



哈尔滨工业大学

海量数据计算研究中心

Massive Data Computing Lab @ HIT

大数据计算基础

第二章 大数据算法

哈尔滨工业大学

刘显敏

liuxianmin@hit.edu.cn





哈尔滨工业大学

海量数据计算研究中心

Massive Data Computing Lab @ HIT

大数据计算基础

第二章 大数据算法——众包算法

哈尔滨工业大学

刘显敏

liuxianmin@hit.edu.cn



本讲内容

众包的定义

众包的实例

众包的要素

众包算法例析



何为众包？

- Outsourcing – 外包
 - 已知的雇员
- Crowdsourcing – 众包
 - 一群不固定，通常数量很大的参与者
 - 将“开源”的思想应用于软件之外

最成功的应用：Wikipedia



WIKIPEDIA
The Free Encyclopedia

众包的定义

- 协调**一个群体**(互联网上的一大群人)做**“微工作”**(每人做一点贡献)来**解决软件或者单个人难以解决的问题**
- 通过一系列的机制和方法来**指导和协调群体的行为**,从而达到目的



宽泛的定义

social



volunteer

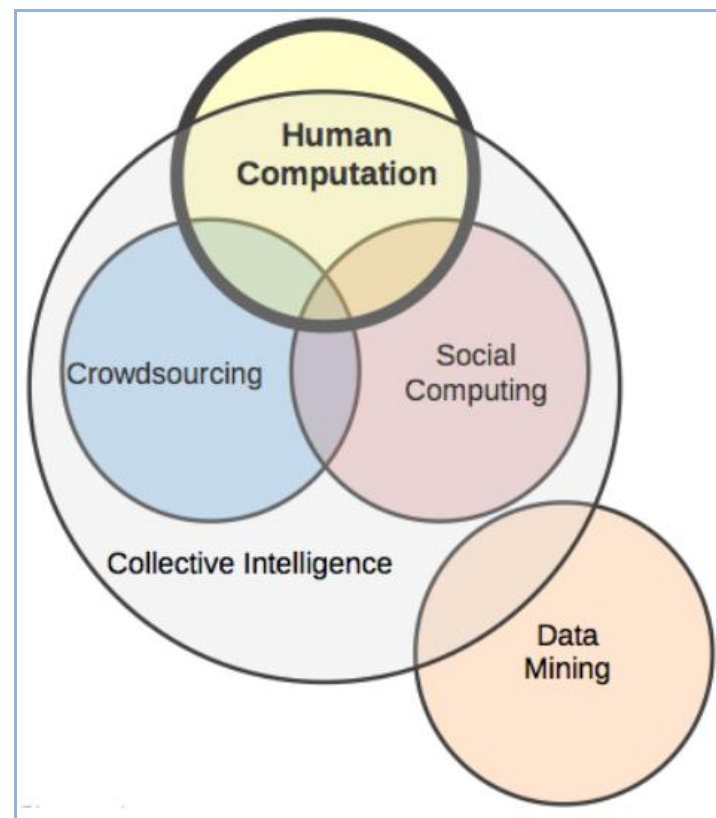


fun



众包 vs 人本计算

- 众包
 - 大任务到微任务
 - 众包极大程度使用了人本计算
- 它们不等价



本讲内容

众包的定义

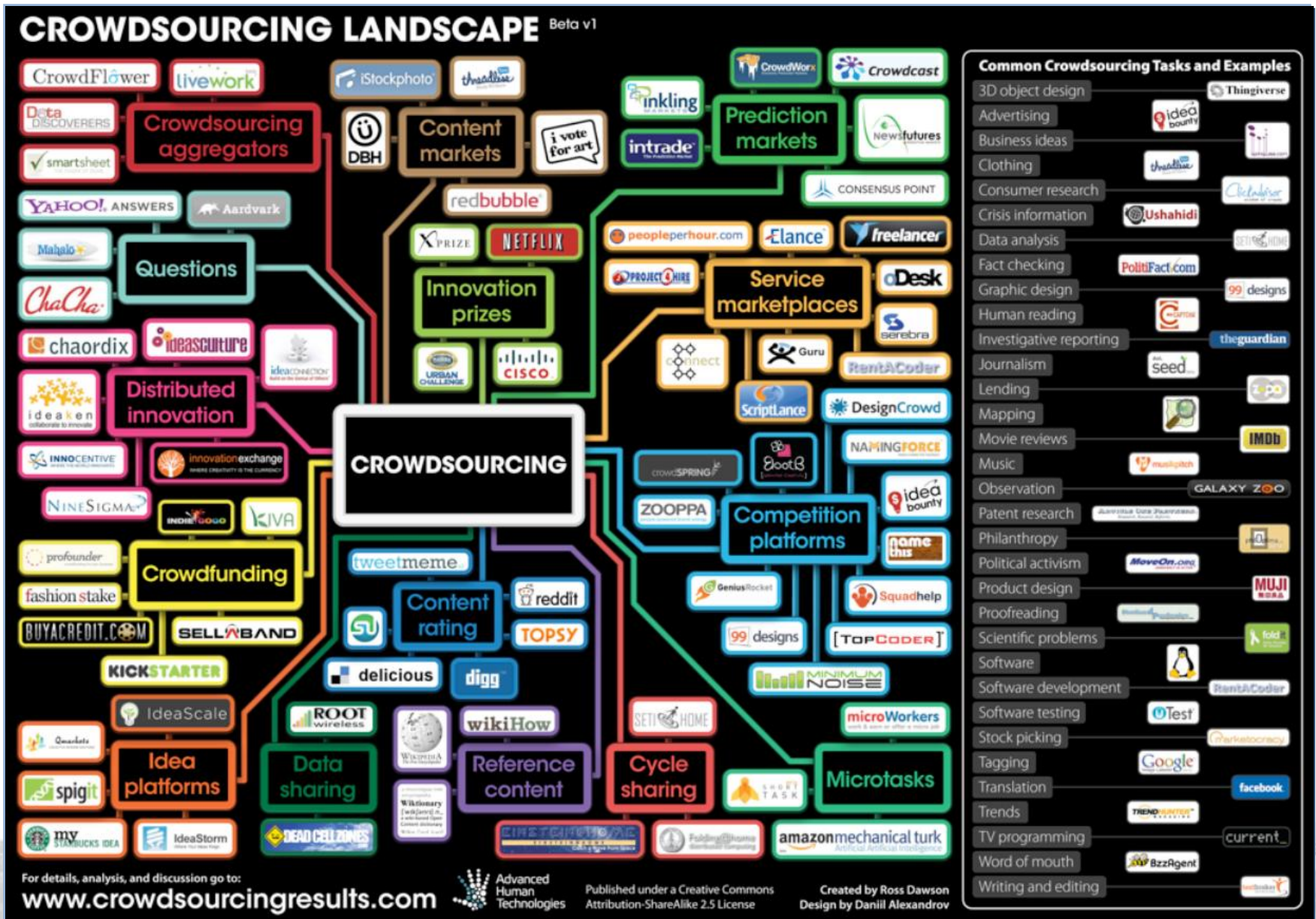
众包的实例

众包的要素

众包算法例析



工業界



例子 – 验证码

► 验证码: 每天200M



NLP例子- 机器翻译

- 机器翻译
 - 人工评估翻译质量慢而且成本高
 - 非专家和专家认同度高
 - 翻译一个句子\$0.10

C. Callison-Burch. “Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon’s Mechanical Turk”, EMNLP 2009.

B. Bederson et al. Translation by Iterative Collaboration between Monolingual Users, *GI* 2010

IR的例子- 图像搜索

Instructions

You are shown two images. You must select the image that is more indicative of suspicious activities.

