**密级：商业秘密 保密期限：三年**



**硕士学位论文**



**题目： 基于参与式感知的数据处理及激励机**

**制的研究与实现**

**学 号： 2013111243**

**姓 名： 王东升**

**专 业： 通信与信息系统**

**导 师： 龚向阳**

**学 院： 网络技术研究院**

**2015年 1月 2日**

独创性（或创新性）声明

本人声明所呈交的论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京邮电大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名： 日期：

关于论文使用授权的说明

学位论文作者完全了解北京邮电大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属北京邮电大学。学校有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。（保密的学位论文在解密后遵守此规定）

保密论文注释：本学位论文属于保密在 三 年解密后适用本授权书。

本人签名： 日期：

导师签名： 日期：

基于参与式感知的数据处理及激励机制的研究与实现

# 摘 要

近年来，在信息检索领域，随着传统检索模式的成熟，智能化、个性化检索等越来越受重视，逐渐成为了学术界及以工业界共同的发展目标。同时，云计算、大数据相关技术飞速发展，也为信息检索的进一步演化提供了坚实的基础。正是在这种背景之下，知识库技术在信息检索领域获得了越来越多的关注，已经成为了学术界与工业界共同关注的研究和应用热点。

文提供了一个可以借鉴的知识库构建及应用实例，为实际应用提供了参考。

关键词：垂直搜索 知识库系统 实体 数据抓取 信息抽取 数据集成

RESEARCH AND DESIGN OF KEY TECHNOLOGY

OF THE DOMAIN-SPECIFIC KNOWLEDGE BASE SYSTEM

FOR VERTICAL SEARCHING

# ABSTRACT

In recent years, with the maturity of the traditional retrieval mode, intelligent

It is very natural and reasonable thing that using knowledge base make the vertical search better, because they are both focused on related entities in a particula.

KEY WORDS: vertical search, knowledge base system，entity, data crawl, information extraction, data integration

# 目录

[摘 要 I](#_Toc437364268)

[ABSTRACT II](#_Toc437364269)

[目录 IV](#_Toc437364270)

[第一章 绪论 1](#_Toc437364271)

[1.1 研究背景与意义 1](#_Toc437364272)

[1.2 研究内容与目标 2](#_Toc437364273)

[1.3 论文结构 3](#_Toc437364274)

[第二章 参与式感知中的激励机制 5](#_Toc437364275)

[2.1 垂直搜索引擎概述 5](#_Toc437364280)

[2.1.1 原理和架构 5](#_Toc437364286)

[2.1.2 垂直搜索引擎评价指标 6](#_Toc437364287)

[2.2 自然语言处理技术概述 7](#_Toc437364289)

[2.2.1 分词及词库 7](#_Toc437364292)

[2.2.2 词性标注 8](#_Toc437364295)

[2.2.3 实体提取 8](#_Toc437364296)

[2.2.4 实体特征及关系发现 8](#_Toc437364297)

[2.3 知识库系统概述 9](#_Toc437364298)

[2.3.1 知识库系统概述 9](#_Toc437364300)

[2.3.2 现有的一些知识库介绍 10](#_Toc437364301)

[2.4 本文用到的开源软件和相关技术 11](#_Toc437364302)

[2.4.1 Stanford NLP 12](#_Toc437364305)

[2.4.2 Scrapy 12](#_Toc437364306)

[2.4.3 IKanalyzer分词器 13](#_Toc437364307)

[2.4.4 Lucene 14](#_Toc437364308)

[第三章 实验平台的需求分析 15](#_Toc437364309)

[3.1 主题知识库系统设计概述 15](#_Toc437364313)

[3.1.1 需求分析 15](#_Toc437364314)

[3.1.2 系统总体设计 16](#_Toc437364315)

[3.2 数据层组件设计 17](#_Toc437364316)

[3.2.1 数据层组件的设计 17](#_Toc437364317)

[3.2.2 数据层组件设计关键点 18](#_Toc437364318)

[3.3 平台层组件设计 20](#_Toc437364319)

[3.3.1 平台层组件的设计 20](#_Toc437364321)

[3.3.2 平台层组件设计中的关键点 21](#_Toc437364322)

