

众包技术研究综述

冯剑红 李国良 冯建华

(清华大学计算机科学与技术系 北京 100084)

摘 要 众包通过整合计算机和互联网上未知的大众来完成机器单独难以处理的任务. 自从 2009 年开始, 众包得到了各个领域的广泛关注, 包括人机交互、信息检索、机器学习、数据库、人工智能、计算机理论等, 已经逐渐成为一个新研究热点, 在基础理论和工程技术方面向科技工作者提出了大量具有挑战性的问题. 文中介绍了众包的基本概念以及工作流程, 对众包在计算机领域中已有的研究工作进行了综述. 基于众包工作流程, 从众包任务准备、任务执行和任务答案整合 3 个方面综述了现有技术研究以及存在的挑战, 最后展望了未来可能的研究方向, 并提供了大量参考文献.

关键词 众包; 机器难问题; 任务准备; 任务执行; 任务答案整合

中图法分类号 TP399 DOI号 10.11897/SP.J.1016.2015.01713

A Survey on Crowdsourcing

FENG Jian-Hong LI Guo-Liang FENG Jian-Hua

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084)

Abstract Crowdsourcing can effectively solve machine-hard tasks by leveraging machine and a large group of people on the web. Many research communities, including human-computer interaction, information retrieval, machine learning, database, artificial intelligence, theory, have been engaged in crowdsourcing since 2009. Crowdsourcing is a promising research topic and has many research challenges. In this paper, we introduce the basic concepts and workflow of crowdsourcing. Based on the workflow, we provide a detailed survey from the aspects of crowdsourcing task preparation, task execution and answers integration. Finally, possible research directions of crowdsourcing and plenty of related references are given.

Keywords crowdsourcing; machine-hard task; task preparation; task execution; answer integration

1 引 言

近年来, 众包得到了工业界和学术界的广泛关注. 众包(crowdsourcing)这一概念是由美国《连线》杂志的记者杰夫·豪(Jeff Howe)在 2006 年 6 月提出的^[1]. 杰夫·豪对“众包”的定义是: “一个公司或机

构把过去由员工执行的工作任务, 以自由自愿的形式外包给非特定的(而且通常是大型的)大众网络的做法. 众包的任务通常由个人来承担, 但如果涉及到需要多人协作完成的任务, 也有可能以依靠开源的个体生产的形式出现.”^①

传统的外包是指以合同(contract)的形式把任务委派给指定的人或者机构完成. 外包业务主要是

收稿日期: 2014-07-22; 最终修改稿收到日期: 2014-11-24. 本课题得到国家自然科学基金(61373024, 61472198)、国家“九七三”重点基础研究发展规划项目基金(2015CB358700, 2011CB302206)、北京市青年英才计划(YETP0105)资助. 冯剑红, 女, 1980 年生, 博士研究生, 主要研究方向为众包数据管理、数据库. E-mail: fengjh11@mails.tsinghua.edu.cn. 李国良, 男, 1980 年生, 博士, 副教授, 中国计算机学会(CCF)会员, 国家优秀青年科学基金获得者, 主要研究方向为数据库、信息检索、众包数据管理. 冯建华, 男, 1967 年生, 博士, 教授, 主要研究领域为数据库、信息集成.

① <http://wiki.mbalib.com/wiki/众包>

借助计算机能完成的任务,例如软硬件的开发和维护.软件提供商通过整合硬件和算法来完成外包的任务,例如制作网站,外包数据库,提供软件服务等.但是很多任务并不能通过简单的算法来实现,比如图像打标签,衡量两条记录是不是一个实体,评价一个商品等等.这类机器难以处理的问题(即机器难问题),可以借助于众包来完成.众包是直接发布到互联网,通过开放式集合(open call)互联网上未知的大众来解决传统计算机单独难以处理的问题^[2-3],例如 Wikipedia、reCAPTCHA^[4]、标记图像、语言翻译^[5]等.因为现实生活中存在着大量类似的问题,因此众包有着广泛的应用前景和市场.

由于众包有很多应用背景,例如,海量图像识别,评价网上商品的好坏等,因此近几年众包在计算机领域的学术界得到了非常广泛地关注,已经成为了研究的热点话题.很多重要会议都包含了众包的主题演讲(keynote)和专题讲座(tutorial),例如 VLDB 2011^[6]、SIGIR 2011^[7]、SIGIR 2012^[8]、WWW 2011^[9]、WSDM 2011^[10].他们主要讲述了众包的基本工作原理,众包技术研究进展以及遇到的问题.此外,目前已经有多个 workshop 专门讨论了众包技术研究存在的问题和面临的挑战,例如 VLDB 的 workshop DBCrowd 2013、KDD 的 workshop HCOMP 2009-2012、WWW 2012、SIGIR 2010-2011、CIKM 2012、NIPS 2011、NIPS 2013、CVPR 2010 的 workshops 等.

目前一些研究人员从不同角度对众包的相关研究工作进行了总结:Yuen 等人^[11]从应用、算法、性能和数据集 4 个方面总结了众包的进展.Kittur 等人^[12]阐述了众包在同步协作、实时响应和动机等 12 个方面所面临的挑战.Doan 等人^[13]回顾了万维网上应用的众包系统,根据问题类型和协作方式等方面对众包系统进行了分类总结.Zhao 等人^[3]从信息、技术、大众和组织机构 4 个角度综述了众包研究.本文不同于上述文章,本文着重介绍了众包的工作流程,以工作流程为基础,从任务准备、任务执行和任务答案整合 3 个方面总结了众包技术研究的重要进展和面临的研究挑战,并展望了未来的研究方向.

本文第 2 节给出了众包概念以及众包的主要研究领域;第 3 节提出了众包的工作流程;第 4 节对众包工作流程中的技术研究进行总结分析;第 5 节给出了众包的未來研究趋势;第 6 节总结了全文.

2 众包概念和主要研究领域

本节主要介绍众包的基本概念和主要研究领域.

2.1 众包基本概念

自 2006 年 Howe 首次提出众包的概念以来,学术界对众包的定义一直没有一个公认的结论.文献^[14]总结了 40 种不同的众包定义,这些定义从不同的角度对众包进行了描述,对比这些定义可以得到众包的基本特征:

- (1) 采用公开的方式召集互联网大众;
- (2) 众包任务通常是计算机单独很难处理的问题;
- (3) 大众通过协作或独立的方式完成任务;
- (4) 是一种分布式的问题解决机制.

根据这些特征,下面我们给出众包的定义.

定义 1(众包). 众包是一种公开面向互联网大众的分布式的问题解决机制,它通过整合计算机和互联网上未知的大众来完成计算机单独难以完成的任务.

根据大众参与众包的不同形式,众包被分为协作式众包(collaborative crowdsourcing)和竞赛式众包(crowdsourcing contest)^[3].协作式众包的任务是需要大众协作来完成的,并且完成任务的大众通常没有奖励回报;而竞赛式众包的任务通常是由个人独立完成,完成任务的个人会得到相应的奖励(如金钱报酬).

协作式众包中典型的成功案例主要包括维基百科 Wikipedia 和 reCAPTCHA.维基百科是开放的、自由的、免费的百科全书编辑平台,任何大众都可以进行编辑和修改.显然维基百科不能通过计算机软件技术来自动完成,也不能通过少数人来完成,因此需要众包的理念来实现.另外一个成功的案例是 reCAPTCHA. CMU 的教授 von Ahn 提出了 reCAPTCHA 项目,通过在验证码中嵌入书籍的扫描信息来完成纸质书籍的电子化^[4].

竞赛式众包中典型的例子是 Amazon Mechanical Turk (Mturk)^①.由于在竞赛式众包模式下,任务请求人需要为任务完成人支付报酬,因此 Mturk 采用市场的理念,为任务请求人和任务完成人构建了一个在线交易平台. Mturk 根据工人(任务完成人)和任务请求人不同的需要,提供了不同的服务. Mturk 平台上最常见的任务有:数据收集、分类工作等.

通过以上 3 个例子,不难看出,众包主要通过整合互联网上未知的大众和机器来解决机器难问题.

2.2 众包主要研究领域

从 2009 年起,众包在很多学科领域得到了广大科研人员的共鸣,其主要研究领域包括:

① <https://www.mturk.com>

(1) 人机交互领域^[15-21]. 由于众包需要和大众进行交互, 良好的人机界面设计对众包有重要影响. 因此人机交互领域主要研究如何设计良好的界面来提高众包的可用性.