Task

Imagine that you are a security guard and you are monitoring two places. Someone informed you that there are suspicious activities in one of the places, but you were not told which one. Which place will you attend to?



Submit

score
16

Matchin
A question of taste.

time
1:32

Which image do you prefer?



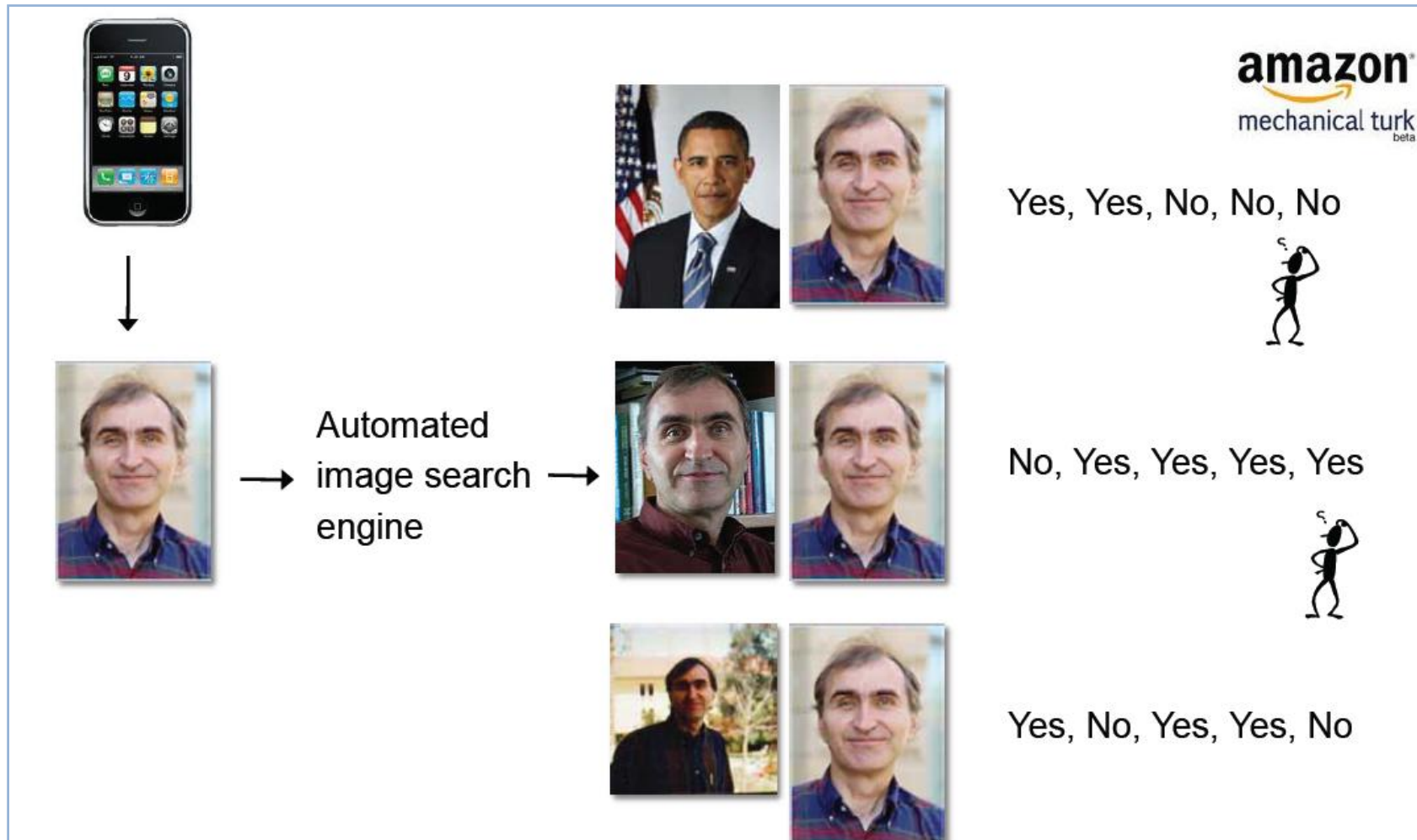
This one!



That one!

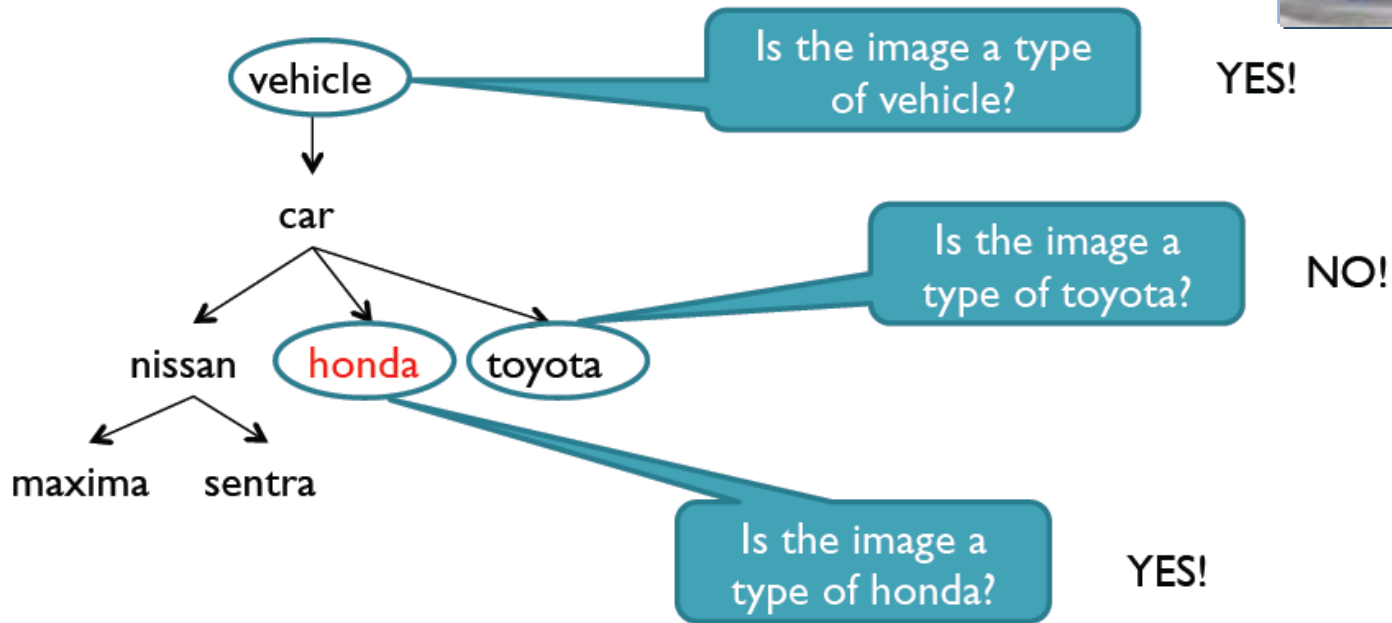
Tingxin Yan, Vikas Kumar, Deepak Ganesan: CrowdSearch: exploiting crowds for accurate real-time image search on mobile phones. MobiSys 2010:77-90

众包搜索：利用众包改进图像搜索



图片分类

- 分类一副图片



target node = intended category
Is the image a type of X? = Is the target node reachable from X?

IR的例子 - 广告

Assignments Completed **0** Accuracy **?** [Send Feedback](#)

You are in preview mode. Remember to accept the HIT before working on it!

How relevant are these 25 advertisements to a search term?

Instructions [Hide](#)


In this task, you will be given a search term and a small advertisement. Please rate how relevant the advertisement is to the search terms. The scale is from 1 to 4, where 1 is not relevant at all and 4 is completely relevant. Below is a description of each rating.

