

《知识图谱: 概念与技术》

第6讲 众包构建

基于众包的知识图谱构建与精化

林欣

华东师范大学

xlin@fudan.edu.cn

本章大纲

- 知识型众包的基本概念
- 知识型众包的研究问题
- 众包在知识图谱构建与精化过程的作用

知识型众包的基本概念

众包的基本概念

- 众包 (crowdsourcing) ——群众外包
- 互联网时代，利用大量的网络用户来获取需要的服务
- 众包的优势：价格低廉、灵活



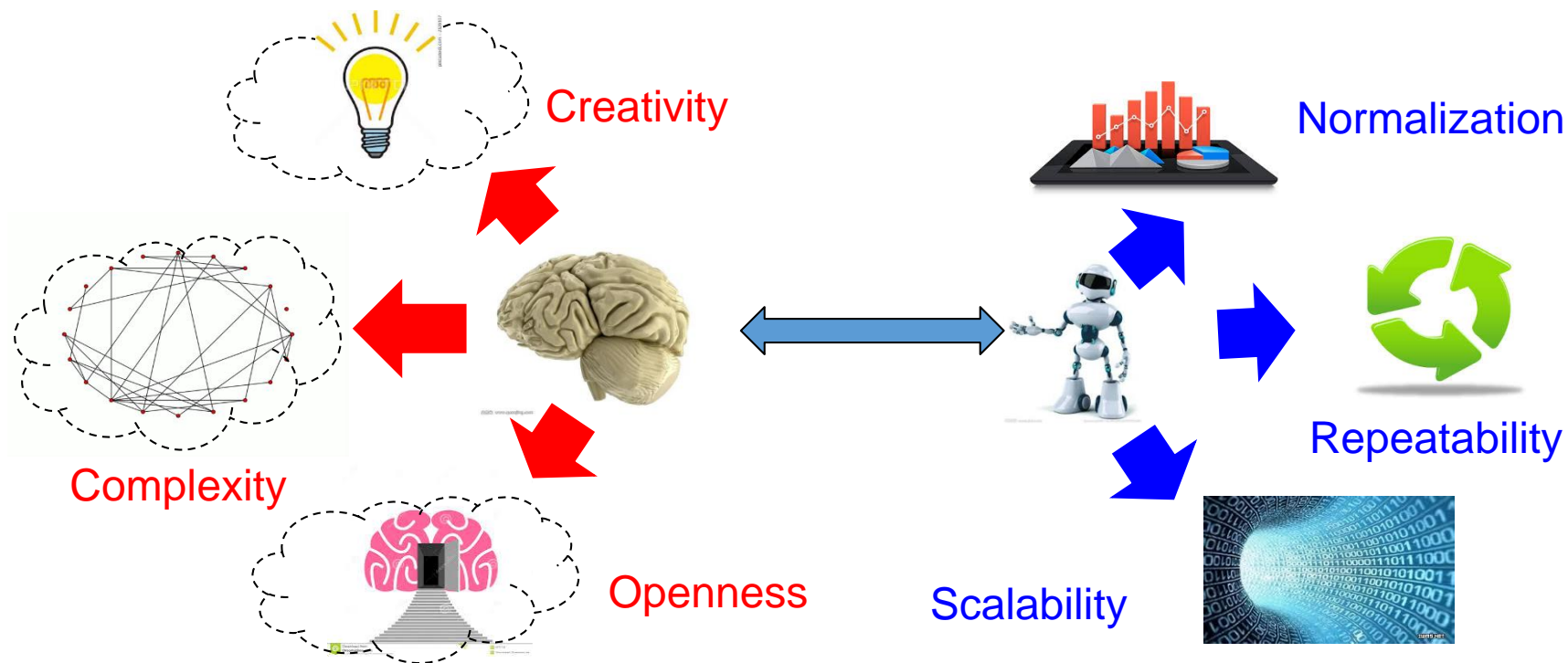
传统众包特点

- 任务单一
- 任务评价方法简单
- 工人要求单一
- 工人门槛较低
- 工人数量相对较多

传统众包的核心问题是优化任务与工人的匹配，提高用户体验度！

知识型众包

- 任务特点：与知识相关
- 是众包的一个分支
- 是沟通机器与人脑之间的桥梁



知识型众包应用特例



reCAPTCHAAs



ImageNet Labeling



Amazon MTurk

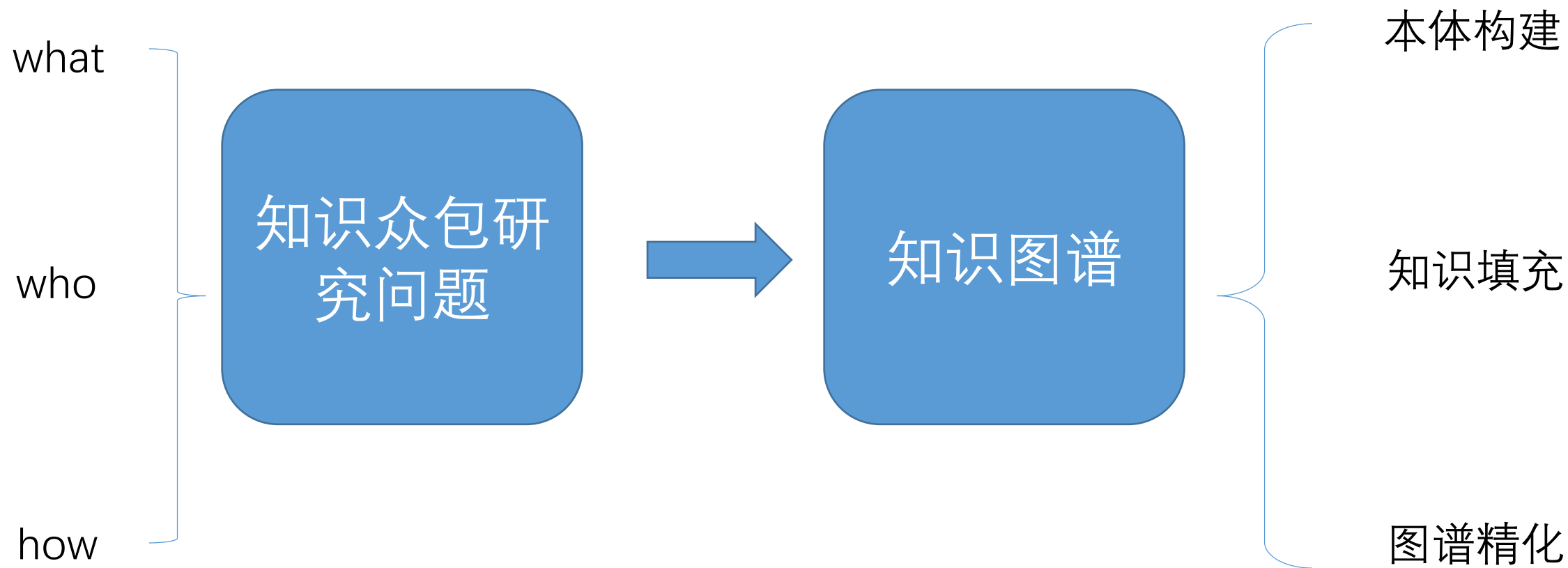


知识型众包已经成为知识收集、数据标注的重要手段

知识型众包特点

- 任务多样性强
 - 数据：图像、文本
 - 任务类型：情感分析、搜索结果排序、数据标注、数据分类、图像或音视频标注
 - 难易程度
- 工人多样性强
 - 文化程度
 - 投入程度
 - 专业领域
 - 完成任务动机
- 任务质量难以评价
 - 没有Ground-truth
 - 很难衡量工人置信度
 - 评价本身的花费高
- 任务质量影响较大
 - 错误知识隐藏较深，很难被定位出
 - 知识推理会扩大错误

授课大纲



知识型众包的研究问题

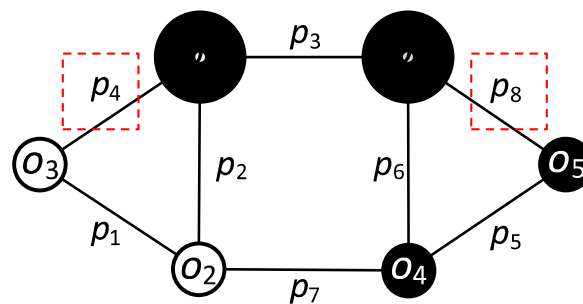
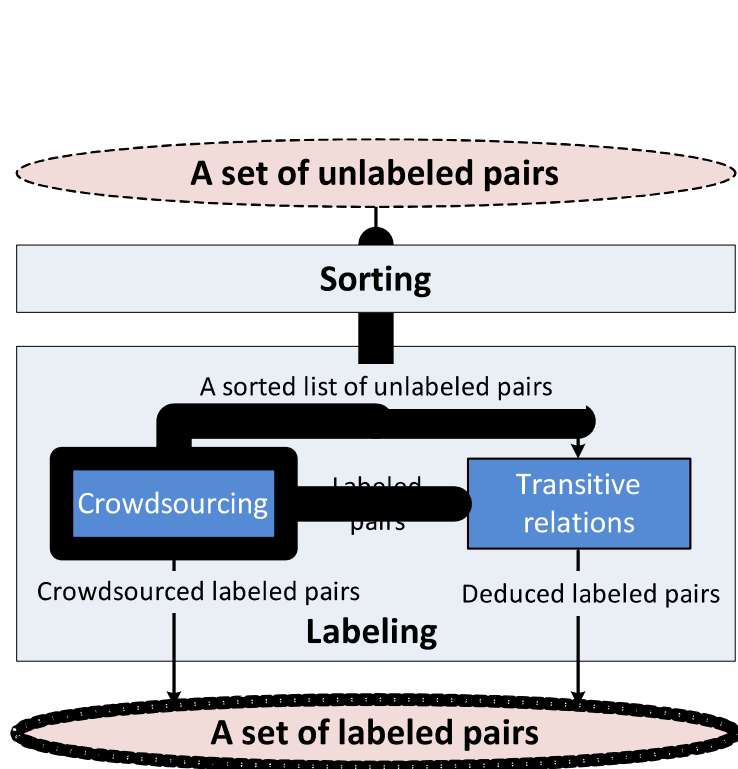
知识型众包研究问题

- 将什么任务交给众包 (What)
- 如何筛选工人 (Who)
- 如何完成众包 (How)
 - 如何设计问题
 - 如何激励工人
 - 如何控制质量
 - 如何最大化利用众包