(2) 数据库领域^[22-42]. 数据库领域主要研究如何提供一个类似于数据库的管理系统来管理众包的数据和大众, 通过查询语言和查询模型来规范化描述和表示众包系统.

(3) 自然语言处理领域^[43-50]. 由于自然语言处理领域的很多问题计算机都不能很好地解决, 因此他们利用大众来完成机器翻译^[43] 或者发现其中的语法错误^[45].

(4) 机器学习和人工智能^[51-56]. 机器学习和人工智能领域主要是通过大众的反馈学习新的知识, 以及利用大众的反应来做决策支持.

(5) 信息检索领域^[57-78]. 信息检索领域主要研究众包的结果质量问题. 由于大众完成任务会获得报酬, 因此部分大众可能会为了骗取报酬进行欺诈, 从而造成众包的结果质量难以保证. 为了解决这个问题, 研究人员通过研究各种有效机制来提高结果质量. 此外, 众包的结果质量问题也是多个领域都非常关心的问题.

(6) 计算机理论领域^[79]. 理论领域主要研究如何协调任务完成者和任务请求者之间的矛盾, 并考虑如何达到平衡.

由于协作式众包和竞赛式众包应用的范围不同(协作式众包通常是针对某种特定的任务构建应用系统, 而用户利用竞赛式众包可以解决不同类型的任务), 因此, 表 1 将各个领域研究的众包类型进行了对比, 统计了不同研究领域(文献[15-79])研究的众包类型数量. 从表 1 可以看出竞赛式众包更加通用, 应用更加广泛, 相应的技术研究更多, 因此本文主要针对竞赛式众包的技术研究进行分析总结.

表 1 不同领域研究的众包类型比较

| | 竞赛式众包 | 协作式众包 |
|-----------|-------|-------|
| 人机交互领域 | 5 | 2 |
| 数据库领域 | 20 | 1 |
| 自然语言处理领域 | 8 | — |
| 机器学习和人工智能 | 5 | 1 |
| 信息检索领域 | 21 | 1 |
| 计算机理论领域 | 1 | — |

3 众包工作流程

3.1 众包工作流程

众包的主要参与者包括任务请求人(requester)

和任务完成人——也叫做工人(worker). 他们通过任务(tasks)联系到一起. 图 1 给出了众包的典型工作流程.

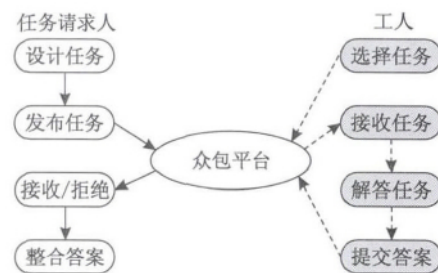


图 1 众包工作流程

如图 1 所示, 当任务请求人打算利用众包完成自己的任务, 需要按照以下步骤来使用众包.

- (1) 设计任务;
- (2) 利用众包平台发布任务, 等待答案;
- (3) 拒绝或者接收工人的答案;
- (4) 根据工人的答案整理结果, 完成自己的任务.

而工人使用众包的主要步骤包括:

- ① 查找感兴趣的任务;
- ② 接收任务;
- ③ 回答任务;
- ④ 提交答案.

从时间维度来考虑, 可以把众包的工作流程分成 3 个阶段: 任务准备、任务执行和任务答案整合. 其中任务准备阶段包括: 任务请求人设计任务、发布任务, 工人选择任务; 任务执行阶段包括: 工人接收任务、解答任务、提交答案; 任务答案整合阶段包括: 请求人接收/拒绝答案、整合答案.

3.2 众包任务分类

众包任务按照其复杂性、难度和粒度分为复杂任务(complex tasks)、简单任务(simple tasks)、宏观任务(macro tasks)和微观任务(micro tasks). 顾名思义, 复杂任务主要是完成一些非常复杂的、不能轻易完成的任务, 例如编写一个软件, 开发一个网站等. 简单任务主要是完成一个较容易完成的任务, 例如设计一个 logo, 翻译一段文字等. 宏观任务是指一个人较容易完成的任务, 宏观任务的粒度较大, 例如写一个饭店的评论. 而微观任务则粒度更小, 例如对一个图像进行打标签, 查找某个酒店的电话号码, 录入一张名片, 判断两条记录是否是同一个实体等.

3.3 众包平台

众包任务的发布和答案的收集是通过众包平台来完成的. 众包平台主要分为两大类: 一类是商用的众包平台; 另一类是社交网络、论坛等社交平台.

目前有很多商用的众包平台,其中主流的众包平台包括 Amazon Mechanical Turk(Mturk)、CrowdFlower^①、samasource^②、CloudCrowd^③等.国内也有很多众包平台,例如脑力库^④、猪八戒^⑤、三打哈^⑥等.商用的众包平台根据任务请求人和工人的不同需求提供相应的服务,并向任务请求人收取一定的管理费用.对于社交平台,如 facebook^⑦等,任务请求人通过在社交平台上嵌入自己的应用来完成.图 2 给出了不同众包平台应用比较,根据文献[15-79](文中第 2 节描述)中研究数据来源统计了众包平台的应用比例.从图 2 可以看出商用众包平台的应用明显多于社交平台,因此文中众包平台泛指商用众包平台.

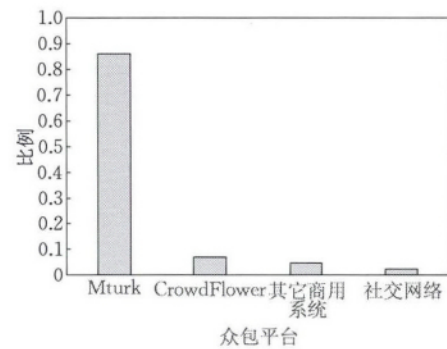


图 2 众包平台应用比较

我们从用户、功能服务、主要任务类型和界面 4 个方面对商用众包平台进行了对比分析,如表 2 所示.

表 2 众包平台对比分析

| 众包平台 | 用户 | 功能服务 | 任务类型 | 界面 |
|-------------|-------------------------------------|---|--------------|------------|
| Mturk | 任务请求人:企业或个人 工人:以个人为主 | 为请求人提供任务发布、接收结果等服务 为工人提供任务搜索、结果提交、收取报酬等服务 | 微观任务 | 图形化、API 接口 |
| CrowdFlower | 任务请求人:企业或个人 工人:其他众包平台上的工人 | 只为请求人提供服务:任务分解、任务设计、任务发布、结果质量检查等 | 微观任务 | 图形化 |
| samasource | 任务请求人:企业 工人:由 samasource 认证的贫困人员 | 为请求人提供任务分解、任务发布、结果整合等完整的服务 为工人进行基本的培训 | 微观任务 | 未公开 |
| CloudCrowd | 任务请求人:企业 工人:以个人为主 | 为请求人提供任务设计、结果整合等完整的业务服务 为工人提供结果提交、收取报酬等服务 | 微观任务 | 图形化 |
| 脑力库 | 任务请求人:以企业为主 工人:专业设计人员 | 为请求人提供任务发布、保密任务等服务,为长期发布任务的请求人提供专栏显示 为工人提供与请求人沟通、收取报酬等服务 | 宏观任务 | 图形化 |
| 猪八戒 | 任务请求人:企业或个人 工人(服务商):企业或个人 | 为请求人提供选择服务商、发布任务需求等服务 服务商可以发布提供的服务内容,服务商可以交换意见 | 宏观任务 微观任务 | 图形化 |
| 三打哈 | 任务请求人:以企业为主 工人:以个人为主 | 为请求人提供推广任务方案、任务渠道等服务 为工人提供任务搜索、工人交流等服务 | 宏观任务 微观任务 | 图形化 |

从表 2 可以看出,在主流的商用众包平台上发布的任务主要是微观任务,每个任务叫做 Micro Task(微任务),其主要原因包括:

(1) 工人更喜欢完成一些粒度比较小的任务,这样可以在很短的时间内完成一个任务.而复杂任务很难在短时间内完成,需要花费工人很长时间.由于工人一般利用较短的空闲时间来完成的任务^[11],因此工人通常不会挑选复杂任务.

(2) 任务请求人很难衡量复杂任务的价格,因此不利于报酬的付给.

(3) 微观任务具有更好的交互性,任务请求人可以多次和工人交互获得更加准确的答案,此外任务请求人还可以通过众包来多次尝试自己新的想法.