- 4 - Completely Relevant Ads**
These are often the exact item
- 3 - Closely Related Ads**
An ad for iPod cases would be
- 2 - Somewhat Related Ads**
For instance, an ad for speakers
- 1 - Irrelevant Ads**
Ads that have **nothing** to do with

Tips

A search query of "sunglasses"

Search Terms: coat size 12



Fashionable Clothing 8-36
Plus Size Gothic Burlesque Fashion
Satin 80s Fancy Dress Party Sale


How relevant is this ad to the search terms? (required)

Not Relevant At All

1 2 3 4

Very Relevant

Search Terms: coat size 12



Juicy Tubes Gift Sets
Huge selection of Juicy Tubes Sets
Full size + minis available only @

How relevant is this ad to the search terms? (required)

1 2 3 4

Omar Alonso, Daniel E. Rose, Benjamin Stewart: Crowdsourcing for relevance evaluation. SIGIR Forum (SIGIR) 42(2):9-15 (2008)

CV的例子 - 绘画相似性

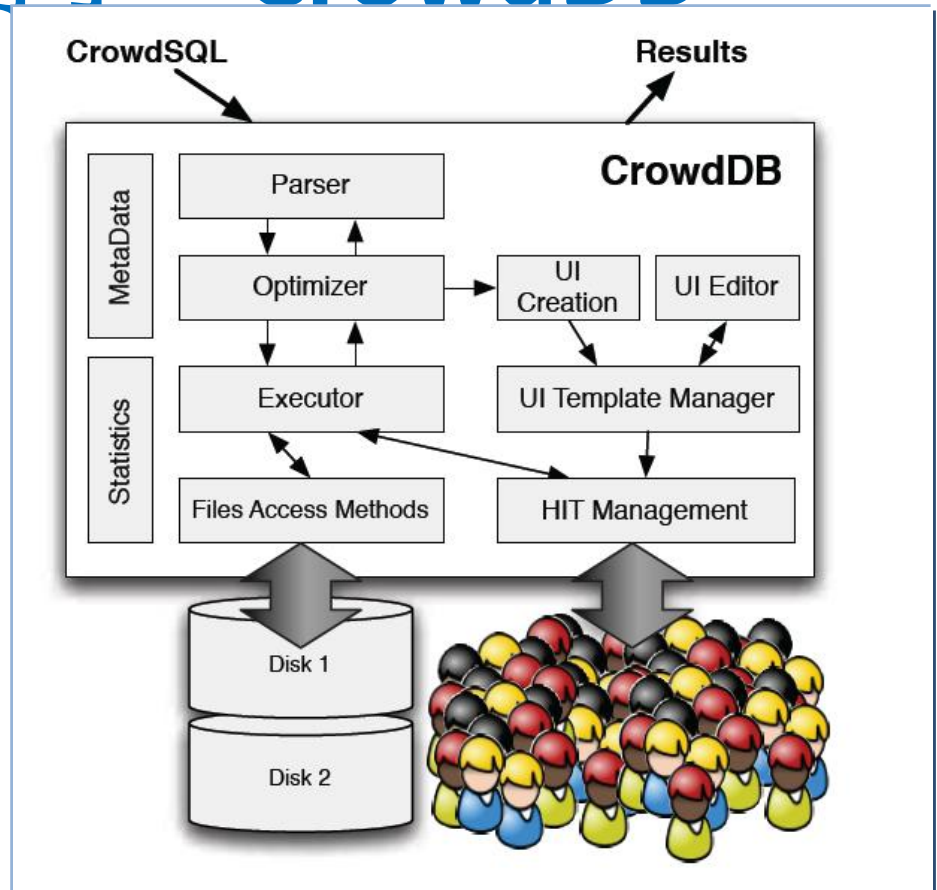


▶ 上边的画在艺术风格上有多大程度相似？

▶ 非常相似、相似、有些不同、非常不同

数据库的例子 - CrowdDB

- 使用众包来回答数据库查询
 - 在哪里使用众包?
 - 如何使用众包?



Michael J. Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, Reynold Xin: CrowdDB: answering queries with crowdsourcing. SIGMOD 2011:61-72

本讲内容

众包的定义

众包的实例

众包的要素

众包算法例析



众包

▶ 请求者

▶ 提交任务



提交任务

收集答案

▶ 平台

▶ 任务管理

发布任务

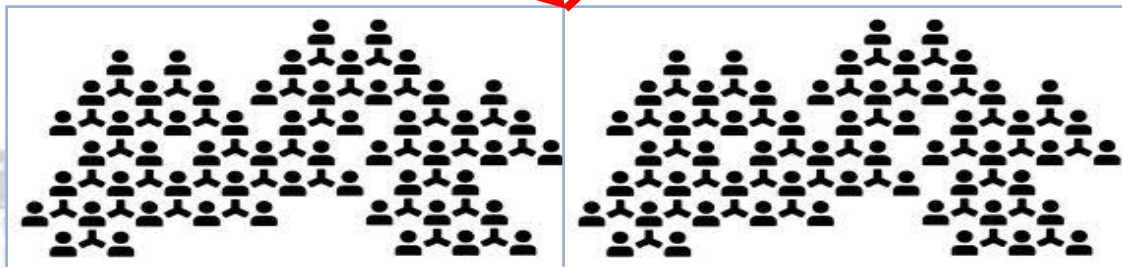


▶ 工人

▶ 在任务上工作的人

查找感兴趣的任務

返回答案



智能任务(HITs)

- 请求者通过Web服务API创建“智能任务”(HITs)。
- 工人
 - 登陆
 - 选择 HITs
 - 执行之。
- 请求者评估结果，给完成的HIT打分(满意度)。

