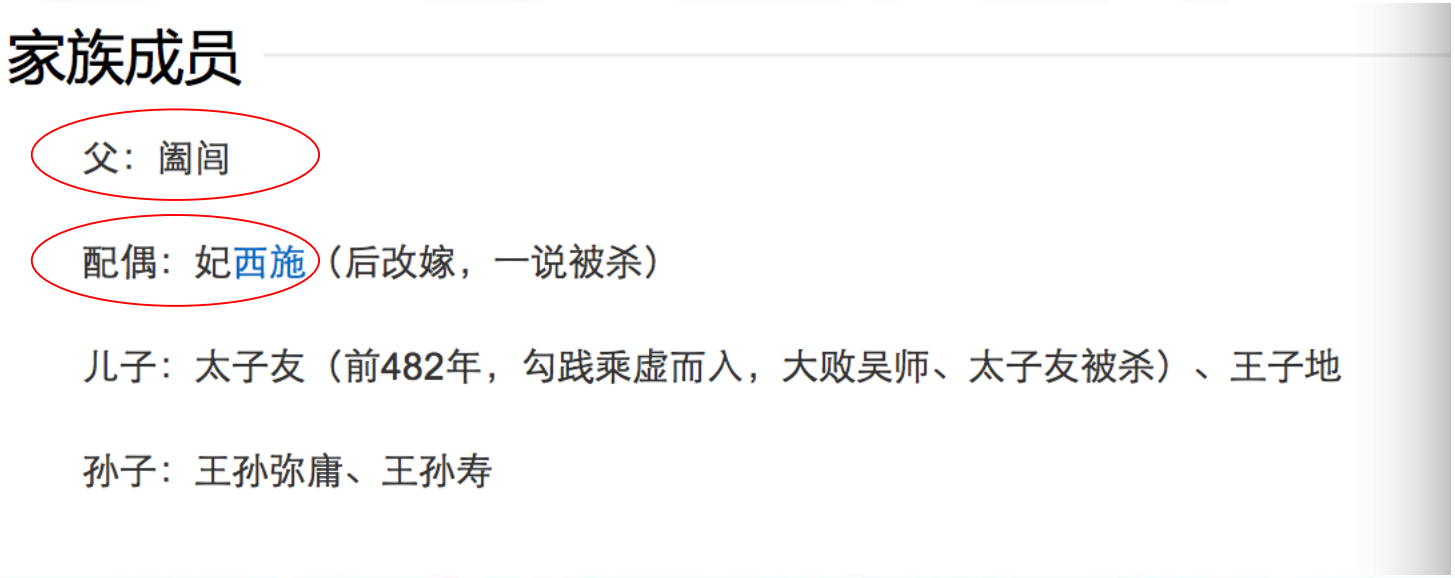
知识图谱构建

- 四种常见方法 [KDD 14]
  - Wiki-based： YAGO、Dbpedia、FreeBase
  - Ontology-based: NELL
  - Open IE: TextRunner, Reverb
  - Taxonomies (is-a): Probase

KDD14. X.Dong et.al. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion

Wiki-Based方法

- 准确率高
- 知识结构性好



家族成员

父：阖闾

配偶：妃西施（后改嫁，一说被杀）

儿子：太子友（前482年，勾践乘虚而入，大败吴师、太子友被杀）、王子地

孙子：王孙弥庸、王孙寿

Wiki-Based 方法

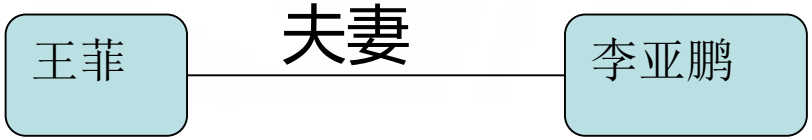
- 缺点
  - 知识长尾效应
  - 覆盖度有限
  - FreeBase中 71%人物没有“出生地”，75%的人物没有国籍