将什么任务交予众包

- 任务选择
 - 目的：节约金钱与时间
 - 选择最重要的任务
 - 选择人最擅长而计算机不擅长的任务
- 与知识图谱相关的任务选择
 - 实体匹配
 - 本体匹配

利用众包进行实体匹配[SIGMOD13]

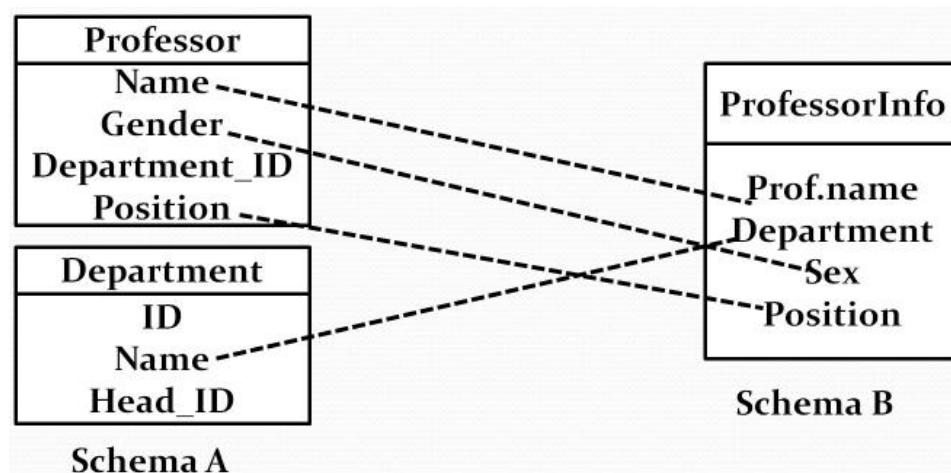


ID	Object
o_1	iPhone 2nd Gen
o_2	iPhone Two
o_3	iPhone 2
o_4	iPad Two
o_5	iPad 2
o_6	iPad 3rd Gen

ID	Object Pairs	Likelihood
p_1	(o_2, o_3)	0.85
p_2	(o_1, o_2)	0.75
p_3	(o_1, o_6)	0.72
p_4	(o_1, o_3)	0.65
p_5	(o_4, o_5)	0.55
p_6	(o_4, o_6)	0.48
p_7	(o_2, o_4)	0.45
p_8	(o_5, o_6)	0.42

合理调整众包任务的顺序和知识推理，降低众包开销！

本体匹配[VLDB15]

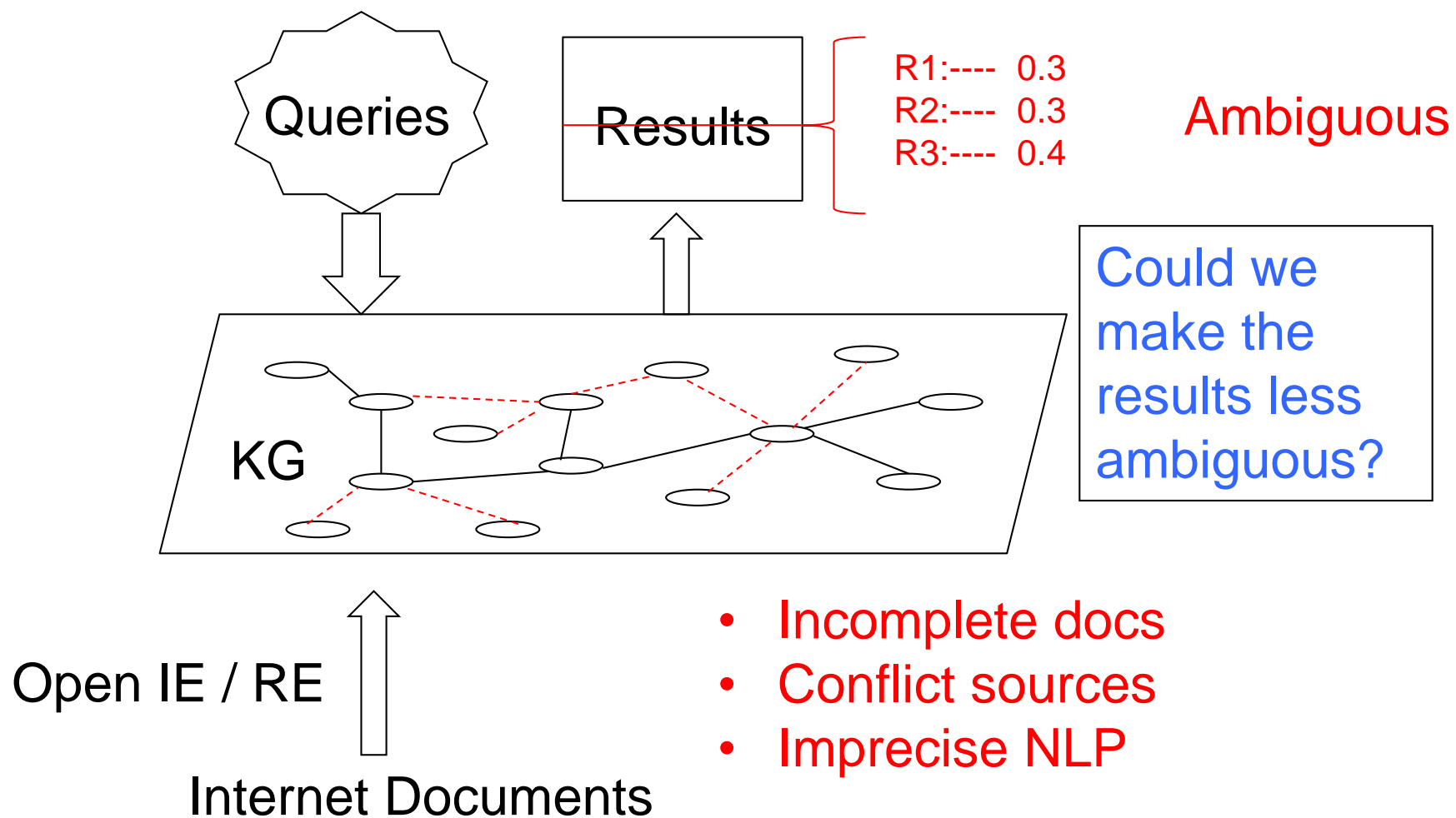


Correspondence	probability
$c_1 = \langle (\text{Professor})\text{Name}, \text{Prof.name} \rangle$.75
$c_2 = \langle \text{Position}, \text{Position} \rangle$.7
$c_3 = \langle \text{Gender}, \text{Sex} \rangle$	1
$c_4 = \langle (\text{Department}) \text{Name}, \text{Department} \rangle$.75
$c_5 = \langle (\text{Department})\text{Name}, \text{Prof.name} \rangle$.25

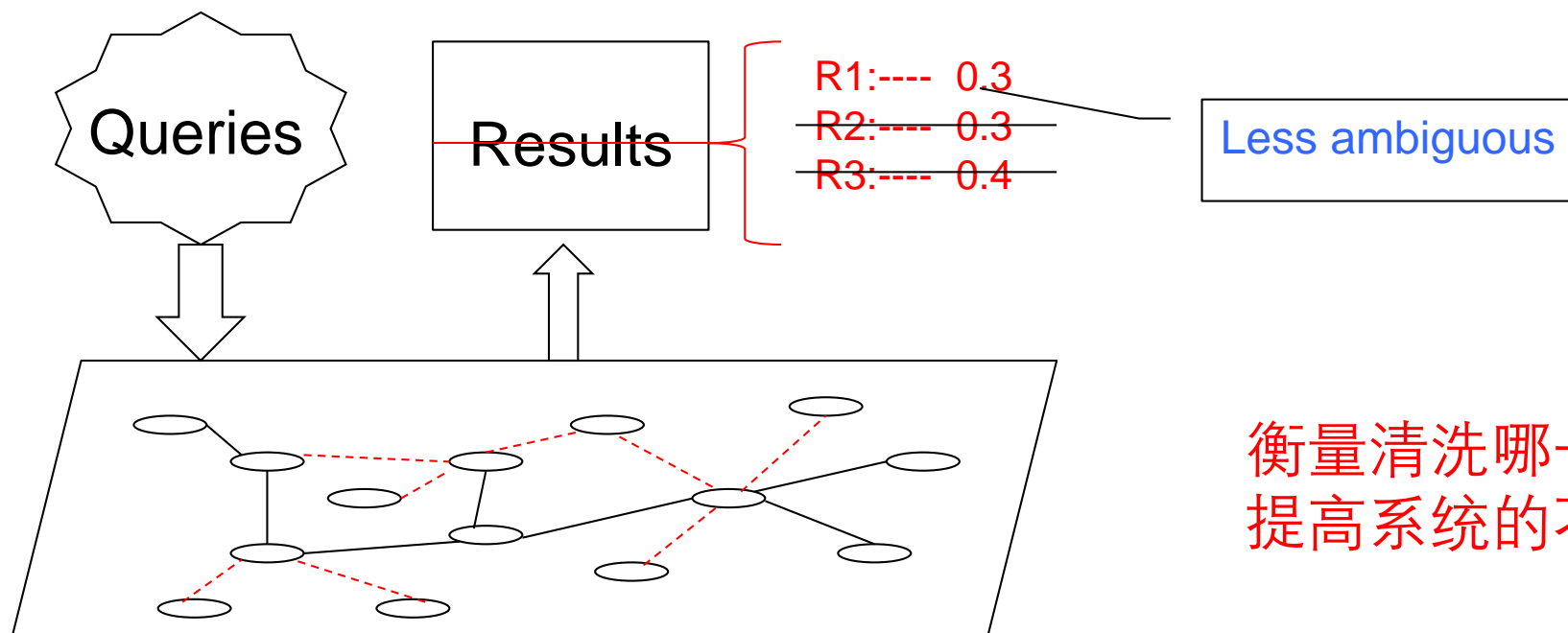
Possible Matchings	probability
$m_1 = \{ \langle (\text{Professor})\text{Name}, \text{Prof.name} \rangle, \langle \text{Position}, \text{Position} \rangle, \langle \text{Gender}, \text{Sex} \rangle, \langle (\text{Department}) \text{Name}, \text{Department} \rangle \}$.45
$m_2 = \{ \langle (\text{Professor})\text{Name}, \text{Prof.name} \rangle, \langle \text{Gender}, \text{Sex} \rangle, \langle (\text{Department}) \text{Name}, \text{Department} \rangle \}$.3
$m_3 = \{ ((\text{Department})\text{Name}, \text{Prof.name}), (\text{Position}, \text{Position}), (\text{Gender}, \text{Sex}) \}$.25