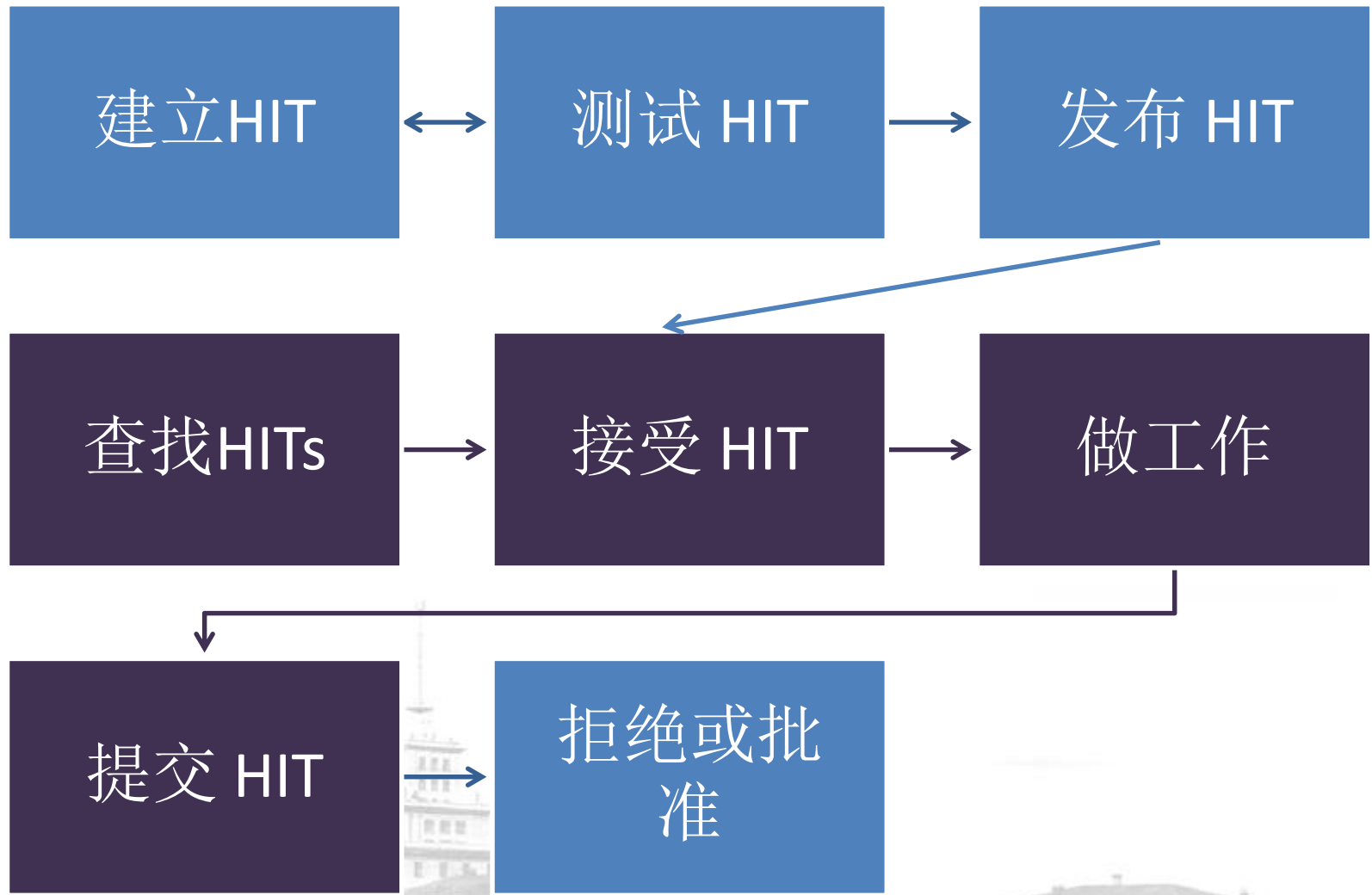
目前 >1,000,000 工人

来自 100 国家

完成数以百万HITs

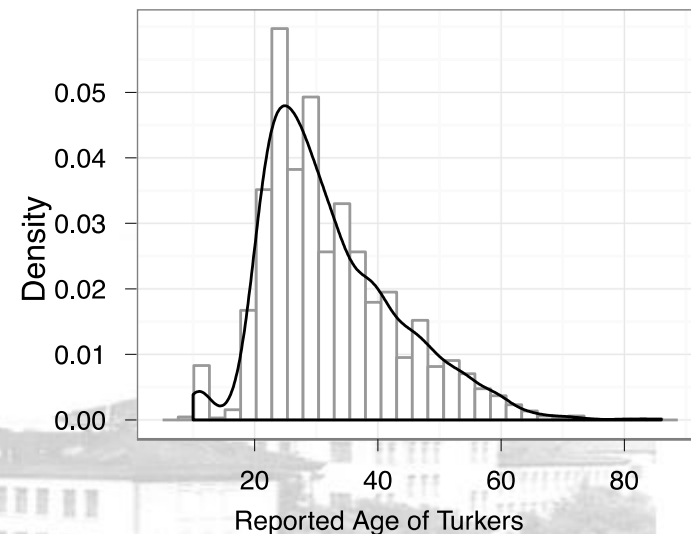


请求者 & 工人



工人的回报

- 报酬 (\$\$\$)
- 有趣（避免无聊）
- 社交
- 赚好评/声望
- 学习新的东西（例如英语）
- 副产品（例如重验证码）
- 创建自服务资源（如维基百科）
- 多重诱因通常是在并行工作



任务的报酬

- 一个HIT多少钱?
- 微妙的平衡
 - 太少了，没兴趣
 - 太多了，吸引垃圾发送者
 - 付出过多构成反激励
 - 钱并不能提高质量，但是（通常）增加参与性

Winter A. Mason, Duncan J. Watts: Financial incentives and the "performance of crowds". SIGKDD Explorations (SIGKDD) 11(2):100-108 (2009)



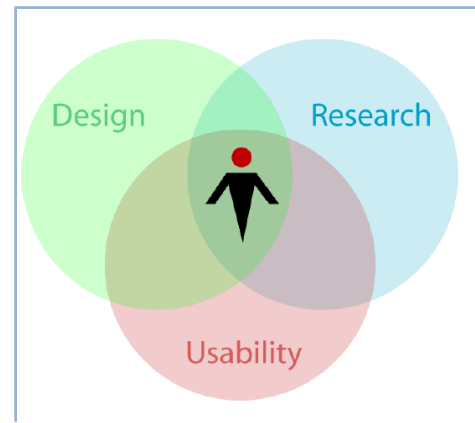
众包中的问题

- 选择（或自主开发）众包平台
- 人机交互
 - 付款/激励，界面和交互设计，通信，名誉，招聘，挽留
- 质量控制/数据质量
 - 信任，可靠性，垃圾邮件检测，标签共识
- 任务管理
- 人群管理
 - 结合人类处理单元（HPU）和CPU



人机交互

- 在人机交互中从用户获取输入是很重要的
 - 调查
 - 快速原型
 - 可用性测试
 - 认知走查(cognitive walkthrough)
- 在第3页及以后的HITs，不会有工人选择。
- 许多这样的HITs放在那里超过一个月，始终没有完成。
- 设计不良的任务发现界面会伤害市场中的每一位参与者！



质量控制

- 工人回答的质量是极为重要的组成部分
- 将它作为“整体”质量 - 不仅仅是工人
- 双向评价
 - 你可能会认为工人做得不好。
 - 同样的工人可能会认为你是一个糟糕的请求者。



质量控制

- 支持率
 - 多数表决
 - 确定始终不同于多数的工人
- 资格考试
 - 问题：减缓处理过程，很难测试相关性
 - 解决方案：创建主题相关的问题，以便用户在开始评估前熟悉此过程
- 缺少保证- 仍然不足以保证获得好的结果



资格测试：优点和缺点

- 优点
 - 利用好的的工具来控制质量
 - 调整及格标准
- 缺点
 - 设计和实施测试需要额外的成本
 - 可能会解雇工人，耽误完成时间
 - 难以核实主观任务
- 尝试创建和任务相关的问题，让工人在开始任务之前熟悉任务



处理坏的工人

- 支付“坏”的工作而不是拒绝它？
 - 赞成：维护声誉，承认设计不良的故障
 - 反对：鼓励欺诈，损害评级系统
- 用奖金作为奖励
 - 最低支付\$0.01且以\$0.01作为奖金
 - 比拒绝0.02美元的任务更好
- 如果垃圾发送者被抓到，则阻塞其未来的工作



任务分配

- 推方法：系统 \Rightarrow 工人
 - 系统采取完全的控制将指定的任务分配给谁。
- 拉方法：工人 \Rightarrow 系统
 - 该系统只设置环境，使工人自己给自己（或彼此）的分配分配。



任务推荐

- 迭代推荐
- 捕捉用户兴趣
- 估计用户质量
- 成本 - 质量权衡



本讲内容

众包的定义

众包的实例

众包的要素

众包算法例析



实体识别

ID	Product Name	Price
r_1	iPad Two 16GB WiFi White	\$490
r_2	iPad 2nd generation 16GB WiFi White	\$469
r_3	iPhone 4th generation White 16GB	\$545
r_4	Apple iPhone 4 16GB White	\$520
r_5	Apple iPhone 3rd generation Black 16GB	\$375
r_6	iPhone 4 32GB White	\$599
r_7	Apple iPad2 16GB WiFi White	\$499
r_8	Apple iPod shuffle 2GB Blue	\$49
r_9	Apple iPod shuffle USB Cable	\$19