Mturk 将微观任务定义为 Human Intelligence Task(HIT),由于 Mturk 的广泛应用(如图 2 所示),因此微观任务的描述通常采用 HIT.

商用众包平台的界面以图形化界面为主(如表 2 所示),图形化界面由于有图形界面的支持,让任务设计简单明了,更适用于新手.而 API 接口则更加灵活,任务请求人可以根据自己的任务需求嵌入相应的程序.图形化界面和 API 接口各有各的好处:图形化界面通过界面的方式帮助任务请求人发布任务,主要包括设计、发布、管理等几个标签.而众包平台提供的 API 接口主要包括创建任务(包括任务需求、支付钱数、任务描述、返回任务 id 和任务类型)、提交任务、接收工人的答案、付给工人报酬、付给满意工人更多的奖励、向工人发送消息等^[28].

- ① <http://crowdfunder.com/>
- ② <http://samsource.org/>
- ③ <http://www.cloudcrowd.com/>
- ④ <http://www.naoliku.com/>
- ⑤ <http://www.zhuhajie.com/>
- ⑥ <http://www.sandaha.com/>
- ⑦ <http://www.facebook.com/>

4 众包工作流程中关键性技术研究

第3节将众包的工作流程分为了3个阶段,下面分别总结了这3个阶段中的关键性技术。

4.1 任务准备

在任务准备阶段,主要挑战包括:

(1) 如何将复杂任务分解成微观任务,从而利用众包来解决复杂任务?

(2) 如何对任务赋予合适的价格?

(3) 如何处理欺诈者?

(4) 如何平衡任务花费、质量和时间?

(5) 如何设计任务界面?

(6) 工人如何挑选感兴趣的任务?

4.1.1 任务设计

(1) 复杂任务分解. 由于众包更适用于解决微观任务,因此为了支持复杂任务,我们需要把复杂的任务分解为微观任务,然后把微观任务分发给工人,并收集工人对微观任务的答案,最后整合工人答案来完成自己的初始任务. 对复杂任务进行分解,需要结合任务的特征,构建各个子任务之间的依赖关系,从而尽可能最大化每个子任务的独立性. Kittur 等人^[17,20]研究了如何分解复杂任务以及如何整合工人的答案来完成初始任务,并提出了基于 MapReduce 的框架来实现任务的分解. 但是他们的方法只适合于特定类型的任务,通用效果不好,可扩展性仍需要解决。

(2) 任务标价. 如何对一个任务赋予合适的标价也是目前研究的热点. Mason 等人^[80]发现,一个任务的价格过高或者过低都会对任务的完成带来一定影响. 如果价格过高可以吸引更多的工人来回答该任务,但是并不会提高任务的完成质量,反而会增加任务请求人的金钱负担;此外,任务价格过高容易吸引欺诈者来完成任务,导致任务的结果质量不高^[81];而如果任务价格过低,工人则不感兴趣,导致任务很难被及时完成. 通过研究发现,微任务价格一般定在 1~5 美分。

任务的难度和任务的颗粒是任务定价中两个重要的因素,任务请求人需要综合这两方面因素为任务定价,比如有些任务难而不烦,有些任务烦而不难. 特别是有些任务要求工人有专业背景知识(例如语言翻译),如果这类任务的价格过低,将很难吸引合格的工人. 任务请求人在定价之前可以参照众包平台上类似任务的标价,再结合任务的特点给出任

务标价。

此外,通过使用 Mturk 会发现:如果任务请求人发现一些工人返回的答案质量较差,拒绝赋予工人报酬,则工人会进行争辩,并发布相关消息,从而影响请求人的口碑,降低其他工人对任务的参与度. 解决这个问题的一般方法是,如果一个任务是 2 美分,我们可以将 1 美分作为报酬,而另外 1 美分作为奖励,只有工人提供的答案质量较高,才付给工人奖励金额,从而提高完成任务的结果质量。

(3) 欺诈者处理. 由于回答每个任务,工人都可以获取一定的报酬,一部分工人会进行欺诈从而获取更多报酬,进而造成了工人提交的答案质量难以保证. 在任务执行之前可以通过添加测试题目(qualification test)来测试工人是否了解该任务的相关知识,这样不但可以剔除一部分欺诈者,而且也可以剔除一部分对任务不了解的工人,还可以让工人更熟悉任务,进而提高结果质量. 但是很多工人则会因为测试,拒绝回答这些任务,造成只有少量工人回答这些任务的情况,导致任务长时间不能被完成. 此外,欺诈者可以先通过测试然后再进行欺诈,从而使得任务执行前的测试失效。

另外,任务请求人可以在任务中随机添加一些常识问题,测试工人是否为欺诈者. 如果工人不能正确回答这些问题,便可以尽早剔除这些欺诈者,从而不但可以提高结果质量,还可以避免金钱的浪费. 但是如何插入常识知识而且不被工人发现是一个非常具有挑战性的问题。

除了增加测试题目之外,Eickho 等人^[82]发现欺诈者通常不会选择比较创新的和抽象的任务,而且,欺诈者通常选择的任务是包含的问题数目较多的任务. 因此,针对欺诈者的特性,任务请求人通过有效地设计任务可以主动地抵制欺诈者参与完成任务。

(4) 花费、质量、时间的平衡. 由于众包平台并不能够保证结果质量,因此结果质量控制一直是科研人员研究的重点. 由于工人和任务请求人二者之间有着不可调和的矛盾(工人目的是得到更多的报酬,任务请求人则是想通过最少的钱快速地完成最多的任务),因此如何平衡任务的花费、任务的结果质量、任务完成时间三者的关系是任务请求人在任务设计阶段面临的重大挑战。

Ipeirotis^[5]发现,大多数任务请求人都将任务的“过期时间”设置为 12 小时或 7 天,在 12 小时这个时间节点,只有大约 50% 的任务被完成,如果等到 7 天,大约 90% 的任务被完成. 也就是说,请求人提

交的任务很可能需要很长的时间才能 100% 被完成。任务完成的时间和很多因素相关,例如任务是否吸引工人、在众包平台上活跃的工人数量、任务的发布时间、任务的类型和任务的价格等等^[83]。Wang 等人^[83]通过基于 Cox 风险比例模型的生存分析方法对任务的完成时间进行了分析,他们利用从 Mturk 上收集的数据研究了众多影响因素(如请求人提交的任务数量等等)与任务完成时间的关系,结果表明预测一个任务的完成时间是非常困难的,因此,请求人设计任务的时候需要综合考虑多方面因素。

一个任务的结果质量和很多因素有关,例如任务的难度、工人的负责程度、工人和任务的相关性等等。此外,由于依赖一个工人给出的答案很难确保任务结果质量,因此任务请求人通常会将任务分配给多个人,然后在任务答案整合阶段利用不同的策略推断任务的最终结果(第 4.3 节会详细介绍推断结果的策略)。

由于任务的金钱花费一般与任务数目和任务标价相关,因此一般是可控的。任务请求人雇佣工人完成任务是需要付费的,尽管每个任务只需要几美分,但是当任务数量很多的时候,单纯依赖工人来完成所有的任务需要的花费是巨大的。在发布任务之前,针对任务要求进行预处理可以减少任务数目。文献^[31, 34, 71]针对众包环境下的 join 操作提出了将机器与人相结合的技术,首先通过机器算法对问题进行过滤,再将剩余的问题交给工人。文献^[34]则利用实体的传递关系,进一步减少任务数量,从而节省花费。Lofi 等人^[37]针对不完整数据集的 skyline 查询,通过“错误模型”对包含缺失数据的数据集进行预处理,得到需要工人回答的问题,从而减少任务的花费。

此外,结合任务特征制定合适的任务提问形式也可以减少任务数量。麻省理工大学的 Marcus 等人^[30]针对 join 和 sort 操作分别提出了不同的任务提问方法。针对 join 操作给出了 3 种方法,第 1 种是简单的比较方法,一个任务只包含一对元素;第 2 种是简单的批量方法,一个任务包含多对元素;第 3 种是智能的批量方法,任务不再是元素对的形式,而是一组单独的元素,工人需要在这些元素中查找匹配的元素。对比这 3 种方法,智能的批量方法产生的任务数量最少。对于 sort 操作,他们提出了基于比较和基于打分的任务提问方法。在基于比较的方法中,需要让工人将一组元素中所有的关系进行比较,根据比较的结果,得到最终的排序。在基于打分的方法