利用概率计算定位出最需要给众包确定的对应关系，来判断哪一组对应概率最大

知识图谱清洗 [TKDE17]



知识图谱清洗



衡量清洗哪一条边会最大程度提高系统的不确定性



众包任务选取原则

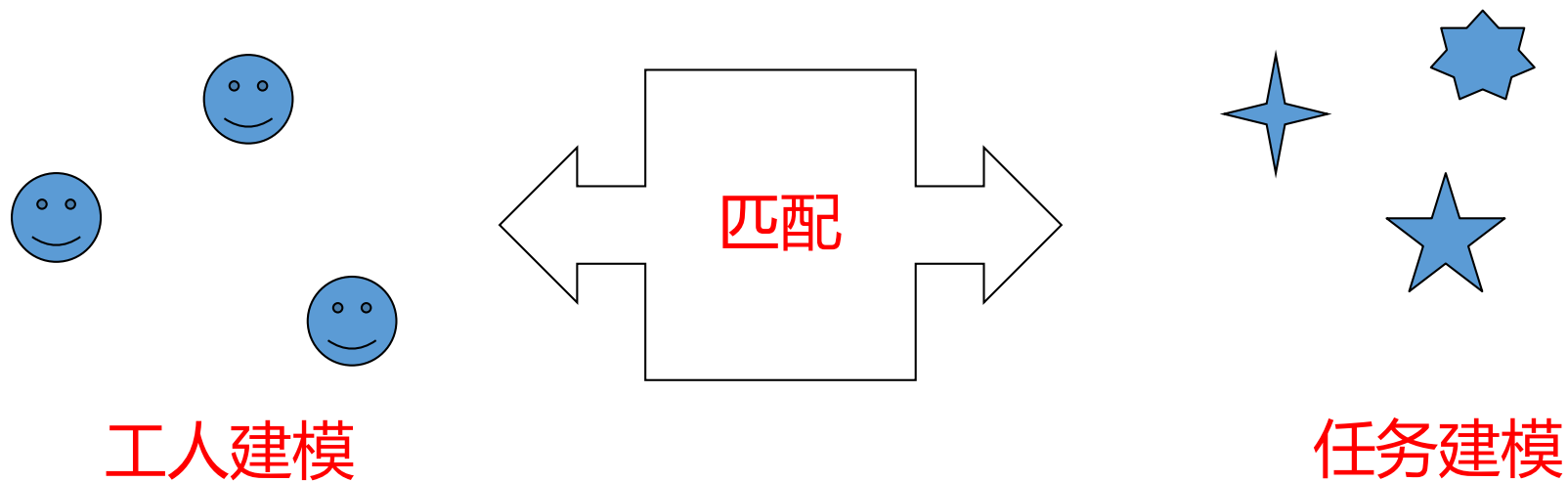
- 知识型众包偏爱小任务
 - 利用碎片化时间、快速收到报酬
- 局部的众包结果会对全局产生影响
- 需要量化这种影响
 - 量化模型可能很复杂
- 不同的任务影响不同
 - 因此对不同任务量化影响是数据管理领域关注的热点

知识型众包研究问题

- 将什么任务交给众包 (What)
- 如何筛选工人 (Who)
- 如何完成众包 (How)
 - 如何设计问题
 - 如何激励工人
 - 如何控制质量
 - 如何最大化利用众包

两种众包工人选择方法

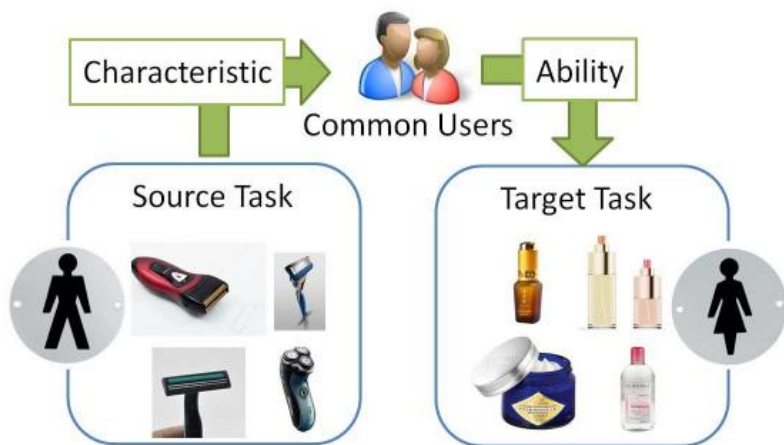
- 被动众包
 - 所有任务由工人方发出选取
 - 工人在正式工作前可能会参与一些技能测试
- 主动众包



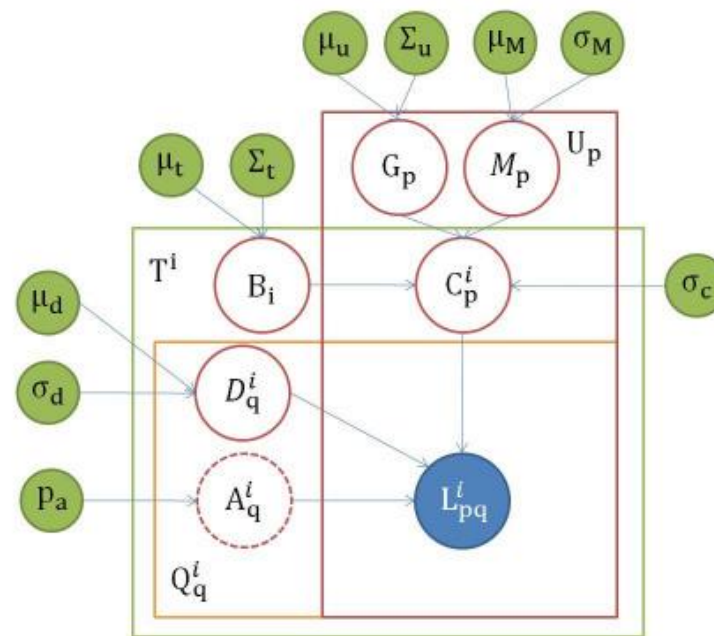
主动众包

- 任务分配
 - 随机分配
 - 按照其他因素排序（时间、工人质量等）
 - 寻找质量最高的工人[SIGMOD 15]
 - 寻找结果预期最有效的工人
 - 寻找最近的工人

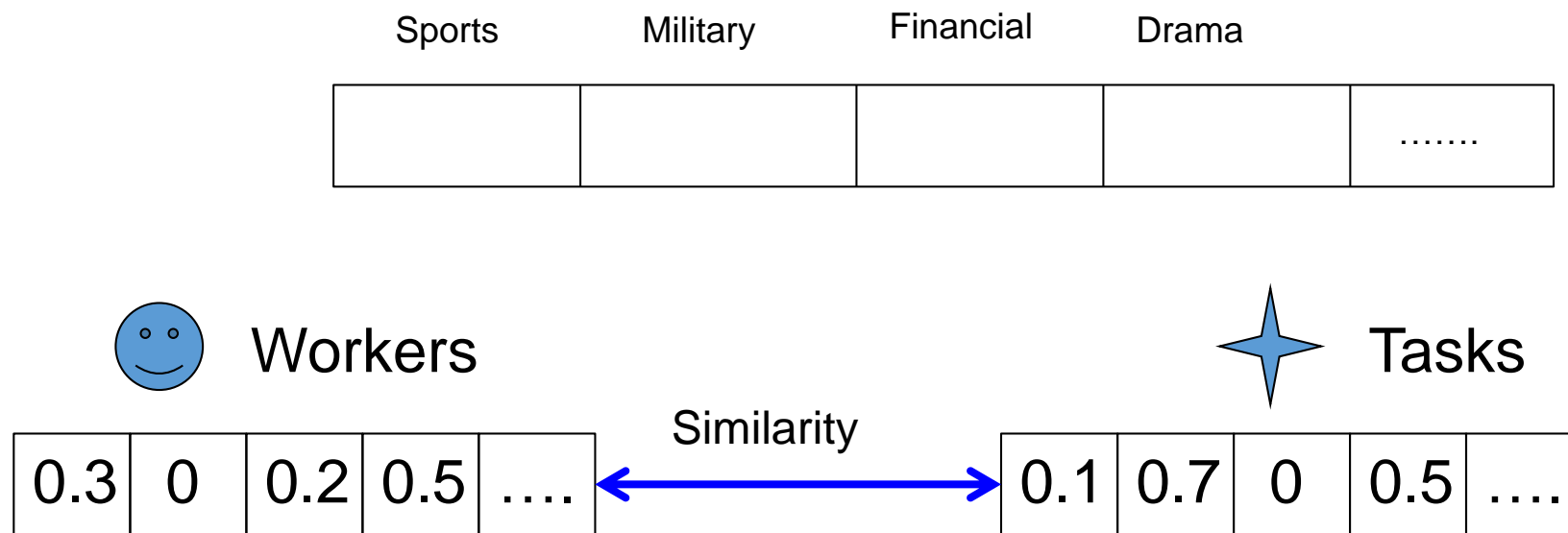
用户建模中的迁移学习[KDD13]