自动挖掘方法

- 知识覆盖度大
- 噪音大、知识准确率难保证

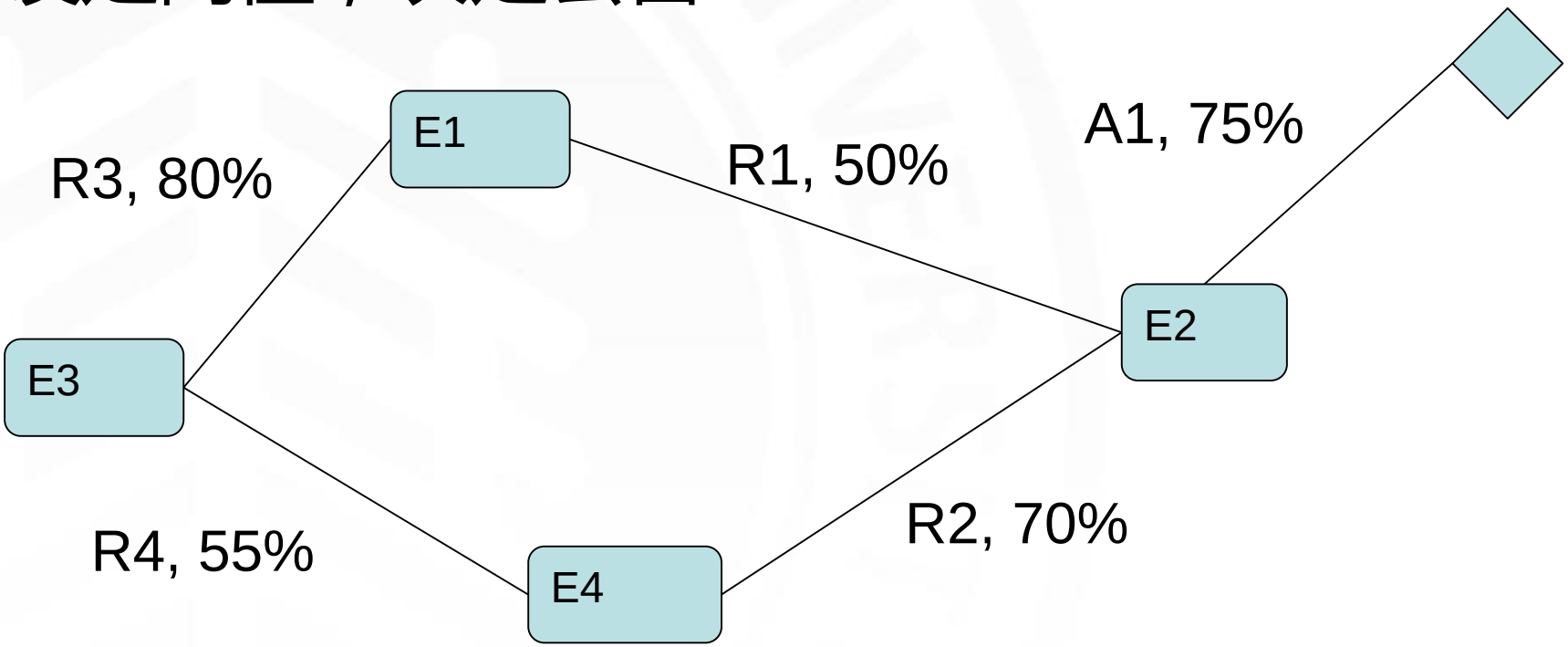
王菲李亚鹏结婚照首度曝光 天山拍摄简单的甜蜜[组图]

2011-12-07 08:01:00 来源: 中国娱乐网 有0人参与 手机看新闻 转发到微博(0)