中,则是让工人直接对待评估的元素进行打分,依次作为排序的依据。对比两种方法,基于打分的方法产生的任务数量明显少于基于比较的方法。

设计问题时,当只考虑花费、质量、时间 3 个方面中的某一个方面,是比较容易控制的,但是花费、质量、时间三者之间的关系是很难衡量的,因此目前很多科研人员研究如何在三者之间进行平衡,从而达到某种优化目的。例如,在给定金钱的情况下,希望最短时间完成任务;在给定质量保证的要求下,希望花费最少的金钱来完成指定的任务等。

例如,在图片分类中,判断一个图片(例如雅阁汽车)属于哪一类(例如机动车、本田、丰田、雅阁),需要问工人一些问题,如何通过问最少的问题(花费最小)来完成分类的任务便是一个需要解决的优化问题,Parameswaran 等人^[41]提出多种优化算法来减少金钱开销。

另外一个例子,在图片搜索中,由于现有算法很难保证图片搜索质量,而人类非常容易进行图片质量的判断。因此 Yan 等人^[84]研究如何在质量控制范围之内(例如要求结果准确率大于 95%),花费最少的金钱来完成图片的搜索。

(5) 任务界面。通过任务界面的描述,工人能够获取有关任务的具体信息,全面地了解任务的相关要求,因此任务界面设计非常重要。任务界面的研究重点是如何设计友好的界面来吸引工人,从而便利工人完成任务。

众包平台上包含了成百上千万的活跃任务,由于当前的众包平台提供的界面设计功能都比较简单,因此很多任务由于界面设计的问题,难以吸引工人,导致很多任务在一个月或者更长时间都没有被完成^[5]。Sakamoto 等人^[18]研究了众包参与者在不同任务类型中常用的交互方式。Heer 等人^[21]研究了如何进行问卷调查,发现通过调查问卷来设计界面更适合于众包工人。由于任务请求人的主要目的是为了向工人直观、完整地展现任务内容,因此界面的简洁、清晰是设计中首要考虑的因素。

4.1.2 任务选择

任务选择的研究重点是如何帮助工人挑选与自己相关的任务。按照任务的发现者来分,可以分为两种主流的任务选择方法:一种是基于拉(pull)的方法,一种是基于推(push)的方法。基于拉的方法是由工人主动查找相关任务,即任务搜索;而基于推的方法则是由众包平台进行任务的分发,即任务推荐。

(1) 任务搜索。主流的众包平台都是采用基于

拉的方式。目前众包平台对任务搜索的支持能力较弱,其中的大多数众包平台只提供任务列表,工人只能通过浏览任务页面来获取自己感兴趣的信息,而不能支持与工人兴趣相关的任务查找;由于众包平台对任务采用分页显示,如果任务出现在第3页或者更靠后的页面,将很难被工人发现和关注,而且工人通常挑选最近发布的任务来完成^[85],从而很容易导致有些任务长时间不能被完成。Mturk等众包平台提供了基于关键字的任务搜索功能,由于只考虑了关键字与任务内容的相关程度,没有考虑工人的兴趣偏好等因素,所以返回的搜索结果对所有的工人都是一样的。总的来说,当前的任务搜索功能还不能满足工人的个性化需求,难以有效地帮助工人查找感兴趣的任務。表3给出了主流众包平台提供的任务搜索功能以及任务内容显示的排序方法。

表3 任务搜索功能和任务显示排序方法比较

| 众包平台 | 关键词搜索 | 地域搜索 | 类别搜索 | 按时间排序 | 按价格排序 |
|------------|-------|------|------|-------|-------|
| Mturk | √ | | | √ | √ |
| CloudCrowd | √ | | | √ | √ |
| 脑力库 | √ | | | √ | √ |
| 猪八戒 | √ | √ | √ | | √ |
| 三打哈 | √ | | √ | √ | √ |

(2) 任务推荐。基于推的方式不需要工人主动地输入查询,主要是根据工人的兴趣爱好,众包平台主动地进行相关任务的个性化推送。在众包平台上保留的历史记录是工人行为的最好体现,比如,工人完成了哪些任务,任务的类型,哪些工人的答案被任务请求人认可等等。Ambati等人^[86]提出了利用任务和工人的历史信息进行任务推荐的方法,通过构建工人的偏好模型为工人推荐可能感兴趣的任務。另外,随着时间的变化,工人的兴趣也会发生改变,最近的浏览记录更能准确地反映工人的当前兴趣。Yuen等人^[87]提出将工人的历史信息和浏览记录相结合来反映工人的兴趣偏好,并对结合后的信息进行细化处理,将工人的行为表现分为6个类别,以此构建“工人-任务”矩阵,最后通过概率矩阵分解(Probabilistic Matrix Factorization)实现任务的个性化推荐。

此外,某些特定类型的任务,比如空间任务(基于地理位置和时间的任务),任务推荐还需要结合工人的位置信息。Deng等人^[39]研究了空间任务的最大化任务选择问题,以完成的任務数量最多为目的,为工人推荐一个任务选择序列。他们综合任务的截止时间和工人的行程消耗两方面因素,提出了

基于动态规划和分支定界的算法,这两种算法主要适用于任务数量不多的情况,随着任务数量的增多,算法的运行时间将指数倍地增长,所以他们又提出了一系列改进算法,从而尽量减少任务选择的时间。

4.2 任务执行

在任务执行阶段,主要挑战是:如何有效地结合工人因素、请求人的任务优化目的进行在线任务分配?

在任务执行的过程中,通过有效的在线任务分配策略,有针对性地将任务分配给工人,可以提高任务完成结果的质量。文献[66]提出了一种基于随机图生成和消息传递的任务分配方法,这种方法的局限性在于只针对了特定类型的任务(任务的难易程度一致),而众包平台上的任务种类繁多,某些任务需要特殊的专业知识,比如语言翻译类任务。针对这一问题,文献[88-89]提出了一种类似关键词竞价广告的方法,该方法基于原始对偶理论,结合了请求人给定的任务预算,依据工人的能力水平,将难易程度相匹配的任务分配给工人。以上的分配方法都只是通过模拟数据或离线数据进行了实验分析,实验主要针对任务结果的质量进行了评价,都没有在真实的众包环境下验证分配方法的结果质量和运行效率。张志强等人^[90]提出了结果评估与替换策略来实现任务动态分配,任务请求人开始只分配部分任务,然后根据返回的答案评估结果的质量和工人的能力,对完成结果质量不高的工人采用替换方法,将这些工人完成的任务进行再次分配,从而提高任务的结果质量。

此外,由于任务请求人通常是在任务设计的时候根据经费预算和任务难度为任务设定标价,定价过程中一般没有工人参与,而任务标价对工人的影响是难以预知的,因此任务标价不应该单纯依赖任务请求人的需求,任务执行过程中可以结合不同工人对任务的预期标价进行任务分配。文献[91]提出了一种在线定价机制,这种定价机制综合考虑了任务请求人与工人两方面因素:一方面,任务请求人不需要提前为每个任务设定标价,只需要提供任务的总体预算或任务的总数量;另一方面,当某个工人申请回答任务的时候,工人需要给出期望的任务标价以及期望完成的任务数量。利用以上两方面信息,价格机制实时地决定任务的实际价格以及分配给工人的任务数量。对比任务准备阶段的定价策略(4.1节),表4给出了两种定价策略的比较。

表 4 任务定价策略的比较

| 定价策略 | 定价依据 | 策略评价 |
|------|----------------|--|
| 固定标价 | 任务难度、任务颗粒、奖励机制 | 优点:实现简单 缺点:价格过低导致任务完成时间增加;价格过高导致任务花费代价大 |
| 在线标价 | 工人预期 | 优点:价格更加合理 缺点:要对任务价格进行多次调整,增加了任务选择时间 |

4.1 节中提到任务的质量、花费和时间很难达到平衡,众包任务通常是以满足某种优化为目的.为实现相应的优化目标,在任务执行的过程中可以采用不同的分配策略和优化算法. Liu 等人^[78]实现的数据分析系统以保证结果质量为主要目的,首先通过一个预测模型给出了分配的任务数量(在确保结果质量的前提下,每个任务需要被工人回答的次数),然后,在任务执行的过程中,通过在线地评估结果质量来判断是否提前终止任务,从而节省花费和时间. Gao 等人^[64]则综合考虑了结果质量与任务花费,提出了一种在线决策方法,通过动态分配任务、预测“花费-质量”和终止任务等模块,实现花费、质量和时间的平衡. Kazemi 等人^[38]针对服务器模式的空间任务特征(在线的工人需要先上传自己的位置信息),提出了三种任务分配方法:贪心策略、位置熵最小优先策略和最近邻优先策略.这三种方法分别实现了不同的优化目标,贪心策略主要考虑当前时刻的信息,可以迅速地分配任务,但是没有综合考虑全局任务信息和工人信息,因此难以保证任务的完成质量.另外两种策略则着重考虑了任务的结果质量,其中,位置熵最小优先策略利用位置熵表示任务被完成的可能性,位置熵越小代表活跃在任务区域内的工人越少,即任务被完成的可能性越小,通过将位置熵较低的任务分配给工人,进而实现全局的最优分配;最近邻优先策略则结合工人行程消耗(工人与任务的距离),选择距离较近的任务分配给工人.