利用领域相似性和迁移学习理论，将用户的领域技能进行迁移推理

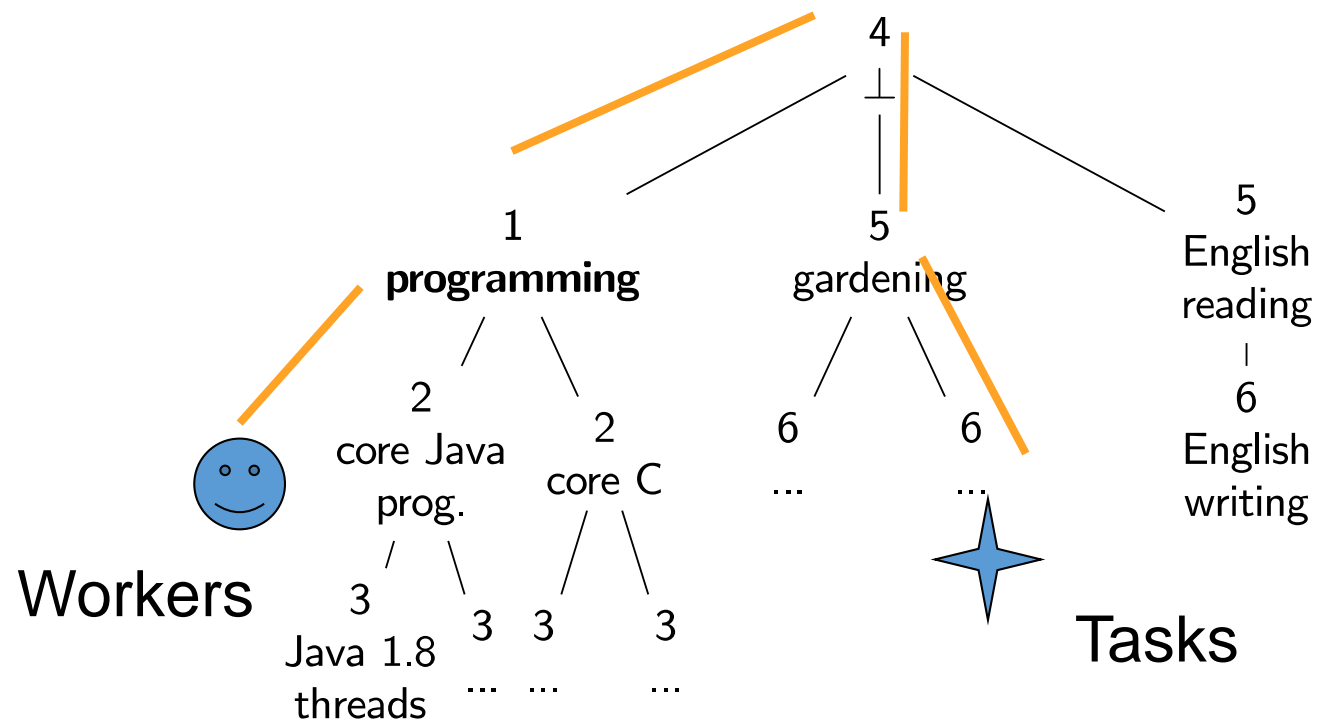


基于领域的匹配方式 [VLDB16]



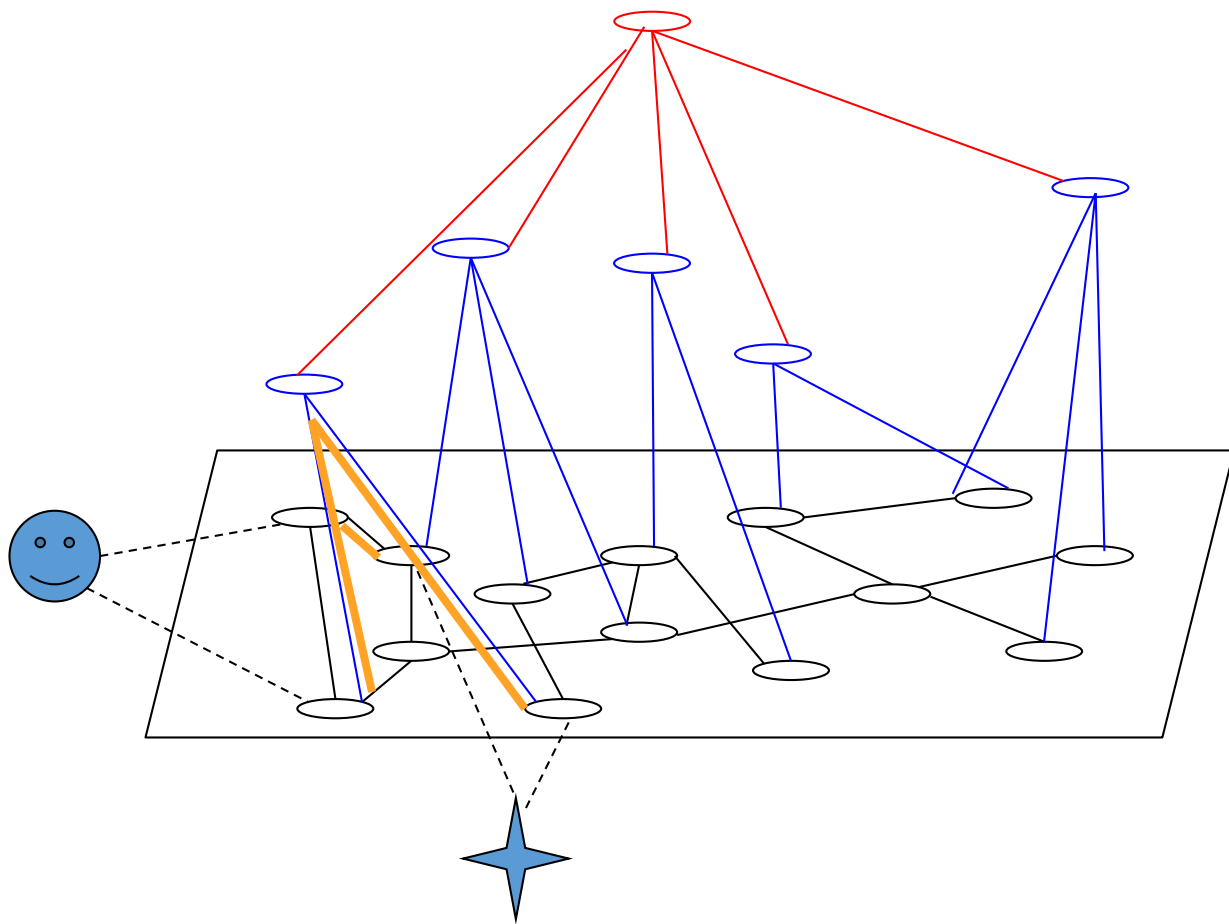
将所有任务分解成13个领域，计算工人与任务在每个领域的相关度

基于技能树的匹配方式 [WWW16]



利用技能树对用户和任务建模，
用树上的距离代表任务和用户
之间的相关度

树-图结合的方式



在某些任务中，如学术评审，领域交叉性强，光用技能树无法很好对任务建模。

知识型众包研究问题

- 将什么任务交给众包 (What)
- 如何筛选工人 (Who)
- 如何完成众包 (How)
 - 如何设计问题
 - 如何激励工人
 - 如何控制质量

众包问题设计的两种思路

- 显式众包
 - 工人知道自己正在做众包
 - 是众包的主流方式
- 隐式众包
 - 工人在不知不觉中完成众包
 - 利用第一任务吸引用户，在第二任务中完成众包
 - 价格低廉、效果更好

显式众包

- 传统原则
 - 小任务最受欢迎
 - 判断题 > 选择题 > 填空题
 - 越少交互越好
 - UI很重要
- 最新研究
 - 在花费和准确性之间做权衡
 - 多选与判断题的权衡
 - 众包 workflow 设计

隐式众包

- 游戏
 - 常识性知识获取
 - 地理位置信息获取
- 秘密获取
 - reCAPTCHAs
 - 自动图像焦点获取
 - 自动图像标注
- 利用心理特征
 - 好奇心
 - 注意力分散

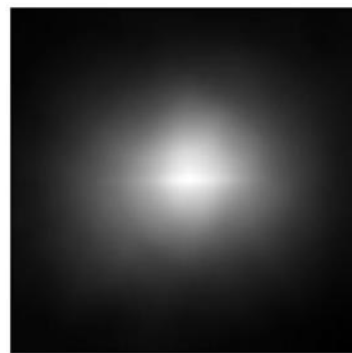
常识性知识获取[CHI06]



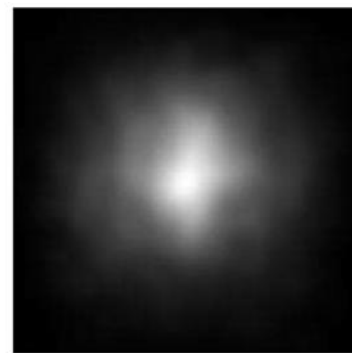
模板:

- ____ is a kind of ____.
- ____ is used for ____.
- ____ is typically near/in/on ____.
- ____ is the opposite of ____ / ____ is related to ____.

视觉焦点获取[TMM 14]