[3.4 应用层组件设计 21](#_Toc437364323)

[3.5 监控及报表系统设计 22](#_Toc437364324)

[第四章 相关技术、设计与实现 24](#_Toc437364325)

[4.1 数据获取技术 24](#_Toc437364327)

[4.1.1 数据获取实现中的难点 24](#_Toc437364328)

[4.1.2 数据获取技术的实现 24](#_Toc437364329)

[4.2 实体关联 30](#_Toc437364330)

[4.2.1 实体关联要解决的问题 30](#_Toc437364331)

[4.2.2 实体关联的实现 31](#_Toc437364332)

[4.3 实体特征发现 34](#_Toc437364333)

[4.3.1 实体特征发现的难点 34](#_Toc437364334)

[4.3.2 实体特征发现模块的实现 34](#_Toc437364335)

[第五章 激励机制仿真实验对比 38](#_Toc437364336)

[5.1 主题知识库与垂直搜索的集成 38](#_Toc437364342)

[5.2 主题知识库在推荐中的应用 39](#_Toc437364343)

[5.2.1 推荐系统设计 39](#_Toc437364344)

[5.2.2 数据验证与测试分析 40](#_Toc437364345)

[5.3 主题知识库在检索排序中的应用 43](#_Toc437364346)

[5.3.1 应用方案设计 43](#_Toc437364347)

[5.3.2 数据验证与测试分析 43](#_Toc437364348)

[第六章 总结与展望 46](#_Toc437364349)

[6.1 工作总结 46](#_Toc437364351)

[6.2 工作展望 46](#_Toc437364352)

[参考文献 48](#_Toc437364353)

[致谢 50](#_Toc437364354)

[攻读学位期间发表的学术论文和科研情况 51](#_Toc437364355)

# 第一章 绪论

## 研究背景与意义

什么是参与式感知，来龙去脉。

近年来，乘着摩尔定律的浪潮，手机性能在飞速的提高，处理能力更强、嵌入的传感器更丰富、存储空间更大、网络传输速率更快。今天的手机已经从仅仅能打电话的功能机阶段进化到有丰富计算、感知和通信能力的智能设备时期。根据ITU的数据【1】，截止到2015年末，全球移动蜂窝用户会达到70亿，其中有20亿智能手机用户，并且据调查，83%的互联网用户更喜欢用他们的移动设备上网。随着手机技术和通信技术的进步，手机已进入到大众生活的各个角落，在此条件下，一种新型的完成大规模感知任务的方案——“参与式感知”应运而生【2】【3】。参与式感知的关键思路是使普通大众有能力对周边环境用手机进行感知，并且分享收集到的感知数据。

什么是参与式感知？

优点、缺点

虽然手机不是特别制造来用作感知的，但事实上手机可以很好地作为复杂的传感器来工作。相机可以作为视频或图片传感器，麦克风是2倍人声频率的音频传感器，嵌入的GPS接收器可以提供位置信息，其他的嵌入传感器如陀螺仪、加速计和接近传感器可以综合利用来感知和估计手机携带者的环境信息，比如手机携带者是在走路还是骑行等等。另外，通过蓝牙或有线连接，外置传感器可以容易的和手机连接，例如空气质量检测仪等，所以手机作为传感器集成平台有灵活的可扩展性。

典型的参与式感知应用工作在一种中心型的方式，即志愿者通过手机传感器采集的数据，通过无线数据通信上传到中央服务器进行处理。手机端的感知任务可以手动触发、自动触发（更多的称作机会感知）或根据环境上下文触发；在服务器端，数据被分析、处理成可用的形式，通过数据可视化技术在个人的手机上或web界面中展示出来。如下图所示

参与式感知对比于传统的传感器网络有四点优势，后者通常需要部署大量固定的无线传感器设备，尤其是在都市区域。第一，因为参与式感知利用现有的感知（手机中的传感器）和通信（蜂窝或WiFi）基础设施，部署开销几乎为零；第二，手机用户固有的移动性使得参与式感知能提供空前的时空覆盖范围，而传统固定传感器网络所感知的范围是固定不变的；第三，手机应用的开发工具和发布平台都很成熟，使得参与式感知应用的开发和部署变得简单；最后，将普通大众吸引到参与式感知活动中，可以宣传环保等概念，这将巨大地改变人们的生活方式。