4.3 任务答案整合

在任务答案整合阶段,面临的最大挑战是:如何处理工人提供的答案?

由于参与完成任务的工人是来自于不同的国家和地区,并且他们的年龄、教育背景不尽相同^[92],而工人回答问题的时候不可避免地受到主观意识和知识背景的影响,因此当工人完成不同类型任务的时候,答案的质量变化较大.为了保证任务结果质量,很多研究人员提出了不同的控制策略.

在工人提交答案后,可以通过各种算法来保证结果的质量.最简单的办法是把一个任务分配给多个(奇数)工人来完成,然后通过多数投票原则(少数服从多数)来获取最终结果,文献^[28,33,35,84]中均采用了这种投票原则.由于多数投票方法是假定每个工人的答题准确率是一致的,没有考虑工人的多样性,而通常不同工人的答题准确率差异较大,比如,欺诈者的答题准确率较低,因此采用这种方法得到的最终结果往往不够准确.

针对这一问题,一些研究做出了改进,将工人的答题准确率运用到结果的估计中,使得最终结果的质量有了很大的提高. Liu 等人^[78]通过增加测试题目得到工人的答题准确率,利用贝叶斯理论将工人的答题准确率和工人给出的答案结合起来得到最终的结果.文献^[71]提出了一种概率模型,该模型基于因子图,通过综合工人答案、工人的答题准确率等因素得到结果.这两种方法都是区分了不同工人的答题准确率,但是都假定了工人的答题准确率是固定的,即:在完成任务的过程中,工人的答题准确率保持不变.然而,随着时间的变化,工人的答题准确率通常是变化的,比如,工人在完成任务过程中,随着对任务了解的增多,他的答题准确率会越来越高.文献^[51,59]提出了一种反映工人答题准确率变化的方法.这种方法是基于 EM (Expectation-Maximization) 算法^[93]得到的最终结果,通过混淆矩阵(confusion matrix)来反映工人的答题准确率. EM 算法需要两个步骤进行迭代计算,直到算法收敛.第 1 步利用已有的工人答题准确率估计值,对所有的问题分别进行计算,得到每个问题结果的估计值;第 2 步是利用第 1 步得到的结果来计算每个工人的答题准确率.由于 EM 算法每次迭代都要重新计算每个问题的结果和每个工人的答题准确率,因此当问题数目较多或参与回答问题的工人数量较多的时候, EM 算法的运行时间较长,代价较大.以上计算工人答题准确率的方法都难以实现准确性与实时性的平衡,文献^[94]提出了一种新的工人模型,通过该模型可以及时准确地得到工人答题准确率,他们利用工人每次返回的新答案,结合答题准确率,设计了两种增量式的策略来推断最终的任务结果,进而实现高效准确地得到任务结果.

此外,我们也可以利用工人的答题准确率来进一步提高多数投票原则得到的结果质量,其基本思想就是根据工人的答题准确率来对每个工人进行打分.回答准确率越高的工人赋予权重越大,相反则权

重较小,最后通过考虑权重进行加权评价工人提供的答案,根据加权分值确定最终结果.如表 5 所示,我们对现有的整合答案方法进行了比较.

表 5 答案整合方法比较

| 推断方法 | 基本思想 | 方法评价 |
|--------------|----------------------------|---|
| 单纯利用工人答案 | 多数投票原则 文献[28,33,35,84] | 优点:方法简单,利于实现 缺点:结果准确率不高 |
| 结合工人答案和答题准确率 | 固定的工人答题准确率 文献[71,78] | 优点:结果准确率较高 缺点:工人准确率的变化导致结果质量降低,需要预知工人答题准确率 |
| | 变化的工人答题准确率 文献[51,59,94] | 优点:结果准确率很高,不需要预知工人答题准确率 缺点:基于 EM 的方法时间代价较高 |

不难发现,任务完成后再进行结果质量评估不能剔除任何欺诈者,因此会白白浪费一部分金钱.尽管任务请求人可以拒绝赋予工人报酬,但是会带来坏口碑,从而影响其他工人来完成自己的任务,导致发布的任务长时间难以完成.

从 4.1 节和 4.2 节的内容可以发现,除了有效地处理工人的答案,任务结果的质量还与任务的设计、分配、预算、时间、工人的个人兴趣和能力等方面均有关联^[95],因此任务的质量控制贯彻了众包工作流程的每个阶段.文献[90]提出了一种阶段式动态的众包质量控制策略,这个策略通过任务设计与管理、工作者组织与管理与结果评估与替换 3 个方面来解决众包任务的质量控制,他们主要从结果评估与替换部分考虑,通过组合式质量评估算法、随机替换工人和设置质量评估监测点 3 个步骤来提高结果的质量.由于在任务执行过程中增加了监测点进行质量评估,必然就增加了任务完成时间,而且,重新分配完成质量不高的任务增加了任务的花费.因此,任务请求人应该结合任务的主要优化目的,在 3 个阶段中分别采用针对性的策略,从而达到任务质量、花费和时间的相对平衡.根据第 4.1 节~4.3 节讨论的众包技术,表 6 给出了 3 个阶段中不同技术对质量、花费、时间影响的比较.

表 6 3 个阶段中不同技术对优化目标的影响

| | 质量 | 花费 | 时间 |
|------|--------|----|----|
| 任务准备 | 复杂任务分解 | ✓ | |
| | 任务标价 | ✓ | ✓ |
| | 欺诈者处理 | ✓ | ✓ |
| | 任务界面 | ✓ | ✓ |
| | 任务数量 | ✓ | ✓ |
| | 任务选择 | ✓ | ✓ |
| 任务执行 | ✓ | ✓ | ✓ |
| 答案整合 | ✓ | | |

5 众包研究趋势

从众包概念的提出到现在只有几年的时间,学术界和工业界对众包已经开展了大量的研究工作,但是目前众包的研究还较初步,其中还存在一些研究问题值得进一步探讨.

5.1 大数据处理

借助众包来完成任务虽然便宜,但并不是免费的^[80],比如每个微观任务需要花费 1~5 美分,当处理大数据量,例如 1 百万级数据,将这些任务全部交给工人来完成则需要花费 1~5 万美元;如果是 1 亿级数据,则需要花费 100~500 万美元,显然这些花费是巨大的.任务请求人发布的任务通常是在受限的经费条件下完成的,因此在经费受约束的情况下如何处理大数据任务是未来研究的方向之一.针对大数据的处理可以从以下两个方面考虑:第 1 个方面,任务数量影响着任务的总体花费,通过有效地设计任务的生成方式可以减少产生的任务数量,进而节省任务花费. Marcus 等人^[30]提出了将每个任务包含多个子问题的策略,但是当任务包含的子问题数量较多的时候,任务的定价要提高,否则容易造成只有少量工人挑选任务的情况,也就是说这种方式虽然减少了任务数量,但是并不能确保减少任务的总体花费.因此如何在经费受限的情况下有效地设计大数据任务是一个具有挑战性的问题.第 2 个方面,利用机器算法给出符合任务要求的备选数据或者过滤掉不影响任务结果的数据,进而减少需要工人回答的问题数量.例如实体匹配任务中先利用机器算法过滤掉明显不匹配的数据^[31].由于不恰当的机器算法会错误地过滤掉影响任务结果的数据,进而影响任务结果的质量,因此如何发现大数据之间的关联规则,通过有效的规则剪枝方法来降低任务金钱开销也是一个不小的挑战.我们研究组提出了大数据群体计算(青年 973 项目),通过众包和机器的高效协作来解决大数据问题.