Touch



Visual



通过用户看到图片时点击屏幕的位置判断图片的焦点。

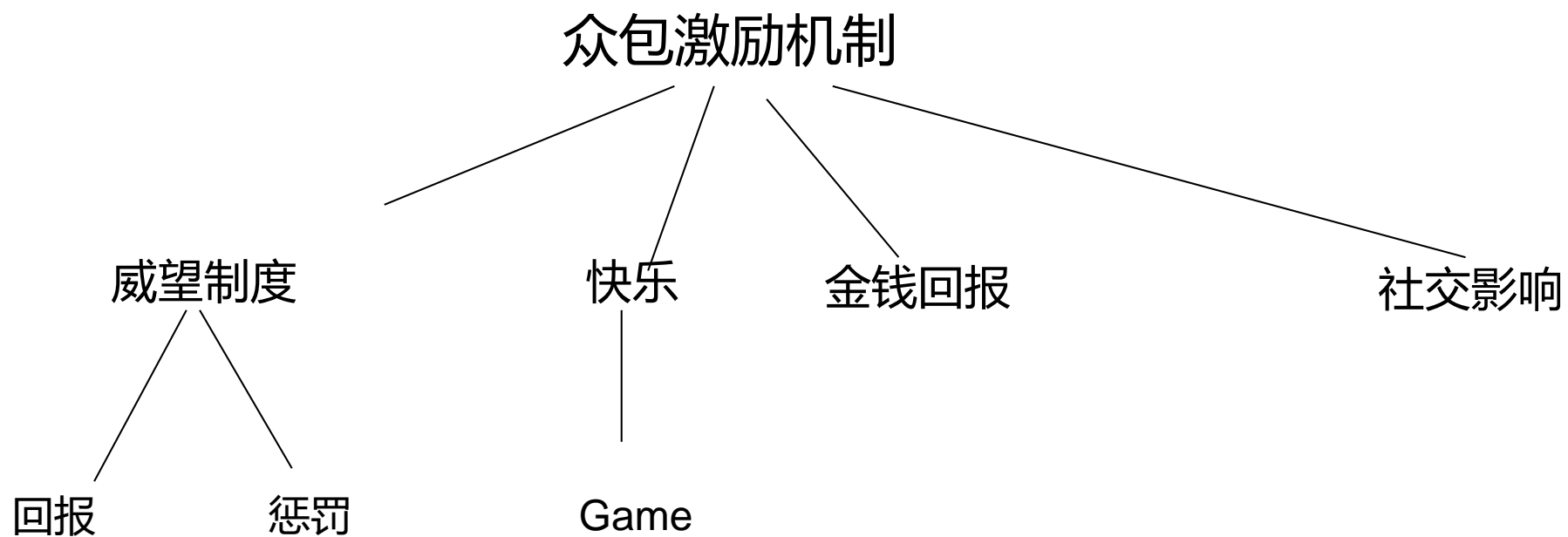
隐式众包原则

- 在无意识中提出任务
- 工人同时是用户
- 第一任务需要首先满足用户的需求，第二任务才是众包任务
- 第一任务的重要性要充分考虑
- 可以利用好奇心激励用户

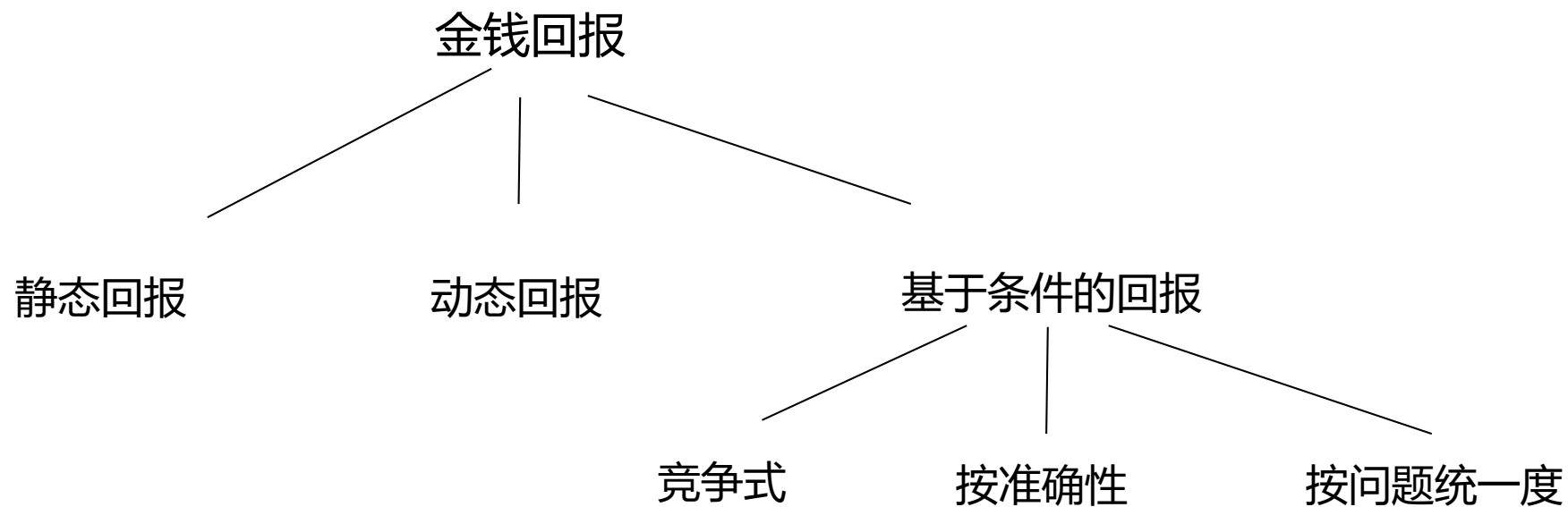
知识型众包研究问题

- 将什么任务交给众包 (What)
- 如何筛选工人 (Who)
- 如何完成众包 (How)
 - 如何设计问题
 - 如何激励工人
 - 如何控制质量

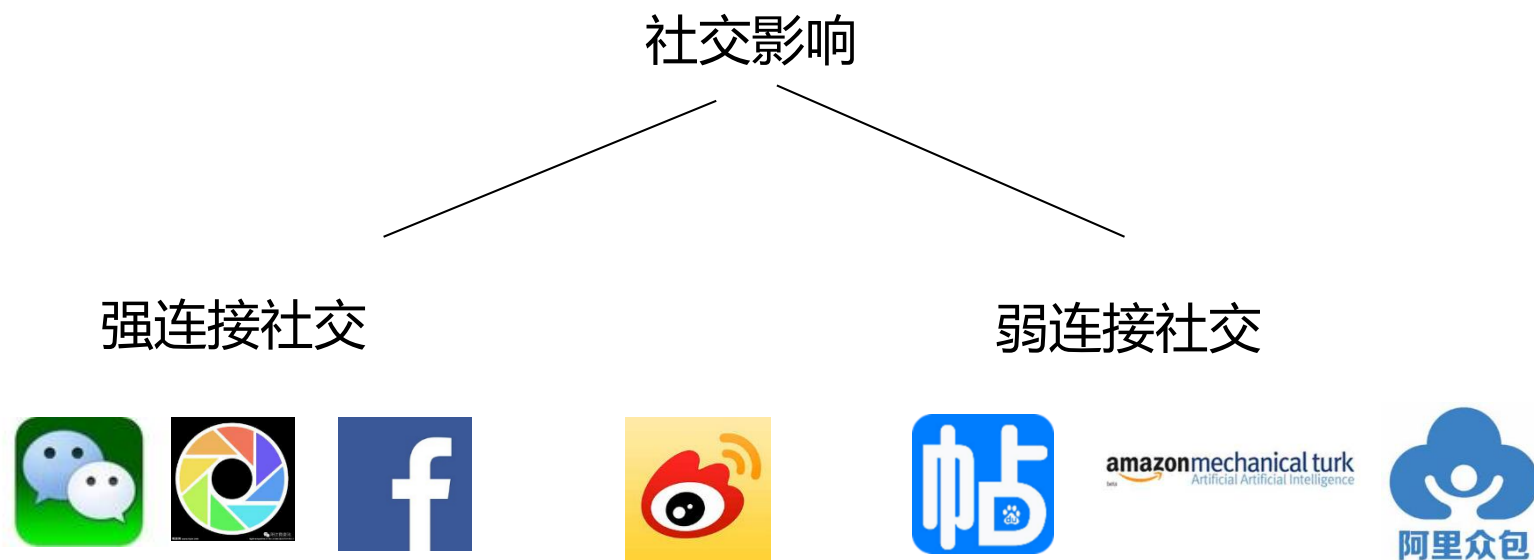
众包激励的分类学



众包激励的分类学



众包激励的分类学



最新研究

- 对于短期任务弱连接社交优于强连接社交
- 混合激励机制



在任务开始阶段利用强社交媒体做宣传，在聚集一定人气后利用弱社交媒体和金钱刺激，在尾段再次利用强社交媒体和金钱刺激手段吸引剩余工人

知识型众包研究问题

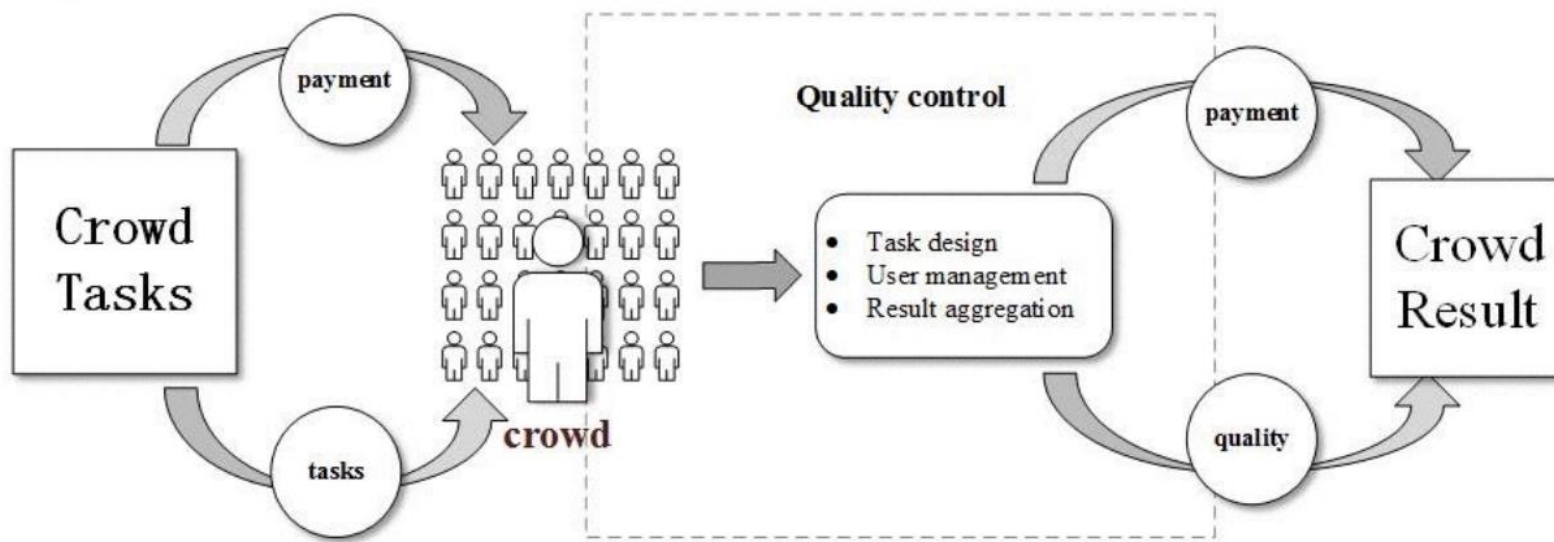
- 将什么任务交给众包 (What)
- 如何筛选工人 (Who)
- 如何完成众包 (How)
 - 如何设计问题
 - 如何激励工人
 - 如何控制质量

众包质量考虑的维度:

正确性
覆盖度
时效性
一致性

众包质量控制

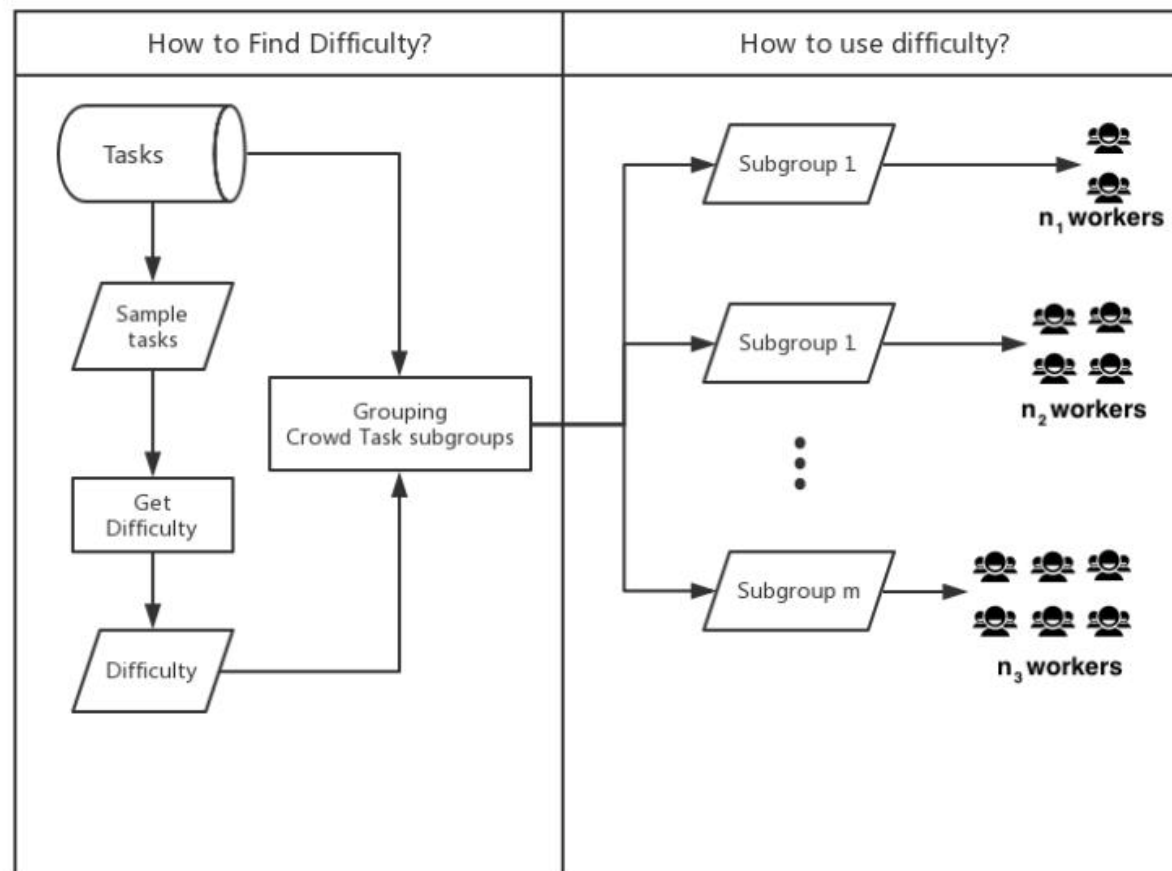
- 众包前质量控制
- 众包过程中质量控制
- 众包后质量控制



众包前质量控制

- 用户管理与分配
[GROUP18]

- 依据任务难度分配工人数目
- 先利用一个小型众包判断哪个特征决定众包的难度
- 根据难度分配众包人数



众包过程中质量控制

- 恶意用户分类

- 假冒资质工人 (Ineligible Worker)——如假冒学历
- 快速欺骗者 (Fast Deceivers) ——为了快速获得金钱回报而胡乱答题
- 规则破坏者 (Rule Breaker) —— 不按照任务设定完成任务
- 聪明的欺骗者 (Smart Deceivers) —— 胡乱答题的时候做了一些掩饰

- 常用方法

- 埋雷法——在题目中安插一些知道答案的任务检验工人质量
- 回溯问题——提问与上一题有关的问题来防止快速欺骗者

众包后质量控制

- 众包后质量主要通过度量答案的可信度来聚合收回的答案
- 个人评估
 - 评估单个工人的可信度
- 群组评估
 - 评估群组工人的可信度
- 基于大数据计算的方法
 - 根据历史统计、工人隐式反馈等方法判断工人可信度

个人评估

- 自评打分
- 交叉打分
- 能力测试
- 个性测试
- 引用标注
- 专家重审

群组评估

- 投票
 - 众数投票
 - 加权投票
- 群组一致性
- 结果聚合

基于大数据计算的方法

- 埋雷+计算
- 异常值检测
- 历史分析
- 隐式反馈

众包在知识图谱构建过程中的作用

知识图谱构建三个阶段

- 本体构建阶段
- 知识挖掘与填充阶段
- 知识图谱精化阶段

本体构建阶段

- 构建本体层次架构
- 构建语义词汇表
- 语义词汇表对齐
- 标注概念说明
- 标注与验证关系

具体应用

- OntoPronto



具体应用

- WikiData
 - Wikipedia的姊妹项目
- CrowdSPARQL
 - 当SparQL查询无法响应时，会重定向至Mturk平台获取知识
- InPhO
 - 利用众包完成概念体系构建

知识图谱构建三个阶段

- 本体构建阶段
- 知识挖掘与填充阶段
- 知识图谱精化阶段

知识挖掘与填充阶段 [KDD 18]

- 机遇与挑战

- 在知识获取和挖掘任务中，人天然比机器有优势
 - 能迅速准确地从自然语言中抽取三元组
 - 能准确对齐异构数据源中的实体
 - 能利用常识丰富知识库
- 然而，完全靠人工十分昂贵
- 人机结合是主要手段

知识挖掘案例

- 基于众包的知识获取
 - 从自然语言中抽取相关实体和三元组
 - 示范系统: HIGGINS
- 基于众包的实体对齐
 - 利用众包实现异构知识来源的实体对齐
 - 示范系统: HIKE
- 基于众包的实体收集
 - 利用众包收集一个开放的实体
 - 示范系统: CrowdEC