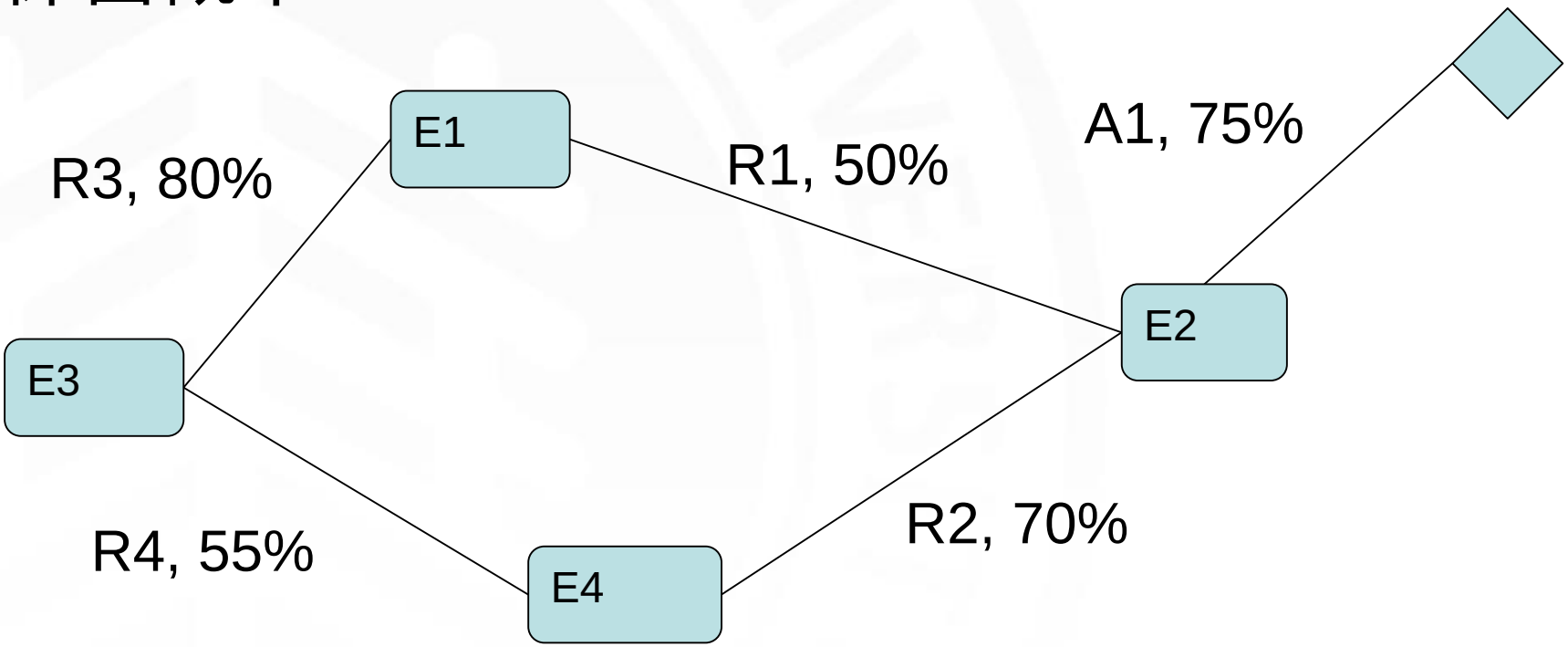
传统非确定性数据解决办法

- 设定阈值，决定去留



非确定知识图谱

- 保留概率

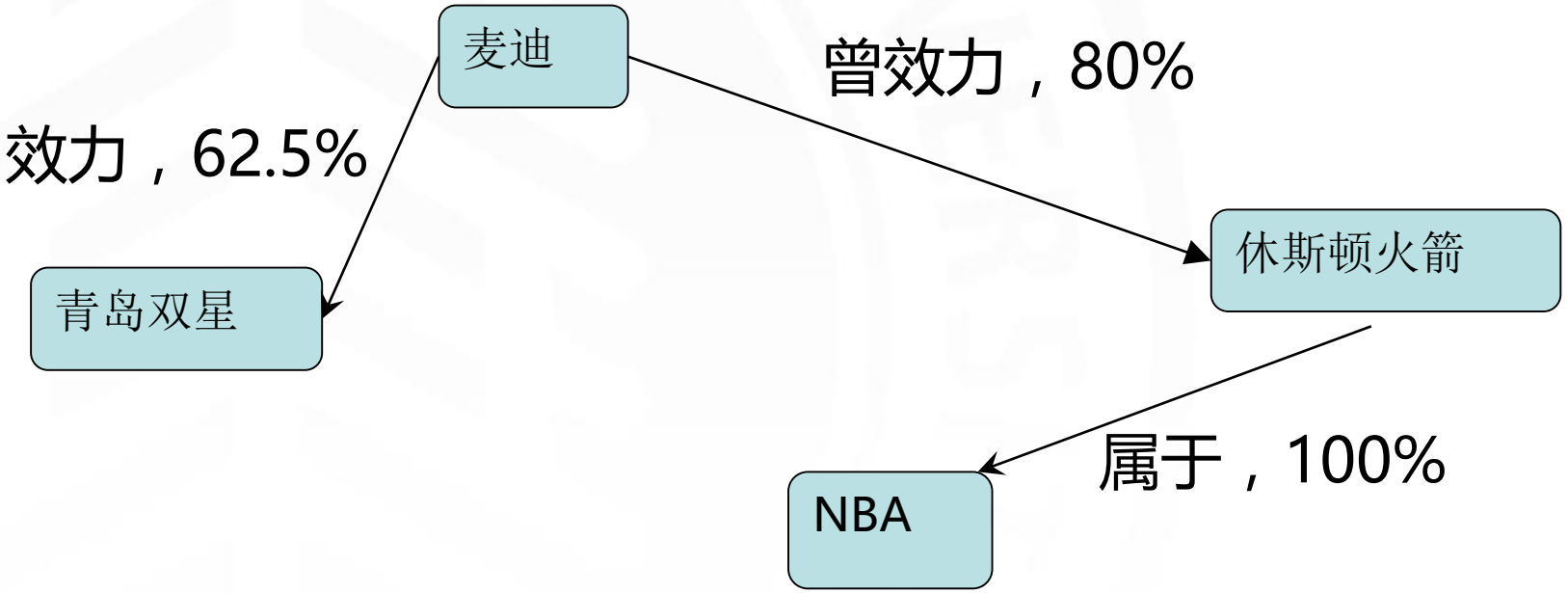


非确定知识图谱

- 优点：
  - 知识覆盖度
  - 信息更丰富（Provenance）
- 缺点
  - 查询质量（Ambiguity)
  - 0.5 问题

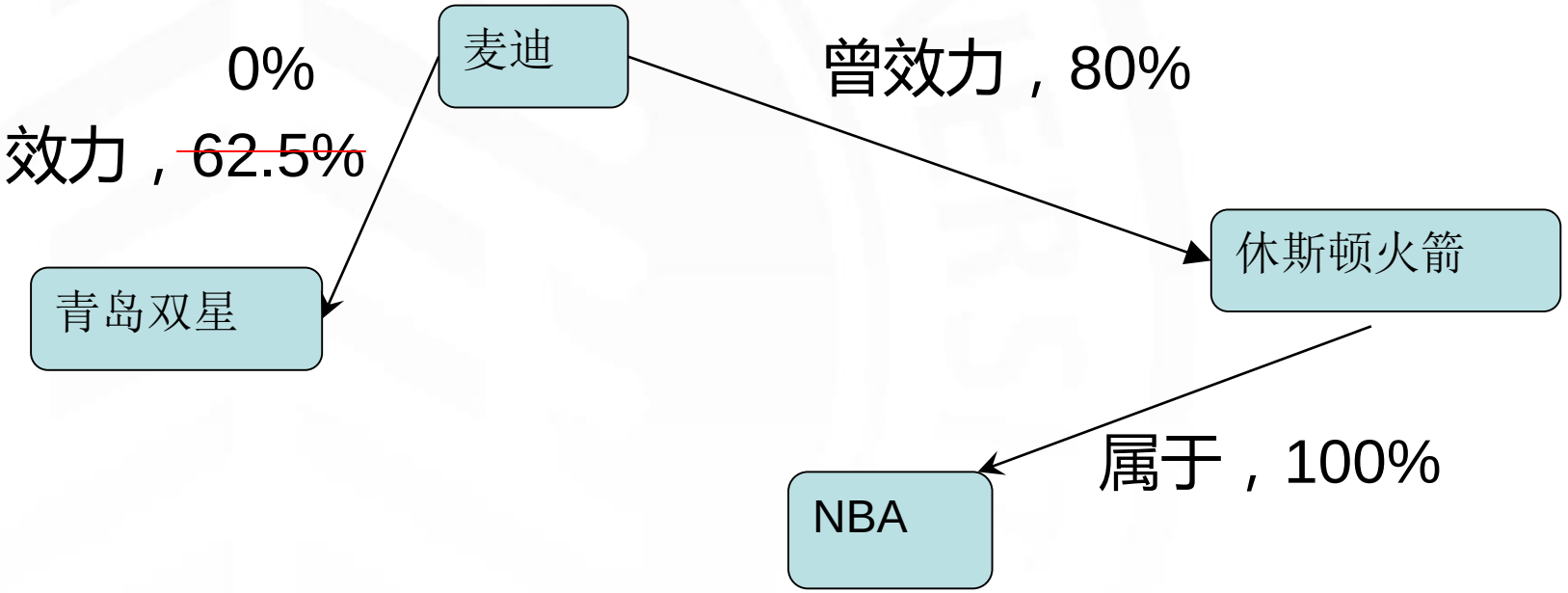