5.2 任务搜索

由于目前的众包平台对任务搜索的支持并不够,因此如何实现高质量的任务搜索仍是未来研究的趋势之一.任务搜索是从工人角度考虑的基于拉的任务发现模型,现有众包平台的任务搜索功能还较弱,通常是工人通过浏览任务列表来查找感兴趣的^[85]任务,个别平台提供了通过关键词搜索任务的功能,如 Mturk,工人可以输入感兴趣的^[85]任务关键词

来搜索相关任务,但是没有考虑不同工人的个人偏好,得到的结果不能充分满足工人的个性化需求.由于每个工人都有自己不同的偏好,因此需要进行个性化搜索.个性化信息需要通过历史任务来获取,通过挖掘工人历史完成的任务信息来获取其个人偏好,然后利用这些兴趣爱好返回他们可能感兴趣的任務.此外工人对不同任务有着不同的熟悉程度,我们可以通过回答任务的准确率来衡量这一指标.因此如何整合这些与工人相关的指标来实现个性化的任务搜索是值得进一步研究的内容.

5.3 任务推荐

任务推荐主要是从众包平台角度考虑的基于推的任务发现模型,众包平台根据工人的偏好推荐相关的任务.主流的众包平台还没有任务推荐的功能.现有的任务推荐研究主要是基于传统推荐的方法^[86-87],主要包括基于内容的推荐,协同过滤以及二者相结合的混合推荐算法.不同于传统推荐算法,我们还需要考虑工人在某个领域回答问题的准确率,即工人擅长的工作类型,从而利用这些信息进行推荐,而不能只关注相关性.此外为了保证每个任务尽快完成,我们还需要考虑更多因素(例如预测哪些用户可能回答这些任务,工人在众包平台上活跃的时间以及工人完成任务的平均时间)进行高质量的任务推荐.

5.4 数据安全与隐私保护

众包的数据安全和隐私保护目前还没有得到很好的解决,例如众包平台上有很多任务会泄露个人隐私^[96].比如有的任务要求工人录入名片的信息,这显然泄露了名片主人的隐私信息.空间任务容易泄露工人的位置信息^[39].另外一个例子,脑力库虽然提供了保密任务的功能(如表2所示),但是任务保密期最高上限是30天,因此任务内容同样会面临被泄露的危险.这些安全隐患限制了众包的可用性,导致一部分人因为数据安全的考虑而拒绝使用众包来完成任務.而在为工人和任务请求人提供了有效的数据安全和隐私保护以后,众包将吸引更多的人来参与发布任务、完成任务,进而推动众包的广泛应用,因此众包的数据安全和隐私保护也是未来研究的趋势之一.其研究重点就是在保证数据安全和隐私的前提下,如何进行任务的发布以及答案的收集与整合?此外,如何将传统的数据隐私保护技术,如基于数据加密的技术等^[97],应用到众包平台?针对复杂任务,在保证分解后的微观任务具有独立隐私的前提下,如何确保整合答案后的整体数据安全?如何防止众包平台利用收集到的任务请求人和工人

信息推测个人隐私?这些都是值得进一步研究的问题.

5.5 与社交网络结合

目前社交网络得到了用户的广泛关注,社交网络主要为用户提供发布信息、相互交流以及传播信息的服务^[98],例如facebook、twitter、微博、校内等等.由于社交网络包含了大量的在线用户,而有些众包任务就是需要用户及时返回任务答案,比如:实时图片搜索应用^[84],因此如何利用活跃在社交网络上的用户来完成众包任务是未来值得研究的课题.众包任务与社交网络结合尚有一系列问题需要解决:如何通过社交用户的发布信息、关注信息以及登陆情况为用户推荐恰当的任务?为社交用户推荐任务的同时,如何减少对用户使用社交网络基本功能的影响?如何调动社交用户参与完成众包任务的积极性?如何为参与完成任务的社交用户提供奖励?总的来说,如何通过整合社交用户、搜索引擎、数据库等技术来完成指定任务是未来研究的方向之一.

6 结束语

本文主要介绍了众包的基本概念和工作流程;从时间维度,讨论了众包在计算机领域中的相关工作进展;由于众包环境下海量任务处理、任务搜索和推荐、数据安全与隐私保护、与社交网络的结合等方面研究成果不多,大量问题尚待解决,因此我们给出了未来研究的趋势.总的来说,大量的众包问题尚待发现,大量的众包应用还尚待开发,众包在未来10~20年将是一个热点的研究方向.

参 考 文 献

- [1] Howe J. The rise of crowdsourcing. *Wired Magazine*, 2006, 14(6): 1-4
- [2] Howe J. *Crowdsourcing*. New York: Crown Publishing Group, 2008
- [3] Zhao Yu-Xiang, Zhu Qing-Hua. Evaluation on crowdsourcing research: Current status and future direction. *Information Systems Frontiers*, 2012, 11(1): 1-18
- [4] von Ahn L, Maurer B, Abraham D, Blum M. reCAPTCHA: Human-based character recognition via web security measures. *Science*, 2008, 321(5895): 1465-1468
- [5] Ipeirotis P G. Analyzing the amazon mechanical turk marketplace. *ACM Crossroads*, 2010, 17(2): 16-21
- [6] Doan A, Franklin M J, Kossmann D, Kraska T. Crowdsourcing applications and platforms: A data management perspective. *Proceedings of the VLDB Endowment*, 2011, 4(12): 1508-1509

- [7] Alonso O, Lease M. Crowdsourcing for information retrieval: Principles, methods, and applications//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 1299-1300
- [8] Lease M, Alonso O. Crowdsourcing for search evaluation and social-algorithmic search//Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. Portland, USA, 2012: 1180
- [9] Ipeirotis P G, Paritosh P K. Managing crowdsourced human computation: A tutorial//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 287-288
- [10] Alonso O, Lease M. Crowdsourcing 101: Putting the WSDM of crowds to work for you//Proceedings of the 4th International Conference on Web Search and Web Data Mining. Hong Kong, China, 2011: 1-2
- [11] Yuen M, King I, Leung K. A survey of crowdsourcing systems //Proceedings of the IEEE 3rd International Conference on Social Computing. Boston, USA, 2011: 766-773
- [12] Kittur A, Nickerson J V, Bernstein M S, et al. The future of crowd work//Proceedings of the 2013 ACM Conference on Computer Supported Cooperative Work. San Antonio, USA, 2013: 1301-1318
- [13] Doan A, Ramakrishnan R, Halevy A Y. Crowdsourcing systems on the world-wide web. Communications of the ACM, 2011, 54(4): 86-96
- [14] Estellés-Arolas E, González-Ladrón-de-Guevara F. Towards an integrated crowdsourcing definition. Information Science, 2012, 38(2): 189-200
- [15] Koch G, Füller J, Brunswicker S. Online crowdsourcing in the public sector: How to design open government platforms //Proceedings of the 4th International Conference on Online Communities and Social Computing. Orlando, USA, 2011: 203-212
- [16] Guy I, Perer A, Daniel T, et al. Guess who ? Enriching the social graph through a crowdsourcing game//Proceedings of the International Conference on Human Factors in Computing Systems. Vancouver, Canada, 2011: 1373-1382
- [17] Kittur A, Smus B, Khamkar S, Kraut R E. CrowdForge: Crowdsourcing complex work//Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology. Santa Barbara, USA, 2011: 43-52
- [18] Sakamoto Y, Tanaka Y, Yu L, Nickerson J V. The crowdsourcing design space//Proceedings of the 6th International Conference on Directing the Future of Adaptive Systems. Orlando, USA, 2011: 346-355
- [19] Noronha J, Hysen E, Zhang H, Gajos K Z. Platemate: Crowdsourcing nutritional analysis from food photographs//Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology. Santa Barbara, USA, 2011: 1-12
- [20] Kittur A, Smus B, Kraut R E. CrowdForge: Crowdsourcing complex work//Proceedings of the International Conference on Human Factors in Computing Systems. Vancouver, Canada, 2011: 1801-1806
- [21] Heer J, Bostock M. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design//Proceedings of the 28th International Conference on Human Factors in Computing Systems. Atlanta, USA, 2010: 203-212
- [22] Parameswaran A G, Garcia-Molina H, Park H, et al. CrowdScreen: Algorithms for filtering data with humans//Proceedings of the ACM SIGMOD International Conference on Management of Data. Scottsdale, USA, 2012: 361-372
- [23] Guo S, Parameswaran A G, Garcia-Molina H. So Who Won ? Dynamic max discovery with the crowd//Proceedings of the ACM SIGMOD International Conference on Management of Data. Scottsdale, USA, 2012: 385-396
- [24] Selke J, Lofi C, Balke W T. Pushing the boundaries of crowd-enabled databases with query-driven schema expansion. Proceedings of the VLDB Endowment, 2012, 5(6): 538-549
- [25] Marcus A, Wu E, Madden S, Miller R C. Crowdsourced databases: Query processing with people//Proceedings of the 5th Biennial Conference on Innovative Data Systems Research. Asilomar, USA, 2011: 211-214
- [26] Parameswaran A G, Polyzotis N. Answering queries using humans, algorithms and databases//Proceedings of the 5th Biennial Conference on Innovative Data Systems Research. Asilomar, USA, 2011: 160-166
- [27] Kanhere S S. Participatory sensing: Crowdsourcing data from mobile smartphones in urban spaces//Proceedings of the 12th IEEE International Conference on Mobile Data Management. Lule, Sweden, 2011: 3-6
- [28] Franklin M J, Kossmann D, Kraska T, et al. CrowdDB: Answering queries with crowdsourcing//Proceedings of the ACM SIGMOD International Conference on Management of Data. Athens, Greece, 2011: 61-72
- [29] Marcus A, Wu E, Karger D, et al. Demonstration of Qurk: A query processor for human operators//Proceedings of the ACM SIGMOD International Conference on Management of Data. Athens, Greece, 2011: 1315-1318
- [30] Marcus A, Wu E, Karger D R, et al. Human-powered sorts and joins. Proceedings of the VLDB Endowment, 2011, 5(1): 13-24
- [31] Wang J, Kraska T, Franklin M J, Feng J. CrowdER: Crowdsourcing entity resolution. Proceedings of the VLDB Endowment, 2012, 5(11): 1483-1494
- [32] Marcus A, Karger D, Madden S, et al. Counting with the crowd. Proceedings of the VLDB Endowment, 2013, 6(2): 109-120
- [33] Lee J, Cho H, Park J W, et al. Hybrid entity clustering using crowds and data. Proceedings of the VLDB Endowment, 2013, 22(5): 711-726
- [34] Wang J, Li G, Kraska T, et al. Leveraging transitive relations for crowdsourced joins//Proceedings of the ACM SIGMOD International Conference on Management of Data. New York, USA, 2013: 229-240