现有的技术



机器

人

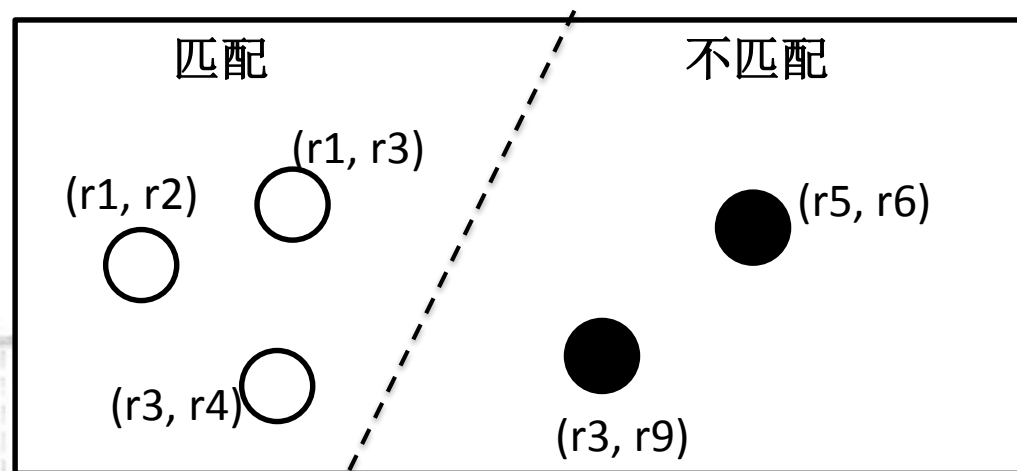


机器

- 基于相似性
 - 相似性函数（例如.Jaccard）
 - 阈值(例如.0.8)

$\text{Jaccard}(r1, r2) = 0.9 \geq 0.8$	✓
$\text{Jaccard}(r4, r8) = 0.1 < 0.8$	✗

- 基于学习





- CrowdDB [Franklin et al. SIGMOD'11]

SELECT p.id,
 q.id
FROM product p,
 product q
WHERE p.product_name \approx q.product_name

CrowdSQL



简单的解决方案

- 智能任务(HIT)

$O(n^2)$ X

Decide Whether Two Products Are the Same or Different

Product Pair #1

Product Name	Price
iPad Two 16GB WiFi White	\$490
iPad 2nd generation 16GB WiFi White	\$469

Your Choice (Required)

☒ They are the same product

☐ They are different products

Reasons for Your Choice (Optional)

$n=10,000$
\$0.01/HIT  \$1,000,000

分批策略

- 基于簇的 HIT

$O(n^2/k^2) \times$

Find Duplicate Products In the Table. ([Show Instructions](#))

Tips: you can (1) **SORT** the table by clicking headers;
(2) **MOVE** a row by dragging and dropping it

Label	Product Name	Price ▲
1 ▼	iPad 2nd generation 16GB WiFi White	\$469
1 ▼	iPad Two 16GB WiFi White	\$490
2 ▼	Apple iPhone 4 16GB White	\$520
▼	iPhone 4th generation White 16GB	\$545

Reasons for Your Answers (Optional)

$n=10000, k=20$
\$0.01/HIT



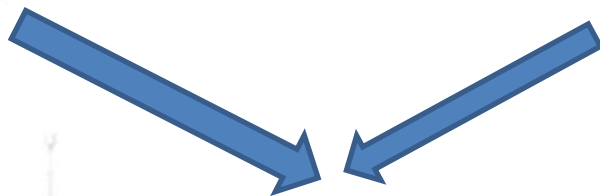
\$2,500

机器+人

钱 | 时间 | 质量



钱 | 时间 | 质量



钱 | 时间 | 质量

混合人机工作流程

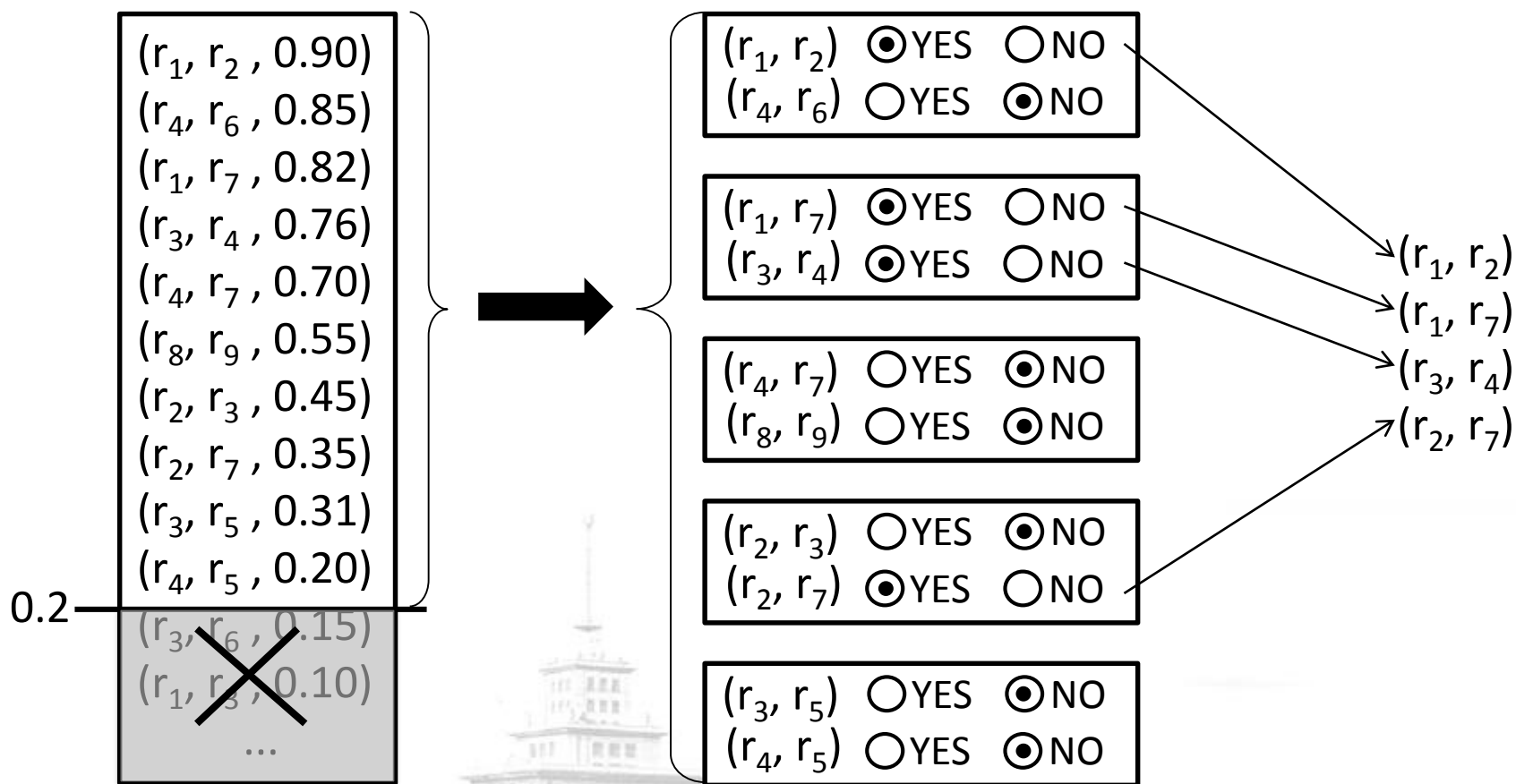
ID	Product Name	Price
r_1	iPad Two 16GB WiFi White	\$490
r_2	iPad 2nd generation 16GB WiFi White	\$469
r_3	iPhone 4th generation White 16GB	\$545
r_4	Apple iPhone 4 16GB White	\$520
r_5	Apple iPhone 3rd generation Black 16GB	\$375
r_6	iPhone 4 32GB White	\$599
r_7	Apple iPad2 16GB WiFi White	\$499
r_8	Apple iPod shuffle 2GB Blue	\$49
r_9	Apple iPod shuffle USB Cable	\$19



$(r_1, r_2, 0.90)$
$(r_4, r_6, 0.85)$
$(r_1, r_7, 0.82)$
$(r_3, r_4, 0.76)$
$(r_4, r_7, 0.70)$
$(r_8, r_9, 0.55)$
$(r_2, r_3, 0.45)$
$(r_2, r_7, 0.35)$
$(r_3, r_5, 0.31)$
$(r_4, r_5, 0.20)$
0.2
$(r_3, r_6, 0.15)$
$(r_1, r_3, 0.10)$
...