0.5问题

- “现在在中国打球的前NBA球员？”



0.5问题

- 怎么办？ 概率清洗



机器还不够聪明，且还会不够聪明很久

众包清洗，用外脑的清洗架构

机器+人脑

双低

机器 + 机器

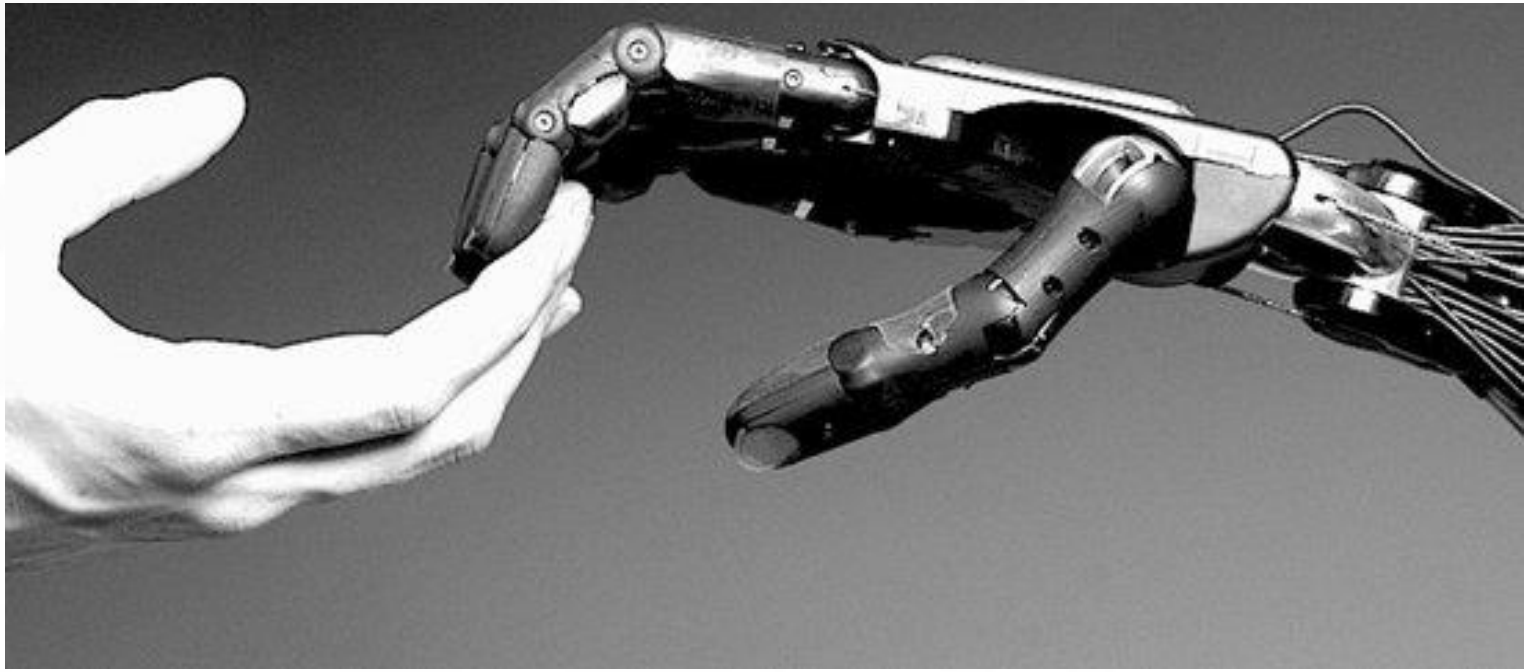
双高

人脑+人脑

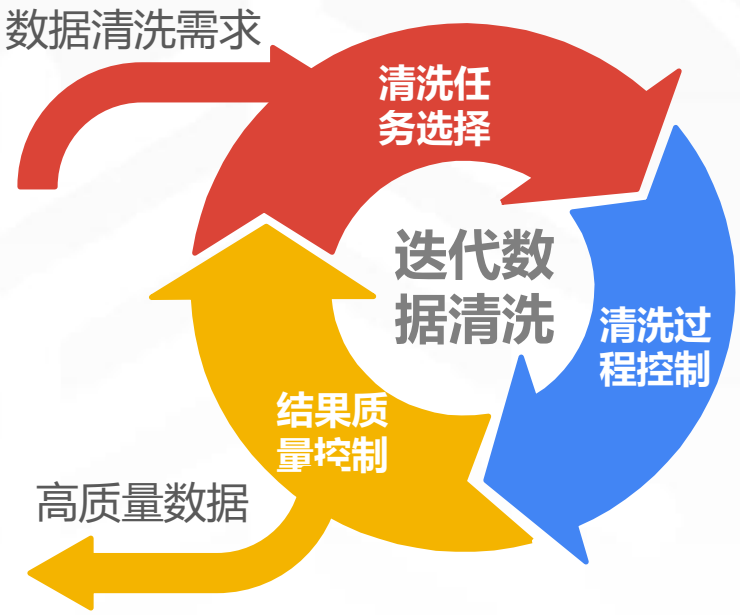
代价

质量

“众包清洗的科学问题”