知识获取与三元组抽取

- 三元组抽取

- 《延禧攻略》是一部由东阳欢娱影视公司与2018年出品的古装宫廷剧。



东阳欢娱影视公司

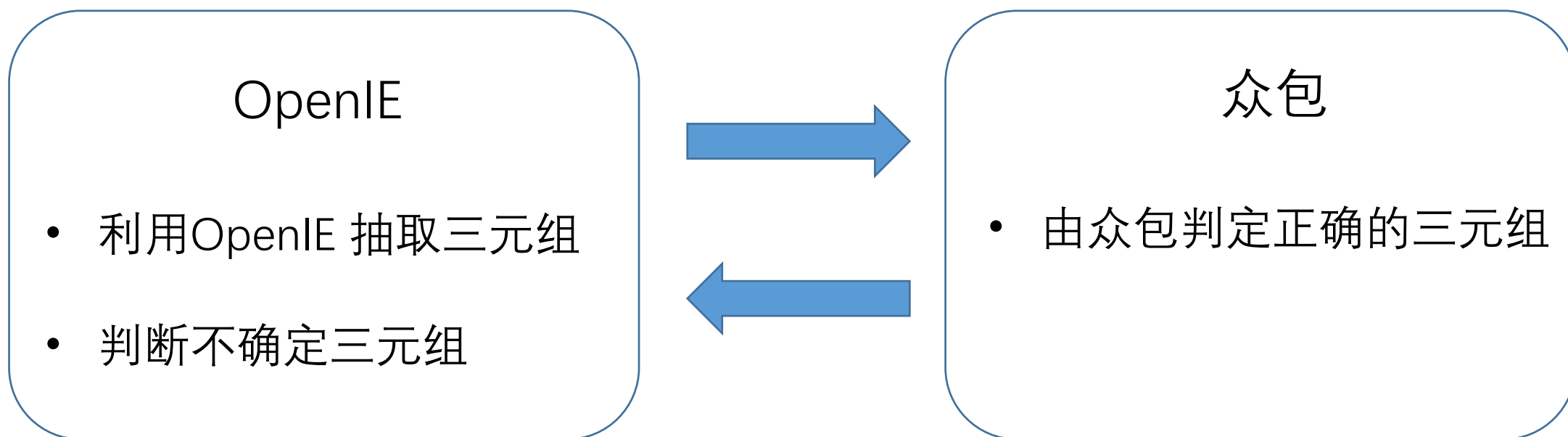
出品

《延禧攻略》

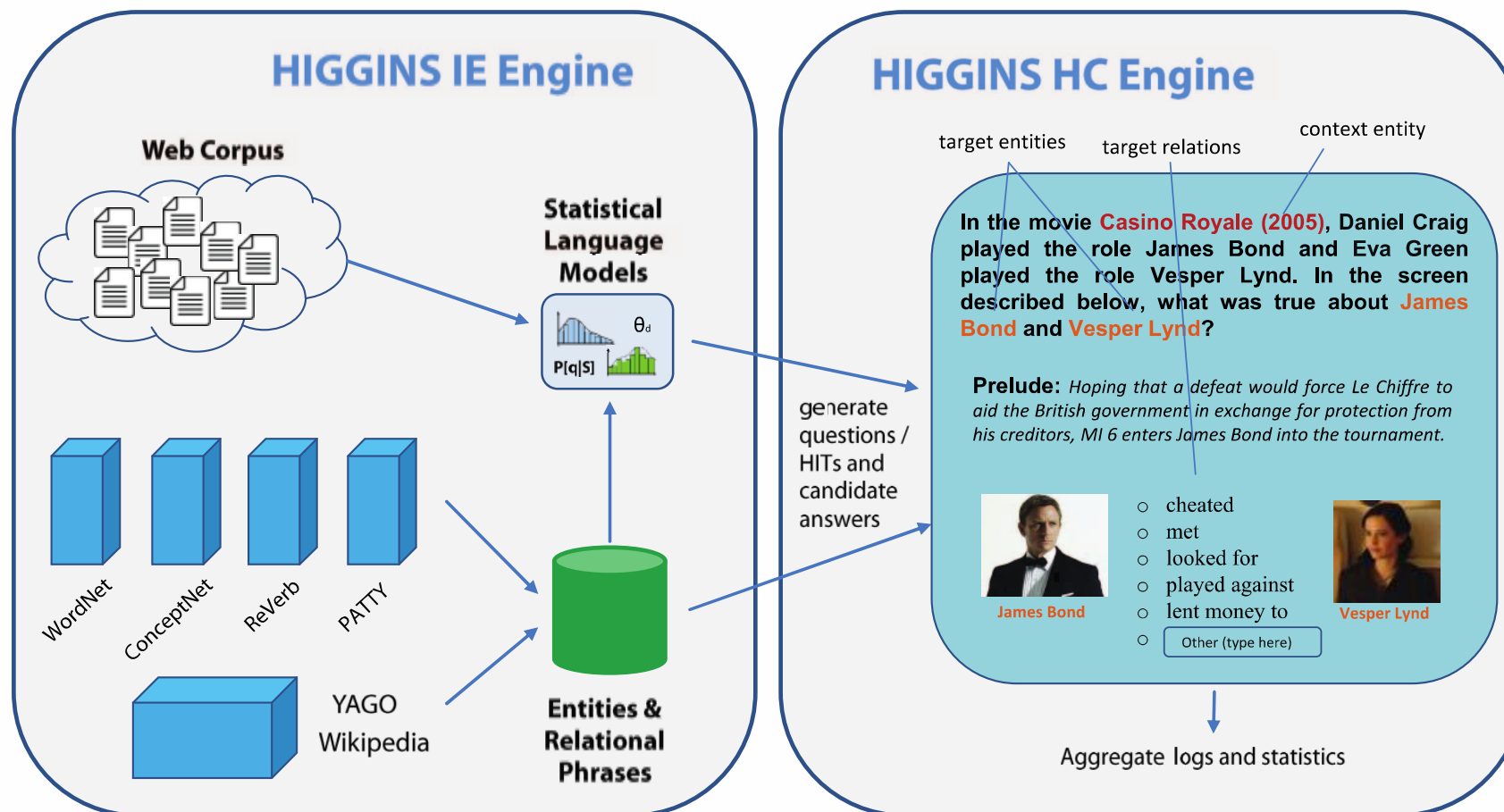
- 现有做法：Open IE + NLP
 - 缺点：存在较多噪声

知识获取与三元组抽取

- 利用众包可以大大降低复杂语句中出错的概率
- 但众包的开销太大



HIGGINS系统 [ICDE 14]



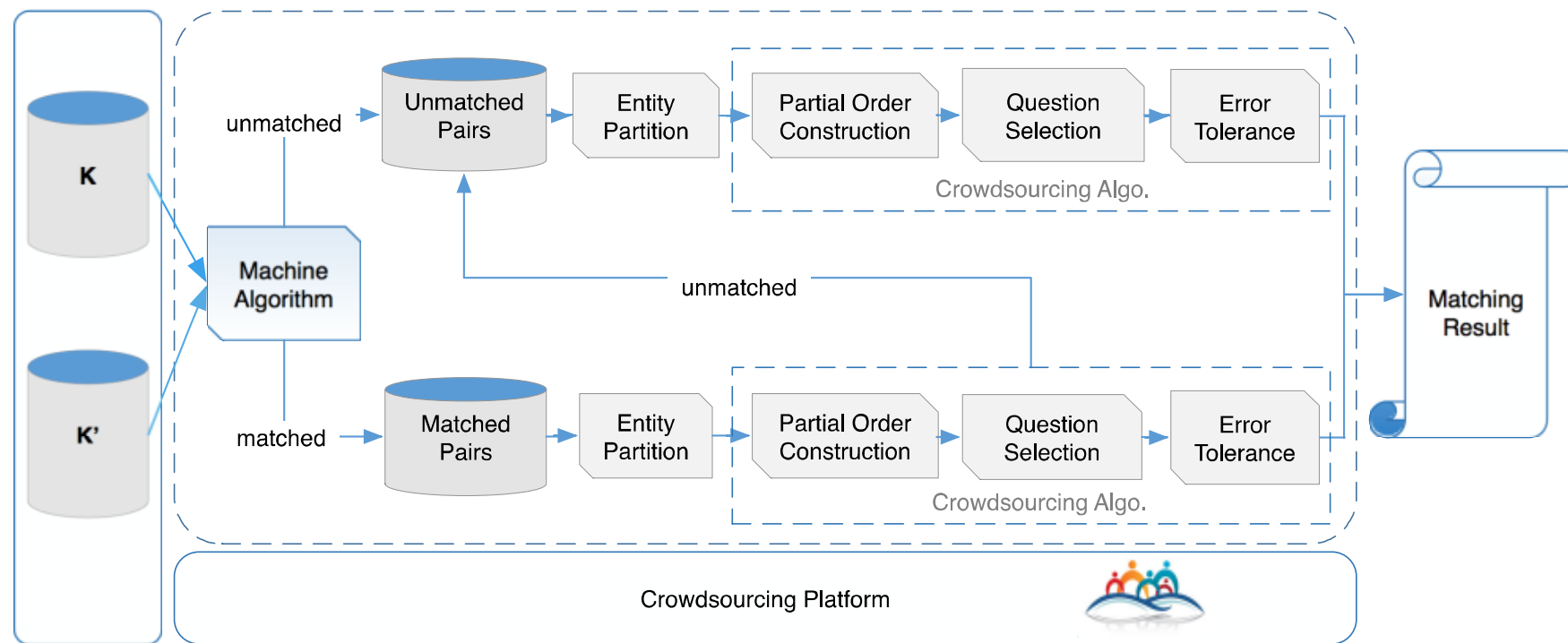
HIGGINS系统

- HIGGINS 信息抽取引擎
 - 识别实体
 - 利用语法规则识别关系词
 - 过滤可能性低的三元组
- HIGGINS众包引擎
 - 生成众包问题
 - 生成候选答案

知识挖掘案例

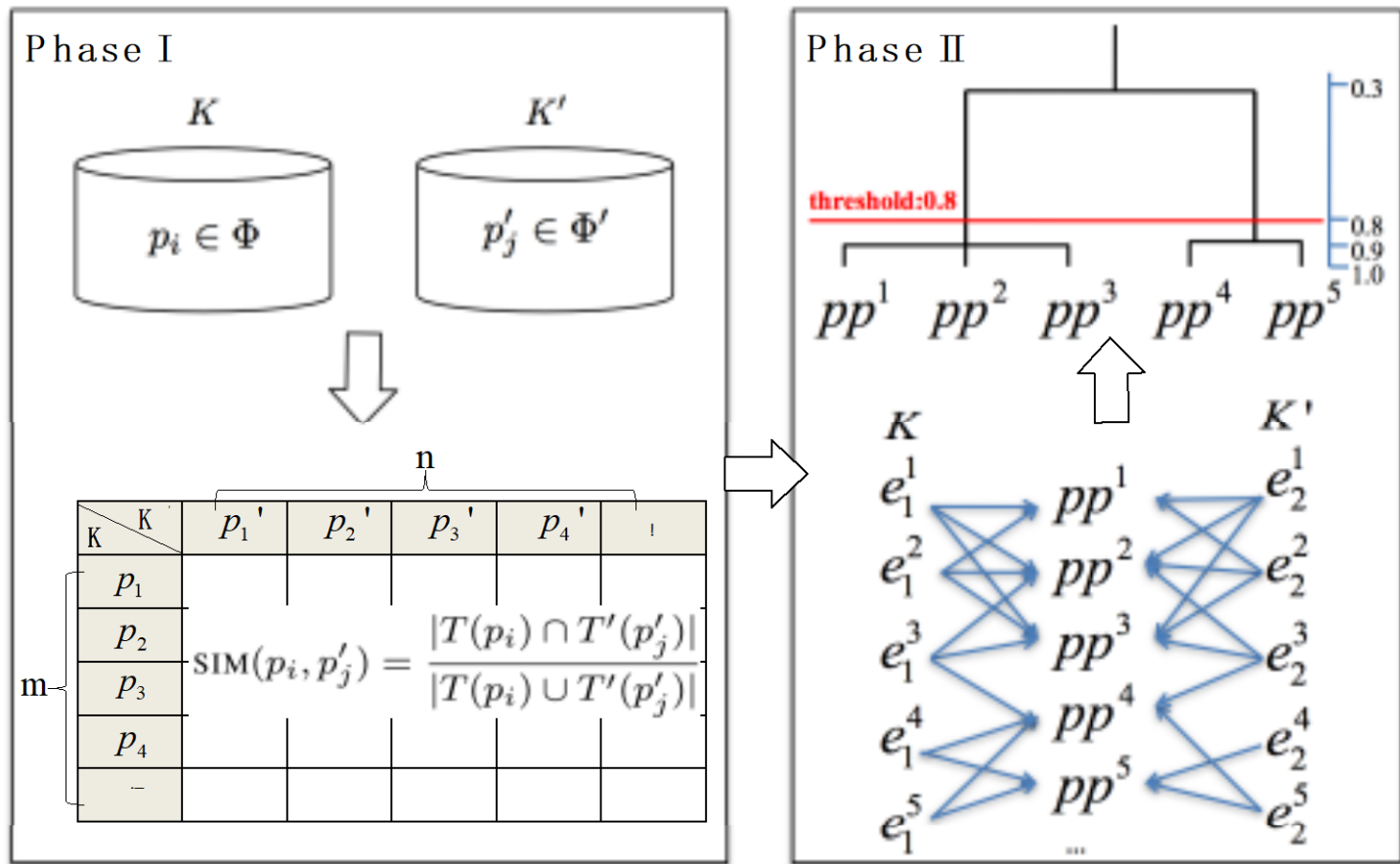
- 基于众包的知识获取
 - 从自然语言中抽取相关实体和三元组
 - 示范系统：HIGGINS
- 基于众包的实体对齐
 - 利用众包实现异构知识来源的实体对齐
 - 示范系统：HIKE
- 基于众包的实体收集
 - 利用众包收集一个开放的实体
 - 示范系统：CrowdEC

基于众包实体对齐系统：HIKE [CIKM17]