- [35] Park H, Garcia-Molina H, Pang R, et al. Deco: A system for declarative crowdsourcing. *Proceedings of the VLDB Endowment*, 2012, 5(12): 1990-1993
- [36] Trushkowsky B, Kraska T, Franklin M J, Sarkar P. Crowdsourced enumeration queries//*Proceedings of the 29th IEEE International Conference on Data Engineering*. Brisbane, Australia, 2013; 673-684
- [37] Lofi C, Maarry K E, Balke W T. Skyline queries over incomplete data-error models for focused crowdsourcing//*Proceedings of the 32nd International Conference on Conceptual Modeling*. Hong Kong, China, 2013; 298-312
- [38] Kazemi L, Shahabi C. Geocrowd: Enabling query answering with spatial crowdsourcing//*Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. Redondo Beach, USA, 2012; 189-198
- [39] Deng D, Shahabi C, Demiryurek U. Maximizing the number of worker's self-selected tasks in spatial crowdsourcing//*Proceedings of the 21st International Conference on Advances in Geographic Information Systems*. Orlando, USA, 2013; 1-10
- [40] Chen Xi, Bennett P N, Collins-Thompson K, Horvitz E. Pairwise ranking aggregation in a crowdsourced setting//*Proceedings of the 6th ACM International Conference on Web Search and Data Mining*. Rome, Italy, 2013; 193-202
- [41] Parameswaran A, Sarma A D, Garcia-Molina H, et al. Human-assisted graph search: It's okay to ask questions. *Proceedings of the VLDB Endowment*, 2011, 4(5): 267-278
- [42] Venetis P, Garcia-Molina H, Huang K, Polyzotis N. Max algorithms in crowdsourcing environments//*Proceedings of the 21st World Wide Web Conference*. Lyon, France, 2012; 989-998
- [43] Zaidan O, Callison-Burch C. Crowdsourcing translation: Professional quality from non-professionals//*Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, USA, 2011; 1220-1229
- [44] Søgaard A, Martinez H, Elming J, Johannsen A. Using crowdsourcing to get representations based on regular expressions//*Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Washington, USA, 2013; 1476-1480
- [45] Madnani N, Chodorow M, Tetreault J R, Rozovskaya A. They can help: Using crowdsourcing to improve the evaluation of grammatical error detection systems//*Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, USA, 2011; 508-513
- [46] Vertanen K, Kristensson P O. The imagination of crowds: Conversational AAC language modeling using crowdsourcing and large data sources//*Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, UK, 2011; 700-711
- [47] Negri M, Bentivogli L, Mehdad Y, et al. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora//*Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, UK, 2011; 670-679
- [48] Callison-Burch C. Fast, cheap, and creative: Evaluating translation quality using amazons mechanical turk//*Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore, 2009; 286-295
- [49] Hsueh P Y, Melville P, Sindhvani V. Data quality from crowdsourcing: A study of annotation selection criteria//*Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*. Stroudsburg, USA, 2009; 27-35
- [50] Laws F, Scheible C, Schütze H. Active learning with amazon mechanical turk//*Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, UK, 2011; 1546-1556
- [51] Yan Y, Rosales R, Fung G, Dy J G. Active learning from crowds//*Proceedings of the 28th International Conference on Machine Learning*. Bellevue, USA, 2011; 1161-1168
- [52] Dai P, Weld D S. Decision-theoretic control of crowd-sourced workflows//*Proceedings of the 24th AAAI Conference on Artificial Intelligence*. Atlanta, USA, 2010; 1168-1174
- [53] Welinder P, Branson S, Belongie S, Perona P. The multidimensional wisdom of crowds//*Proceedings of the 24th Annual Conference on Neural Information Processing Systems*. Vancouver, Canada, 2010; 2424-2432
- [54] Salek M, Bachrach Y, Key P. Hotspotting—A probabilistic graphical model for image object localization through crowdsourcing//*Proceedings of the 27th International Conference on Artificial Intelligence*. Washington, USA, 2013; 1-7
- [55] Law E, Zhang H. Towards large-scale collaborative planning: Answering high-level search queries using human computation//*Proceedings of the 25th AAAI Conference on Artificial Intelligence*. San Francisco, USA, 2011; 1-6
- [56] Brew A, Greene D, Cunningham P. Using crowdsourcing and active learning to track sentiment in online media//*Proceedings of the 19th European Conference on Artificial Intelligence*. Lisbon, Portugal, 2010; 145-150
- [57] Kazai G. In search of quality in crowdsourcing for search engine evaluation//*Proceedings of the 33rd European Conference on IR Research*. Dublin, Ireland, 2011; 165-176
- [58] Alonso O, Baeza-Yates R. Design and implementation of relevance assessments using crowdsourcing//*Proceedings of the 33rd European Conference on IR Research*. On Aggregating Lab Dublin, Ireland, UK, 2011; 153-164
- [59] Ipeirotis P G, Provost F, Wang J. Quality management on amazon mechanical turk//*Proceedings of the ACM SIGKDD Workshop on Human Computation*. Washington, USA, 2010; 64-67
- [60] Blanco R, Halpin H, Herzig D M, et al. Repeatable and reliable search system evaluation using crowdsourcing//*Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Beijing, China, 2011; 923-932