问题1：基于众包的迭代清洗机制



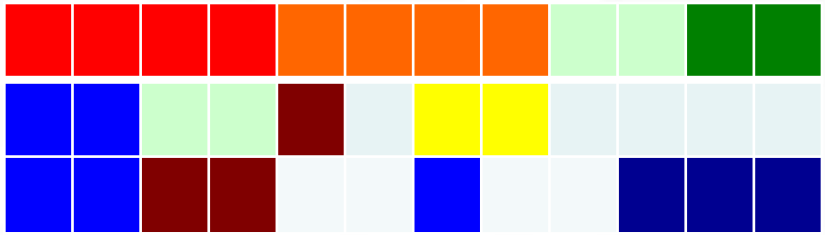
问题描述：

数据清洗的过程是不断提高数据质量的过程，而这种提升往往不能通过一次清洗而到位。正确的过程是通过若干次群智计算与机器学习的交织迭代，不断提高数据质量。

研究重点：

1. 迭代终止条件（收敛程度监测？）
2.迭代过程自动化（包括子步骤间的衔接）

问题2：清洗收益的高效算法



问题描述：

众包更擅长清洗较小的数据，如识别两张图片，识别一个单词。而小数据之间存在关联，清洗每份小数据对全局质量提升量不同（如左图），需要将有限的资源投放给清洗收益最高的小数据。

研究重点：

1. 单个小数据清洗收益算法；
2. 多个关联小数据清洗综合收益算法

Reducing Uncertainty of Schema Matching via Crowdsourcing. C.Zhang, *et.al*, *VLDB*,2013

问题3：众包决策的正确性判决

问题描述：

系统无法控制众包返回的结果，因此，每个工人的结果可能差异性较大，而上层应用只需要众包平台通过正确性判决返回统一的结果。

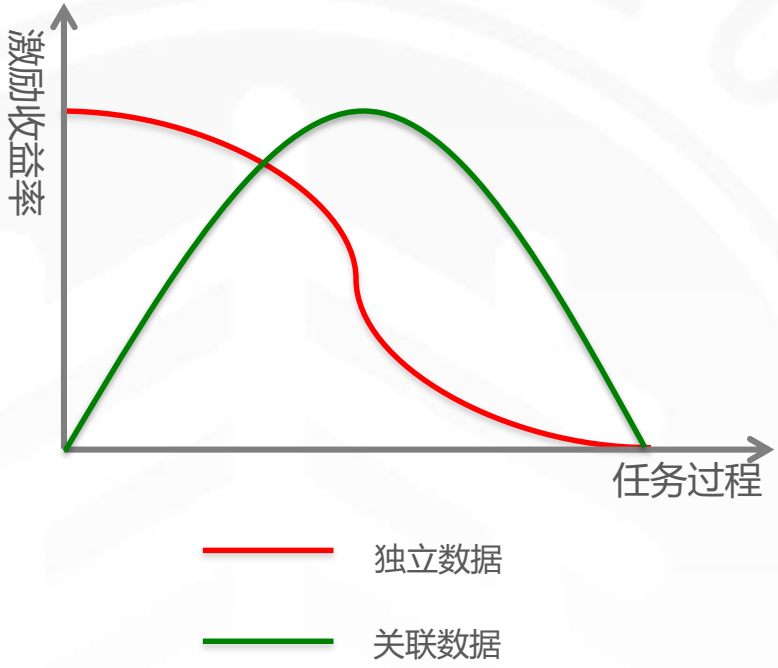
研究重点：

1. 工人建模分析：包括使用工人背景、众包提交历史记录、互评价等数据，为清洗筛选工人或为工人建模；

2. 结果正确性判断：综合所有返回的结果、工人错误率等模型，获得最终结果

Entity Resolution with Crowd Errors. V.Verroios, et.al , ICDE,2015

问题4：动态激励机制研究



问题描述：

众包的激励机制（如金钱、积分）是工人完成任务的动力，越多激励投入，任务收益也越大。然而在任务进行过程中，激励的收益率是动态的，也和具体数据相关，如独立性数据收益率会越来越小，关联性数据收益率呈抛物线。因此，需要研究如何利用最少的代价获得最多的清洗收益。