(a) 去除可能性 < 0.2的对

混合人机工作流程



HIT 生成

- 为了众包，必须将给定记录对的集合组合到HITs中。
 1. 基于对的 HITs
 2. 基于簇的 HITS



基于对的HIT

Decide Whether Two Products Are the Same ([Show Instructions](#))

Product Pair #1

Product Name	Price
iPad Two 16GB WiFi White	\$490
iPad 2nd generation 16GB WiFi White	\$469

Your Choice (Required)

- ☒ They are the same product
☐ They are different products

Reasons for Your Choice (Optional)

Product Pair #2

Product Name	Price
iPad 2nd generation 16GB WiFi White	\$469
iPhone 4th generation White 16GB	\$545

Your Choice (Required)

- ☐ They are the same product
☐ They are different products

Reasons for Your Choice (Optional)

Submit (1 left)

Figure 3: A pair-based HIT with two pairs of records

基于簇的 HIT

Find Duplicate Products In the Table. ([Show Instructions](#))

Tips: you can (1) **SORT** the table by clicking headers:
(2) **MOVE** a row by dragging and dropping it

Label	Product Name	Price ▲
1 ▼	iPad 2nd generation 16GB WiFi White	\$469
1 ▼	iPad Two 16GB WiFi White	\$490
2 ▼	Apple iPhone 4 16GB White	\$520
▼	iPhone 4th generation White 16GB	\$545

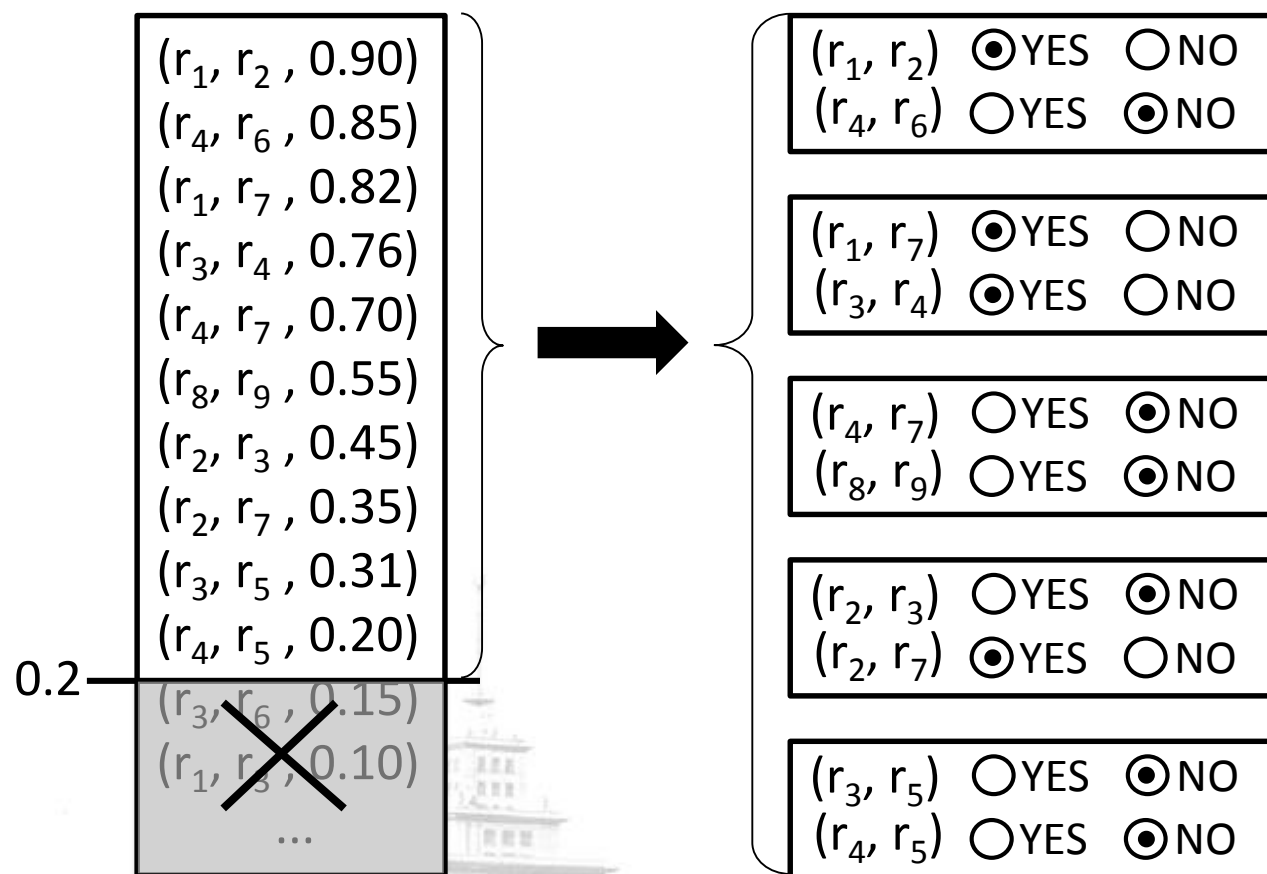
1
2
3
4

Reasons for Your Answers (Optional)

Submit (1 left)

Figure 4: A cluster-based HIT with four records.