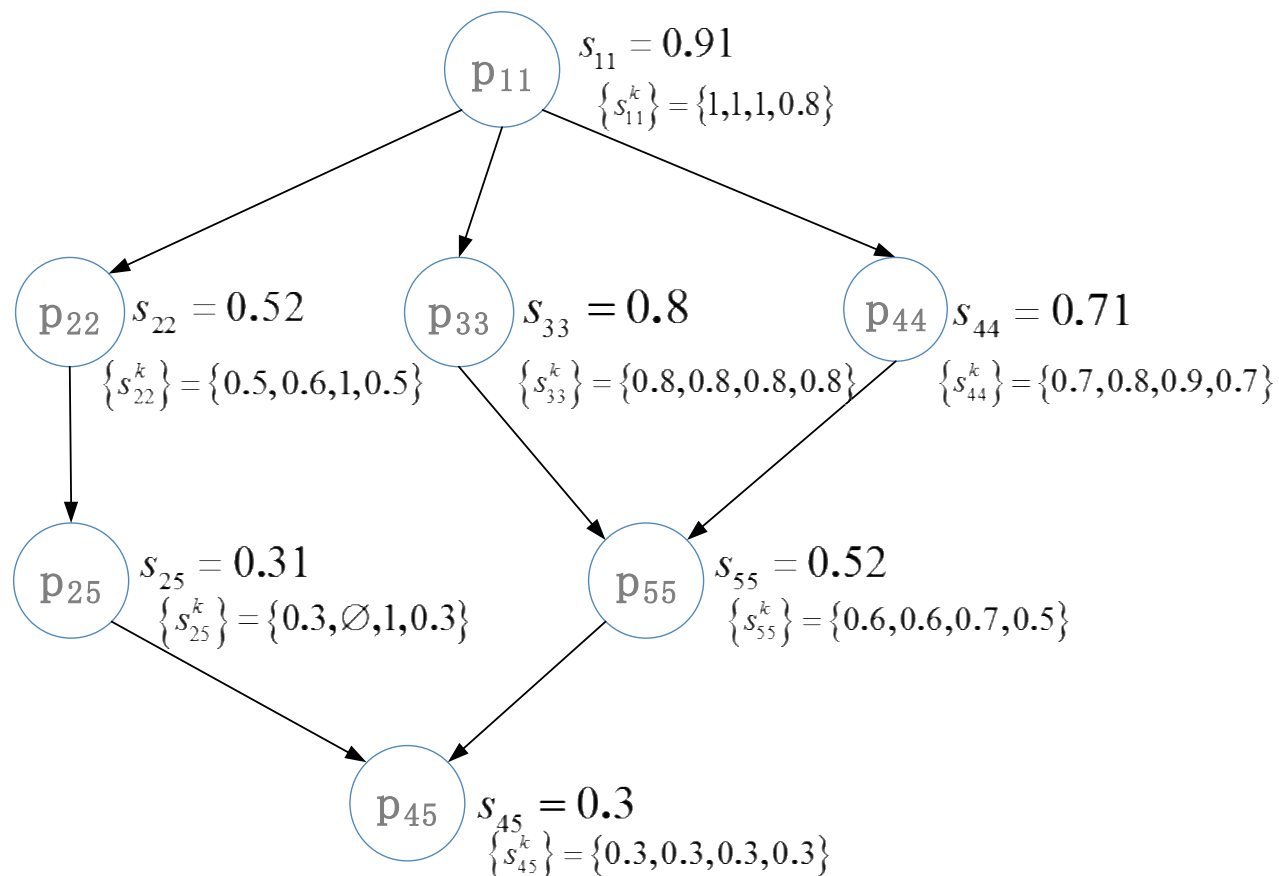
- 四个步骤：
1. 实体分块
 2. 偏序对构建
 3. 问题生成
 4. 错误处理

HIKE: 实体分块



- 实体分块的作用是减少可能的实体对
- 利用与相似谓词关联的紧密程度进行分块

HIKE: 偏序图



偏序关系的利用:
假设 实体对 $p_1 > p_2$,
如果 p_2 可以匹配, 则
 p_1 也一定匹配; 如果
 p_1 不能匹配, p_2 也不
能匹配

知识挖掘案例

- 基于众包的知识获取
 - 从自然语言中抽取相关实体和三元组
 - 示范系统：HIGGINS
- 基于众包的实体对齐
 - 利用众包实现异构知识来源的实体对齐
 - 示范系统：HIKE
- 基于众包的实体收集
 - 利用众包收集一个开放的实体
 - 示范系统：CrowdEC

开放类实体收集

请列举说有在中国打过球的前NBA球员

1. 麦迪
2. JR 史密斯
3. 马布里

1. 麦迪
2. 马布里
3. 斯科拉

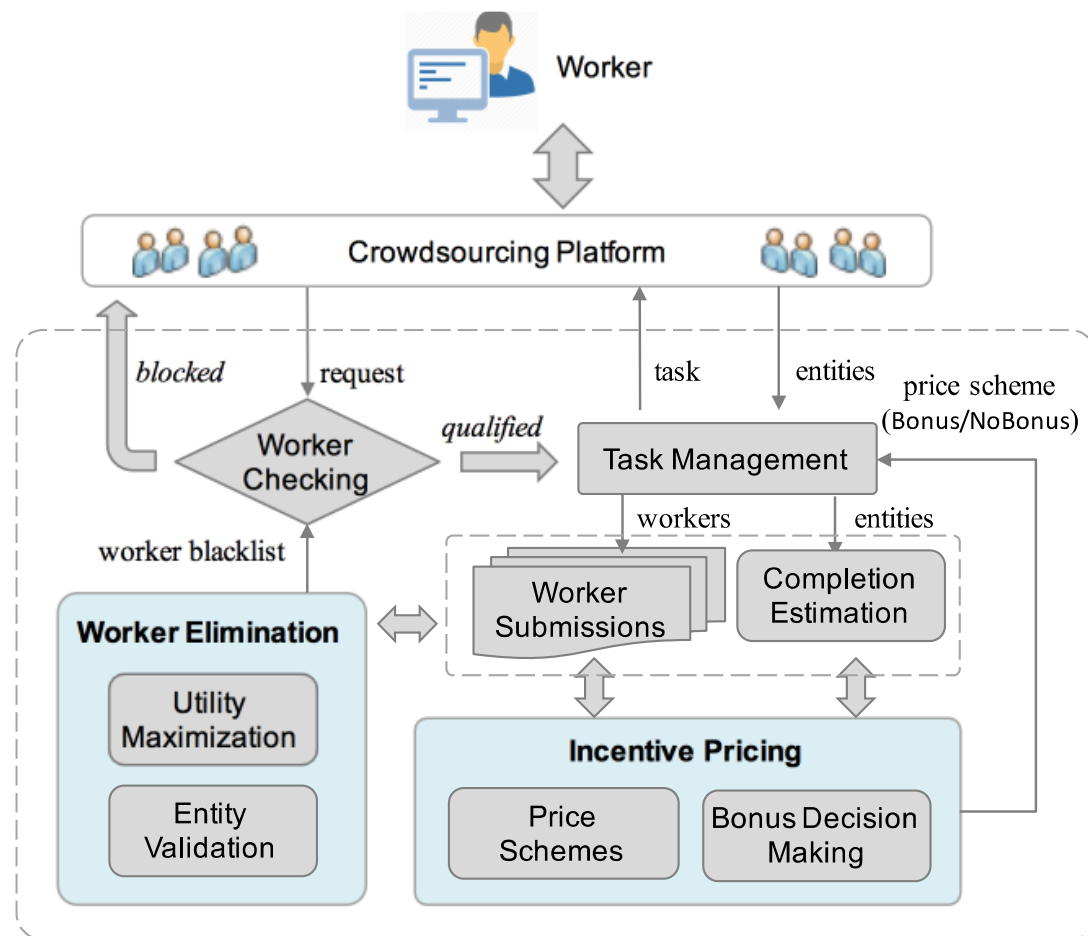
1. 弗兰西斯
2. 科比

面临的问题:

1. 重复
2. 遗漏
3. 错误

易建联? ? ?

CrowdEC系统 [ICDE 18]



特点：利用定价原则引导工人提供不重复的答案

知识图谱构建三个阶段

- 本体构建阶段
- 知识挖掘与填充阶段
- 知识图谱精化阶段

知识精化阶段 [SWJ 2016]

- 为什么需要引入众包帮助知识图谱精化
 - 自动化手段很难实现准确度和覆盖度的双高
 - 网络中的文档存在长尾效应
 - 自动化处理技术存在缺陷
- 众包精化的手段
 - 补缺
 - 纠错

众包补缺

- 所有百科网站
- 常识性知识输入
 - Cyc
 - OpenMind
- 基于链路预测技术的众包验证
- 领域知识图谱的补缺

众包纠错

- 两种常用方法
 - 公开所有数据，由众包自由挖掘错误，如谷歌。但以上方法需要网站拥有超多人流量。
 - 先由机器定位疑似错误，再交予众包确认。适合流量不大的网站和机构。
- 公开数据法
 - 多级审核
- 人机结合
 - 多知识库冲突检测
 - 对偶判断
 - isA闭环检测

知识图谱构建中众包利用原则

- 知识图谱的基本架构不应该交予众包
- 众包更擅长做知识图谱的评价
- 众包培训对知识图谱项目十分重要
- 人机结合精化是大多数知识图谱项目必须考虑的问题
- 质量控制尤其关键
- 众包的开销控制是所有研究的重点
- 即使众包也存在长尾效应

Reference

- [CHI06] L. Ahn, et.al. Verbosity: A Game for Collecting Common-Sense Facts. CHI, 2006
- [ICDE18] Chengliang Chai, Ju Fan, Guoliang Li: Incentive-based Entity Collection using Crowdsourcing. ICDE 2018
- [CIKM17] Y. Zhuang, G. Li, Z. Zhong, J. Feng: Hike: A Hybrid Human-Machine Method for Entity Alignment in Large-Scale Knowledge Bases. CIKM 2017.
- [ICDE 14] S. K. Kondreddi, P. Triantafillou, G. Weikum: Combining information extraction and human computing for crowdsourced knowledge acquisition. ICDE 2014

Reference

- [KDD 18] C. Chai, J. Fan, G. Li et.al: Crowd-Powered Data Mining. KDD tutorial, 2018.
- [SWJ 16] H. Paulheim: Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. Semantic Web Journal, 2016.
- [GROUP 18] Y. Jiang, Y. Sun, J. Yang, X. Lin, L. He: Enabling Uneven Task Difficulty in Micro-Task Crowdsourcing. GROUP, 2018.
- [KDD 13] K.Mo. Cross-task Crowdsourcing. KDD, 2013.
- [TMM14] B. Ni,et al. Touch Saliency: Characteristics and Prediction[J]. IEEE Transactions on Multimedia, 2014, 16(6):1779-1791.

Reference

- [VLDB15] C. Zhang, et.al. Reducing uncertainty of schema matching via crowdsourcing. VLDB, 2015
- [SIGMOD13] J. Wang, et.al. Leveraging Transitive Relations for Crowdsourced Joins. SIGMOD, 2013.
- [WWW16] P. Mavridis, et.al. Using Hierarchical Skills for Optimized Task Assignment in Knowledge-Intensive Crowdsourcing. WWW, 2016
- [VLDB16] Y. Zheng, et.al. DOCS: Domain-Aware Crowdsourcing System. VLDB, 2016
- [TKDE 17] X. Lin, Y. Peng, et.al, Human-Powered Data Cleaning for Probabilistic Reachability Queries on Uncertain Graphs. TKDE, 2017.