研究重点：

1. 动态激励收益率计算方法；
2. 最佳激励分配机制的近似算法；

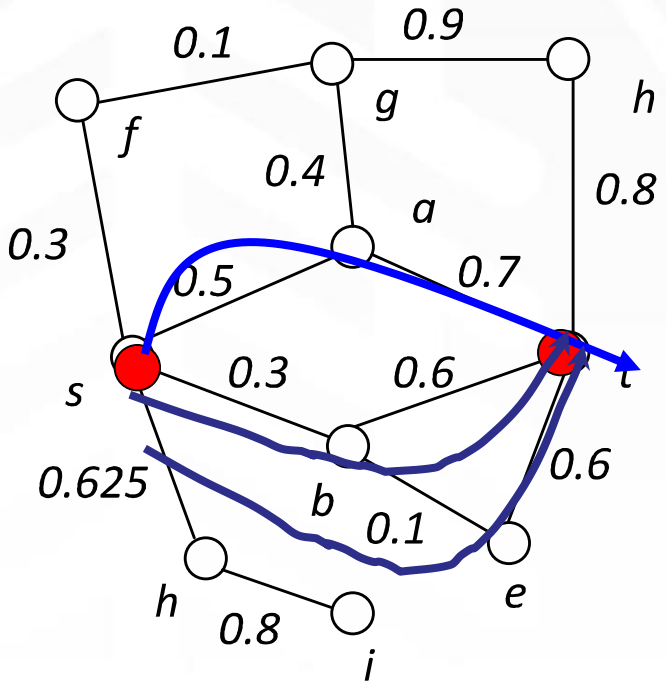
Financial Incentives and the ‘Performance of Crowds’. W.Mason, et.al , SIGKDD Explorations News letter,2010

我们的工作

- 1. 众包关系选取
- 2. 众包平台开发

众包关系选取

- 特定查询优化（可达性查询）



Q: 点s和t是否3跳可达？

A: P= 0.47

目标：选一条边清洗，使得清洗后的答案的不确定性最低（靠近0或者1）

问题定义

- 不确定性度量
  - 熵理论
  - H(A) = - P(A)logP(A) – (1-P(A)) log(1-p(A))
  - 清洗后的熵$H_{e=1}(A)$，$H_{e=0}(A)$
- 期望熵
  - E = P(A) $H_{e=1}(A)$+(1-P(A))$H_{e=0}(A)$