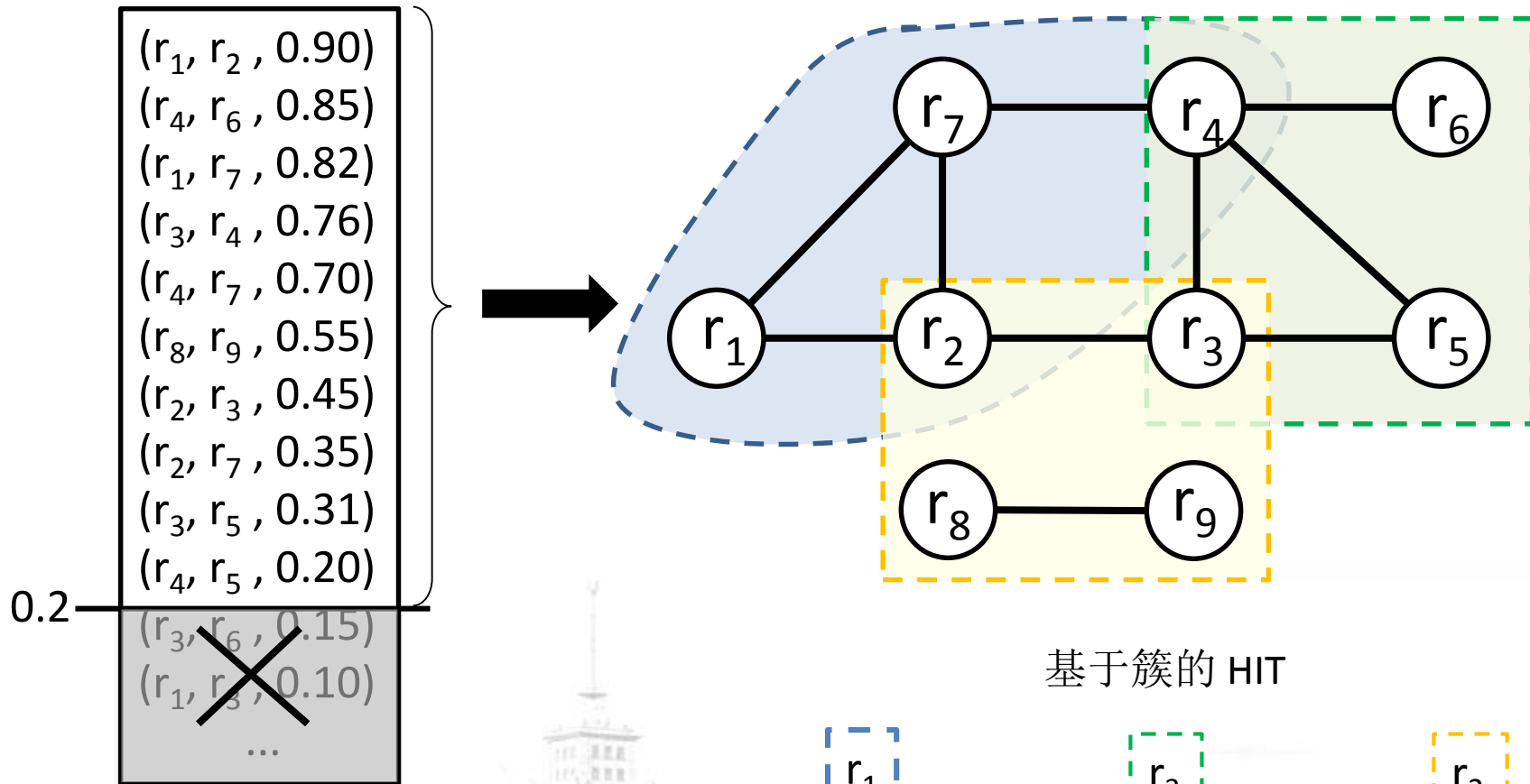
基于对的 HIT生成



基于簇的 HIT 生成

- 目标：给定一个记录对集合 P 和簇大小阈值 K ，基于簇的HIT生成问题的优化目标是生成最小数目基于簇的HIT: H_1, H_2, \dots, H_h , 满足如下两个约束:
 1. 对于任何 $\ell \in [1, h]$, $|H_\ell| \leq k$, 其中 $|H_\ell|$ 表示 H_ℓ 中的记录数
 2. 对于任何 $(r_i, r_j) \in P$, 存在 H_ℓ ($\ell \in [1, h]$) 使得 $r_i \in H_\ell$ 和 $r_j \in H_\ell$

基于簇的 HIT 生成



基于簇的 HIT

簇大小阈值 k
最小化 HITs 的数量

NP-Hard

r_1
 r_2
 r_4
 r_7

HIT₁

r_3
 r_4
 r_5
 r_6

HIT₂

r_2
 r_3
 r_8
 r_9

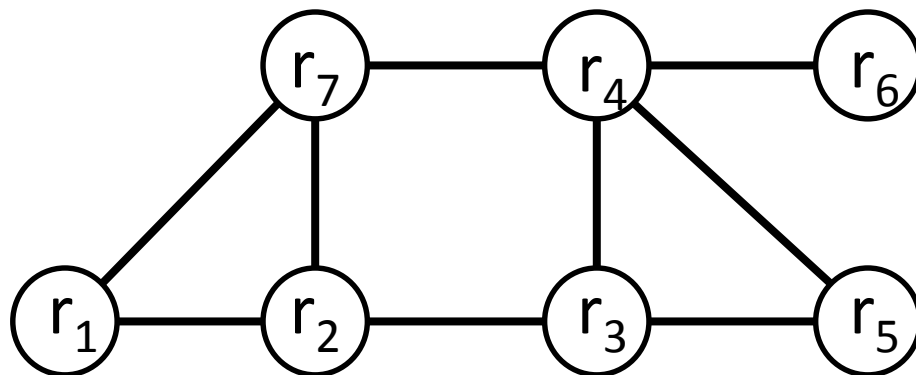
HIT₃

双层法

$(r_1, r_2, 0.90)$
$(r_4, r_6, 0.85)$
$(r_1, r_7, 0.82)$
$(r_3, r_4, 0.76)$
$(r_4, r_7, 0.70)$
$(r_8, r_9, 0.55)$
$(r_2, r_3, 0.45)$
$(r_2, r_7, 0.35)$
$(r_3, r_5, 0.31)$
$(r_4, r_5, 0.20)$
$(r_3, r_6, 0.15)$
$(r_1, r_5, 0.10)$
...



大连通分量(LCC)



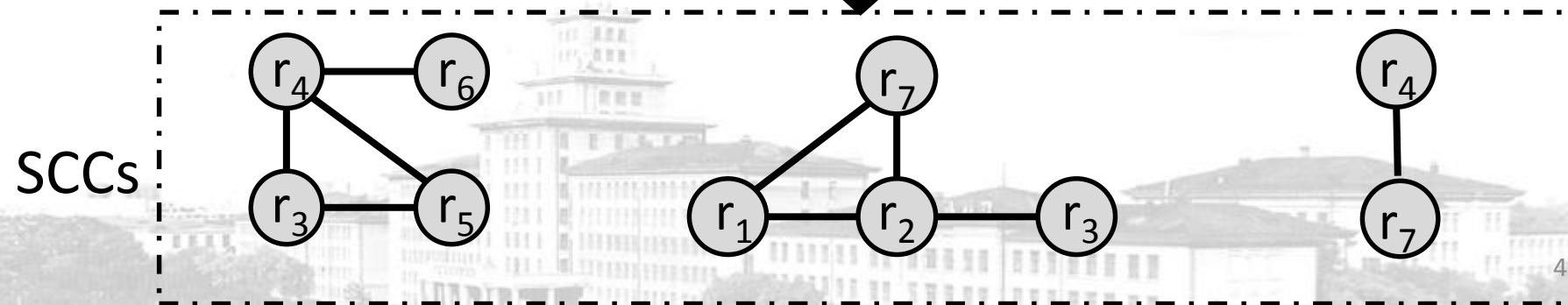
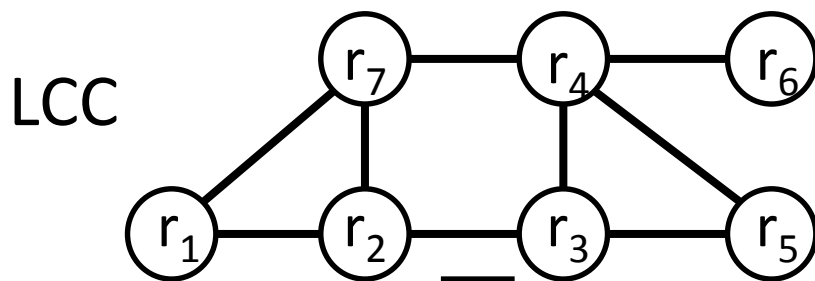
小连通分量(SCC)