- [61] Fan J, Lu M, Ooi B C, et al. A hybrid machine-crowdsourcing system for matching web tables//Proceedings of the 30th International Conference on Data Engineering. Chicago, USA, 2014: 976-987
- [62] Zuccon G, Leelanupab T, Whiting S, et al. Crowdsourcing interactions: Using crowdsourcing for evaluating interactive information retrieval systems. *Information Retrieval*, 2013, 16(2): 267-305
- [63] Kondreddi S K, Triantafillou P, Weikum G. Combining information extraction and human computing for crowdsourced knowledge acquisition//Proceedings of the 30th International Conference on Data Engineering. Chicago, USA, 2014: 988-999
- [64] Gao J, Liu X, Ooi B C, et al. An online cost sensitive decision-making method in crowdsourcing systems//Proceedings of the ACM SIGMOD International Conference on Management of Data. New York, USA, 2013: 217-228
- [65] Singla A, Krause A. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms//Proceedings of the 22nd International World Wide Web Conference. Rio de Janeiro, Brazil, 2013: 1167-1178
- [66] Karger D R, Oh S, Shah D. Iterative learning for reliable crowdsourcing systems//Proceedings of the 25th Annual Conference on Neural Information Processing Systems. Granada, Spain, 2011: 1953-1961
- [67] Alonso O, Schenkel R, Theobald M. Crowdsourcing assessments for XML ranked retrieval//Proceedings of the 32nd European Conference on Information Retrieval Research. Milton Keynes, UK, 2010: 602-606
- [68] Ma H, Chandrasekar R, Quirk C, Gupta A. Improving search engines using human computation games//Proceedings of the 18th ACM Conference on Information and Knowledge Management. Hong Kong, China, 2009: 275-284
- [69] Kazai G, Kamps J, Koolen M, Milic-Frayling N. Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 205-214
- [70] Crescenzi V, Merialdo P, Qiu D. A framework for learning web wrappers from the crowd//Proceedings of the 22nd International World Wide Web Conference. Rio de Janeiro, Brazil, 2013: 261-271
- [71] Demartini G, Difallah E D, Cudré-Mauroux P. ZenCrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking//Proceedings of the 21st World Wide Web Conference. Lyon, France, 2012: 469-478
- [72] West R, Leskovec J. Human way finding in information networks//Proceedings of the 21st World Wide Web Conference. Lyon, France, 2012: 619-628
- [73] Wang G, Wilson C, Zhao X, et al. Serf and turf: Crowdturfing for fun and profit//Proceedings of the 21st World Wide Web Conference. Lyon, France, 2012: 679-688
- [74] Lease M, Yilmaz E. Crowdsourcing for information retrieval: Introduction to the special issue. *Information Retrieval*, 2013, 16(2): 91-100
- [75] Ghosh A, McAfee R P. Crowdsourcing with endogenous entry//Proceedings of the 21st World Wide Web Conference. Lyon, France, 2012: 999-1008
- [76] Bozzon A, Brambilla M, Ceri S. Answering search queries with CrowdSearcher//Proceedings of the 21st World Wide Web Conference. Lyon, France, 2012: 1009-1018
- [77] Heymann P, Garcia-Molina H. Turkalytics: Analytics for human computation//Proceedings of the 20th World Wide Web Conference. Hyderabad, India, 2011: 477-486
- [78] Liu X, Lu M, Ooi B, et al. CDAS: A crowdsourcing data analytics system. *Proceedings of the VLDB Endowment*, 2012, 5(10): 1040-1051
- [79] Chawla S, Hartline J D, Sivan B. Optimal crowdsourcing contests//Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms. Kyoto, Japan, 2012: 856-868
- [80] Mason W A, Watts D J. Financial incentives and the "performance of crowds"//Proceedings of the ACM SIGKDD Workshop on Human Computation. Paris, France, 2009: 77-85
- [81] Sorokin A, Forsyth D. Utility data annotation with amazon mechanical//Proceedings of the Computer Vision and Pattern Recognition Workshops. Anchorage, USA, 2008: 1-8
- [82] Eickho C, Vries A P. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*, 2012, 16(2): 121-137
- [83] Wang J, Faridani S, Ipeirotis P G. Estimating the completion time of crowdsourced tasks using survival analysis models//Proceedings of the WSDM Workshop on Crowdsourcing for Search and Data Mining. Hong Kong, China, 2011: 31-34
- [84] Yan T, Kumar V, Ganesan D. CrowdSearch: Exploiting crowds for accurate real-time image search on mobile phones//Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services. San Francisco, USA, 2010: 77-90
- [85] Chilton L B, Horton J J, Miller R C, Azenkot S. Task search in a human computation market//Proceedings of the ACM SIGKDD Workshop on Human Computation. Washington, USA, 2010: 1-9
- [86] Ambati V, Vogel S, Carbonell J. Towards task recommendation in micro-task markets//Proceedings of the 25th AAAI Workshop in Human Computation. San Francisco, USA, 2011: 80-83
- [87] Yuen M C, King I, Leung K S. TaskRec: Probabilistic matrix factorization in task recommendation in crowdsourcing systems//Proceedings of the 19th International Conference on Neural Information Processing. Doha, Qatar, 2012: 516-525
- [88] Ho C J, Vaughan J W. Online task assignment in crowdsourcing markets//Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto, Canada, 2012: 1-10

- [89] Ho C J, Jabbari S, Vaughan J W. Adaptive task assignment for crowdsourced classification//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA, 2013: 534-542
- [90] Zhang Zhi-Qiang, Pang Ju-Sheng, Xie Xiao-Qin, Zhou Yong. Research on crowdsourcing quality control strategies and evaluation algorithm. Chinese Journal of Computers, 2013, 36(8): 1636-1649(in Chinese)
(张志强, 逢居升, 谢晓芹, 周永. 众包质量控制策略及评估算法研究. 计算机学报, 2013, 36(8): 1636-1649)
- [91] Singer Y, Mittal M. Pricing mechanisms for crowdsourcing markets//Proceedings of the 22nd International Conference on World Wide Web. Riode Janeiro, Brazil, 2013: 1157-1166
- [92] Mason W, Suri S. Conducting behavioral research on Amazon's mechanical turk. Behavior Research Methods, 2012, 44(1): 1-23
- [93] Dawid A P, Skene A M. Maximum likelihood estimation of observer error-rates using the EM algorithm. Applied Statistics, 1979, 28(1): 20-28
- [94] Feng J, Li G, Wang H, Feng J. Incremental quality inference in crowdsourcing//Proceedings of the 19th International Conference on Database Systems for Advanced Applications. Bali, Indonesia, 2014: 453-467
- [95] Lease M. On quality control and machine learning in crowdsourcing//Proceedings of the Human Computation Workshops at the 25th AAAI Conference on Artificial Intelligence. San Francisco, USA, 2011: 97-102
- [96] Schmidt L A. Crowdsourcing for human subjects research//Proceedings of the World's First Conference on the Future of Distributed Work. San Francisco, USA, 2010: 1-7
- [97] Zhou Shui-Geng, Li Feng, Tao Yu-Fei, Xiao Xiao-Kui. Privacy preservation in database applications: A survey. Chinese Journal of Computers, 2009, 32(5): 847-861(in Chinese)
(周水庚, 李丰, 陶宇飞, 肖小奎. 面向数据库应用的隐私保护研究综述. 计算机学报, 2009, 32(5): 847-861)
- [98] Boyd D M, Ellison N B. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 2007, 13(1): 210-230



FENG Jian-Hong, born in 1980, Ph.D. candidate. Her research interests include crowdsourcing data management and database.

LI Guo-Liang, born in 1980, Ph.D., associate professor. His research interests include database, information retrieval and crowdsourcing data management.

FENG Jian-Hua, born in 1967, Ph.D., professor, Ph.D. supervisor. His research interests include database and data integration.

Background

With the rapid growth of computer technology, computer has been widely applied in many areas. However, there are many problems that are still rather hard for computers but easy for human, such as image recognition and sentiment analysis. To address this problem, crowdsourcing is proposed to coordinate the crowd (a large group of people on the web) to solve problems that machine or a small number of users cannot do. Crowdsourcing has many real-world applications, such as Wikipedia, reCAPTCHA and Amazon Mechanical Turk. In recent years, crowdsourcing has attracted significant attention from industrial and academic communities. In this paper, we provide a comprehensive survey of existing studies on crowdsourcing. First, the definition of crowdsourcing is presented according to its basic characterizations. Then the

workflows of crowdsourcing are given by its typical applications. The key research works on crowdsourcing is discussed from three aspects: task preparation, task execution and answers integration. Finally, this paper gives research challenges and directions of crowdsourcing in the future, including big data processing, task search, task recommendation, privacy on crowdsourcing, and the combination of social networks and crowdsourcing.

This work is partially supported by the National Natural Science Foundation of China (Nos. 61373024, 61472198), the National Basic Research Program(973 Program) of China (Nos. 2015CB358700, 2011CB302206), the Beijing Higher Education Young Elite Teacher Project (No. YETP0105).