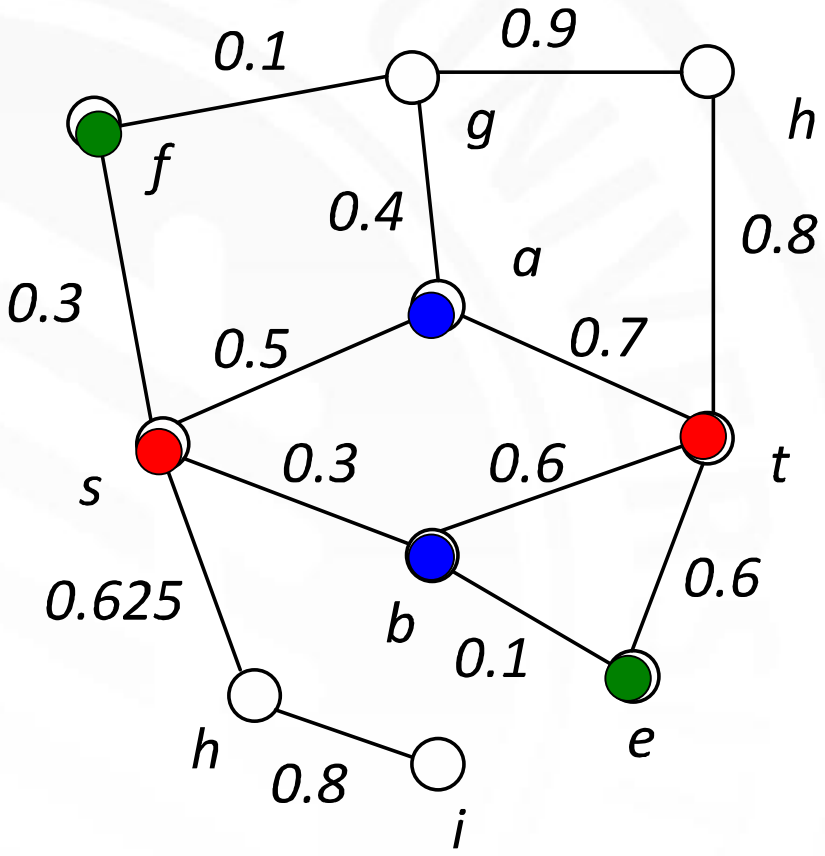
基本思路

- P*因子
  - 所有过e的路径连通，而不过e的路径都不连通的概率
  - 证明：P*与E正相关
- P*快速排序算法

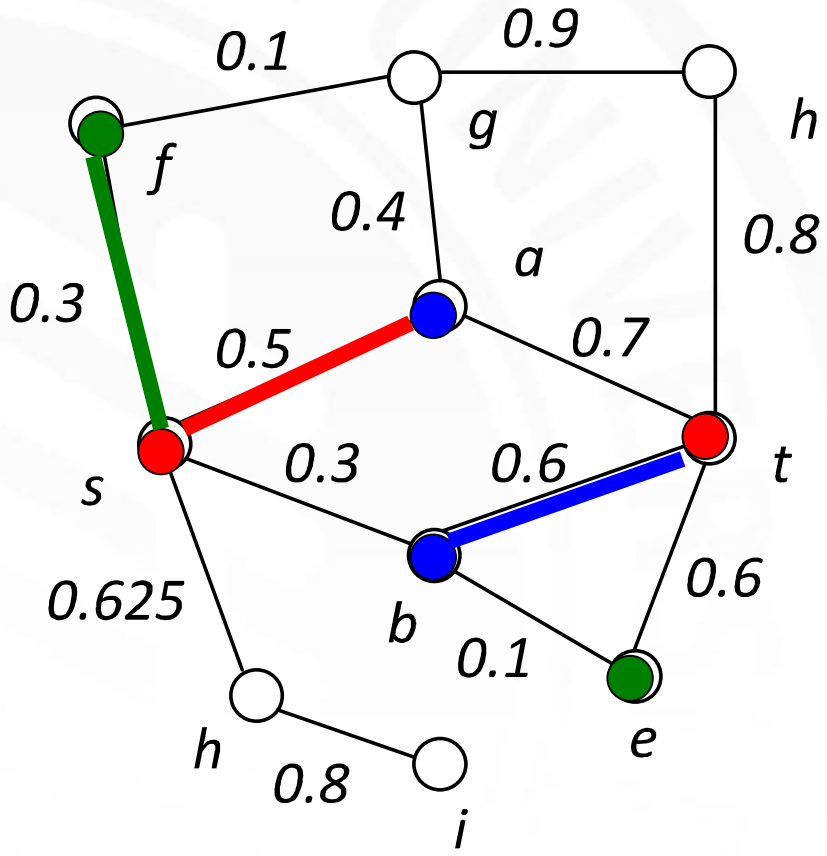
问题扩展

- 全图清洗
- 多边清洗
- Noisy Crowd
- 其他布尔查询（Connectivity、K-core）
- 其他枚举型查询（Subgraph Matching）