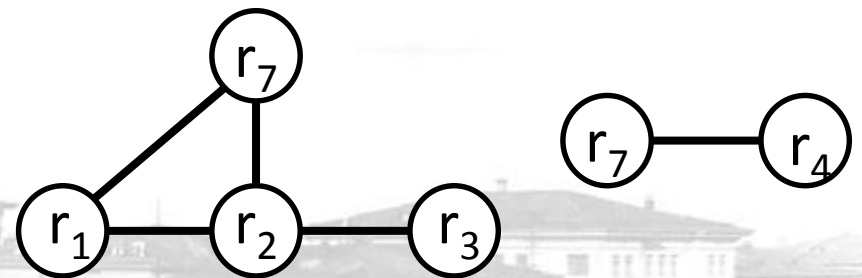
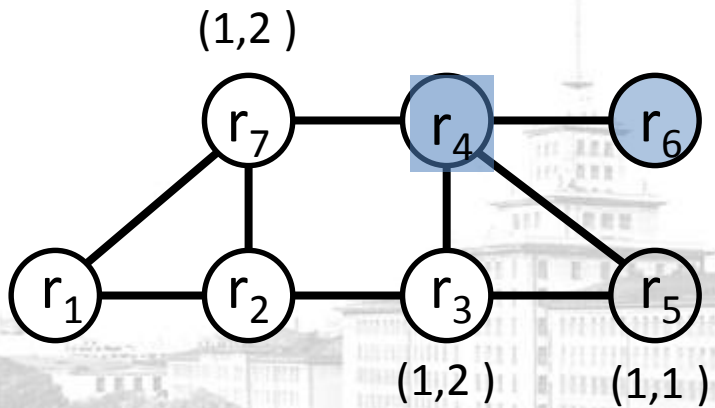
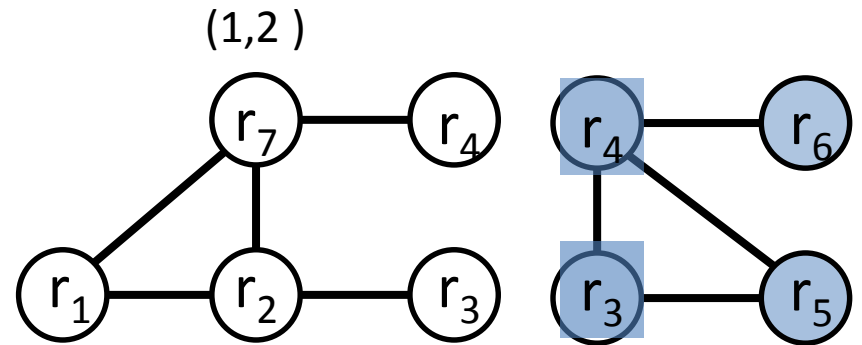
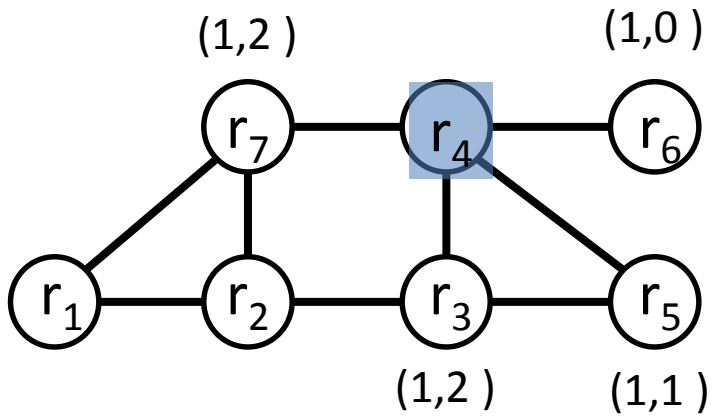
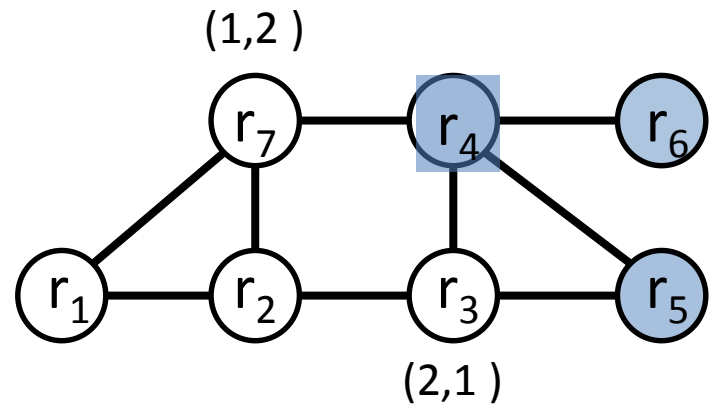
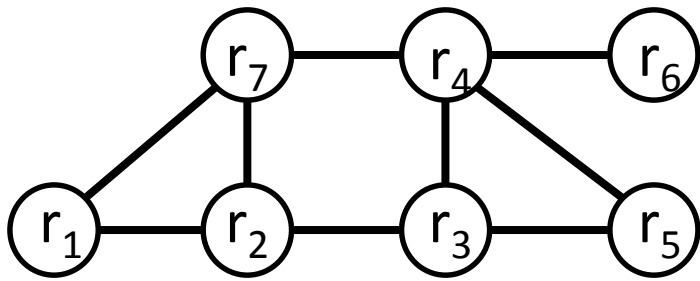


假设 $k=4$

步骤1: LCC 分区

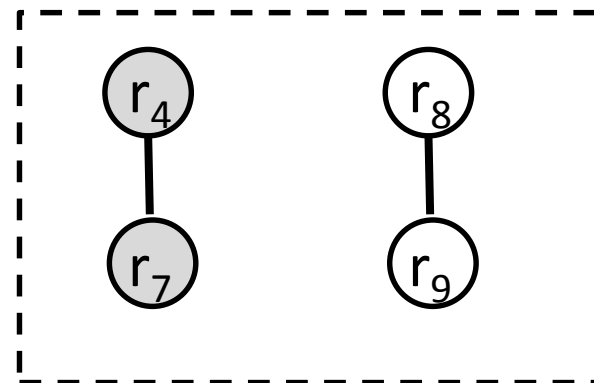
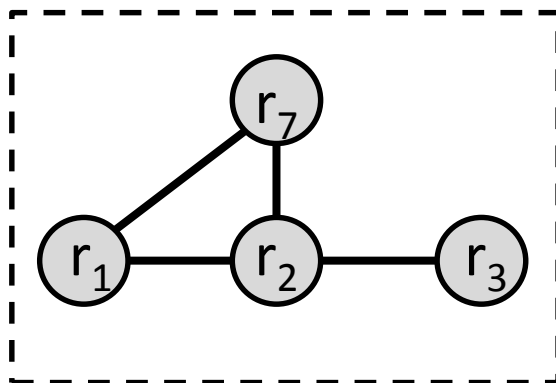
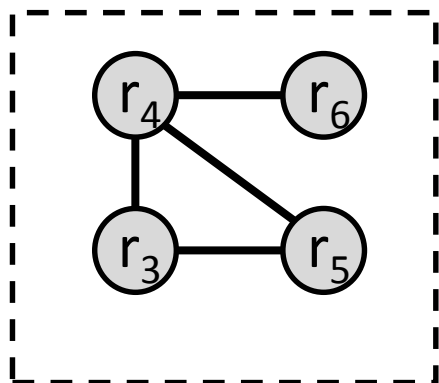
- 基本思路
 - 将LCC划分成高度连接的SCC
- 贪心算法





步骤2: SCC包装

- 一维下料问题[Gilmore et al. OR'61]



r_3
 r_4
 r_5
 r_6

HIT₁

r_1
 r_2
 r_3
 r_7

HIT₂

r_4
 r_7
 r_8
 r_9

HIT₃

总结