全图清洗



多边清洗



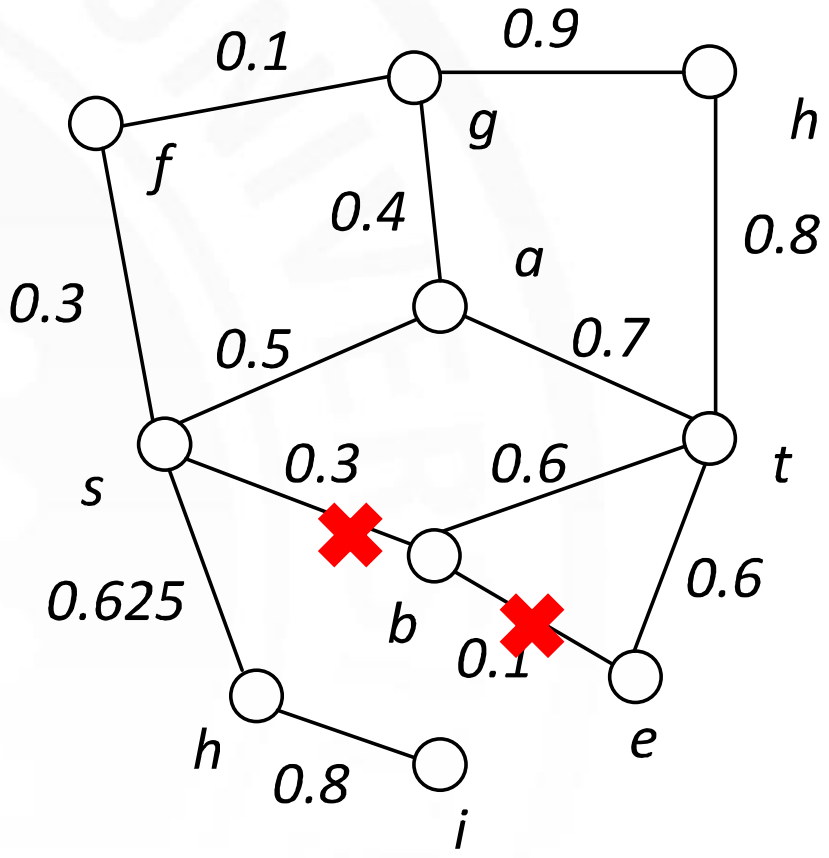
Noisy Crowd

- 众包并不能保证完全正确 [ICDE15]
- 假设Crowd返回的结果和真实值有一定偏差
- 在此假设下选取最优边

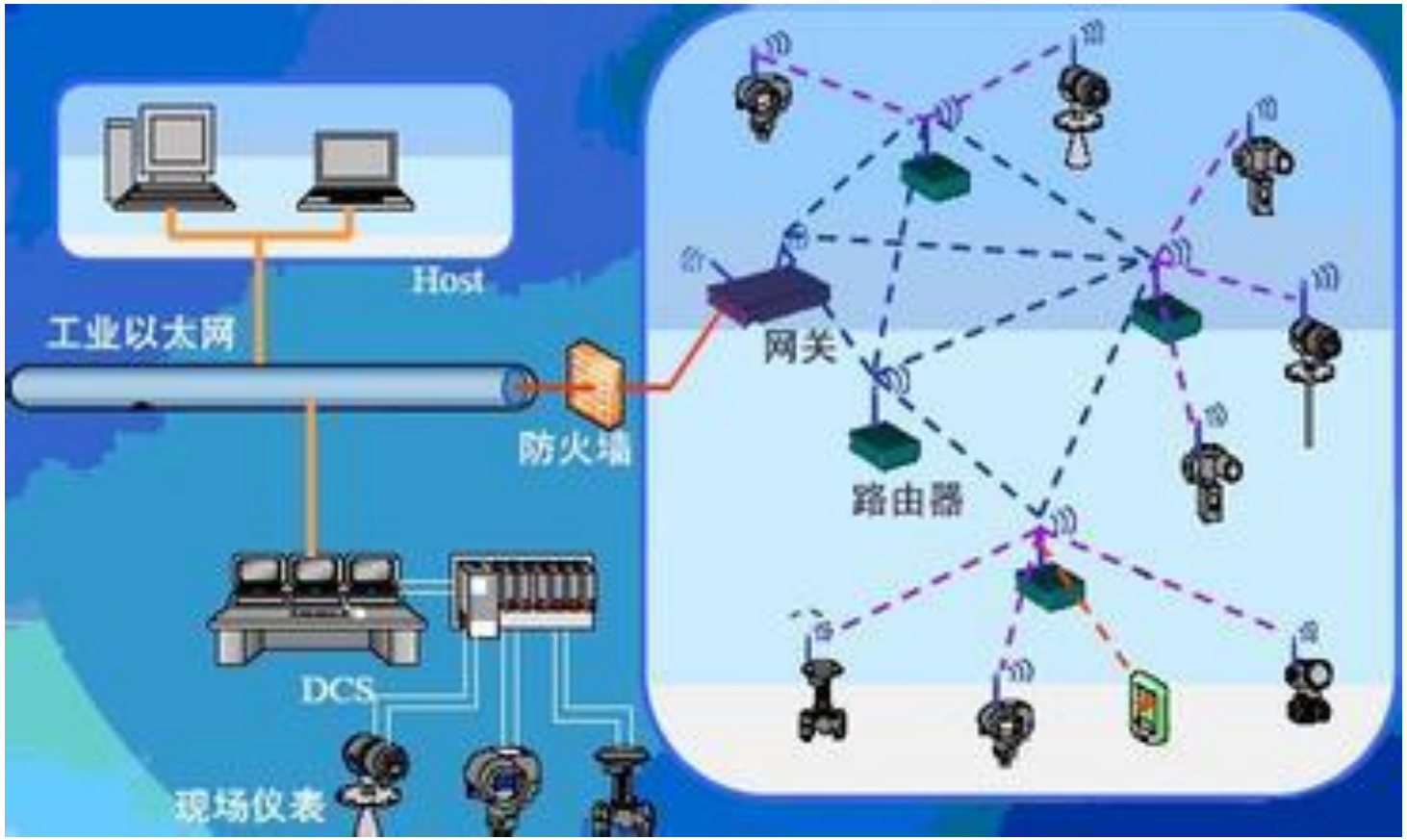
ICDE15. C. zhang et.al. Cleaning Uncertain Data with a Noisy Crowd

其他布尔查询清洗

- 连通性查询
  - 2-Connectivity
- K-Core



应用（无线传感器网络）



枚举型查询的清洗

- 不仅仅只有Yes & NO两种答案
- 每种答案都有概率，组成了信息熵
- SPARQL 查询： 找出所有在中国打球的前NBA球员

众包平台搭建


众包平台
——改变您过去的生活方式


众包系统
登录/注册模块
工作模块
统计模块
注册/登录
微信绑定
测试/工作模式
快速做题
工作推荐
任务发布
质量控制
任务进度监督
任务结果报表


工作统计

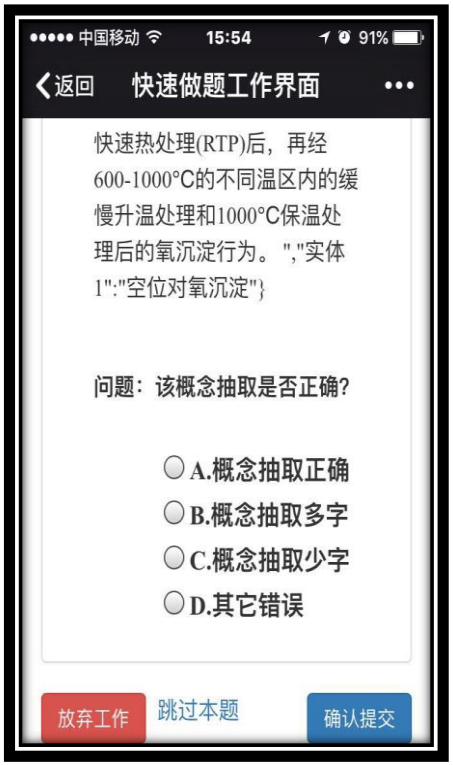

众包平台
嗨，你是哪位？
发布者
工作者

移动端界面


中国移动 15:52 91%
返回
绑定账户
众包平台
已有账号:
立即绑定
还未注册:
立即注册


中国移动 15:53 91%
返回
工作推荐:
序号 名称 元/题 题数
6 知识图谱测试数据 0.1 100
7 知识图谱内部测试数据1 0.05 100
8 多人做题测试 0.1 17
16 知识图谱任务九 0.05 50


中国移动 15:53 91%
返回
快速做题工作介绍
工作详细介绍
注出来)。
任务: 请您通过所学的知识或者查阅相关资料判断出概念的抽取是否正确。一共有四个答案可供选择:
A.概念抽取正确;
B.概念抽取多字;
C.概念抽取少字;
D.其它错误。;
下面是我们给您准备的一些实例:
快速做题数量: 100题
每题报酬: 0.1元
需用时: 5分钟
开始做题


中国移动 15:54 91%
返回
快速做题工作界面
快速热处理(RTP)后，再经600-1000°C的不同温区内的缓慢升温处理和1000°C保温处理后的氧沉淀行为。 ","实体1":"空位对氧沉淀"}
问题: 该概念抽取是否正确?
A.概念抽取正确
B.概念抽取多字
C.概念抽取少字
D.其它错误
放弃工作
跳过本题
确认提交

Thank you!