参与式感知应用场景。

【2】最早提出参与式感知的概念，构想将参与式感知应用到公众健康、城市规划、社会现象记录和自然资源管理等方面。【4】中将参与式感知应用分为两类：个人中心型和环境中心型。个人中心型中，DietSense用手机记录饮食信息，参与者可以将信息分享给医生或营养专家；PEIR (Personal Environmental Impact Report)使得用户通过手机了解到自己所处的环境的污染程度；BikeNet提供了监测参与者骑行体验的系统；PetrolWatch通过参与者对车外的加油站服务牌进行拍照，上传到中央处理器后处理分析得到汽油价格，人们可以检索感兴趣区域的虽便宜的汽油价格。环境中心型中，Haze Watch利用外接传感器测量、、 、 的浓度，与气象站对比，手机测量精确度可能要差，但是参与式感知的方式可以提供更大的覆盖范围，参与者的移动性也有机会观测到突发的污染情况，这是固定观测站无法提供的服务，所以可以用参与式感知的方式，获取细粒度覆盖范围的观测样本作为高保真固定观测站数据的补充；类似的，EarPhone通过手机中的麦克风采集声音，绘制城市级别的噪音地图来研究噪音和相关社会行为的关系；Nericell利用嵌入的加速计传感器、麦克风和定位系统（GPS、GSM基站辅助定位）综合判断路况条件，比如坑洼、碰撞、刹车和鸣笛（可以进一步判断交通拥堵等信息）可以通过手机来采集并上传。概括起来，参与式感知应用的普遍目标是尽可能精确地观察、发现一些现象、过程或状态，然后分发给感兴趣的人们，取之于民、用之于民。

基于参与式感知的应用如雨后春笋般出现，展示出巨大的活力和潜力，但是这些应用真正的成功部署到人们的日常生活中还需要解决一系列问题。很多参与式感知应用的成功依赖于大量用户的参与和贡献足够数量和质量的数据，如何设计激励机制吸引用户参与是亟待解决的问题。由于系统无法控制参与者的行为，平台收到的数据在时间和空间上具有随机性，会造成数据集的不完整。而很多应用需要推断用户所处环境上下文和活动状态信息，这需要综合各种传感器的数据结合机器学习技术还识别人类活动模式。应用采集了用户的各种周边环境数据，很多都会揭露用户隐私，例如位置信息、轨迹数据和拍摄的图片、视频等，如何保护用户隐私是消除用户疑虑、保证其积极参与的重要工作。同时，参与式感知应用还要有能力鉴别数据的真实性、准确性，否则良莠不齐甚至充斥伪数据的数据集将毫无价值。最后但也是很重要的一点，参与式感知应用要考虑减少手机的资源开销，参与者对于手机电量和流量的消耗比较敏感。

## 研究内容与目标

正如上节所提到的，参与式感知在展示巨大潜力、美好前景的同时，距离真正发挥作用还有很多基础工作要研究。本文主要关注如何设计激励机制吸引用户参与到各种感知活动中来，并且利用这种机制促进用户上传优质数据，维持整个社区健康发展。

为什么需要激励机制

所谓激励机制[5]，原本是经济学中常见的原理，是指在组织系统中，激励主体系统运用多种激励手段并使激励规范化和相对固定化，从而与激励客体相互作用、相互制约的结构、方式、关系及演变规律的总和。虽然参与式感知活动主要靠参与者或者称作志愿者的主动参与，但是感知过程不可避免的会产生经济开销（手机的电量消耗、数据流量传输的费用等）和心理开销（定时采集数据需要人工干预、操作流程复杂枯燥等），如果没有适当的激励机制补偿参与者的经济及心理开销，参与者的参与热情必然会不断衰减，最终导致参与式感知应用的失败。

现有激励机制的主要方向介绍

激励机制可以利用实际的报酬方式或者虚拟的积分方式。利用报酬方式时，付给参与者的报酬必须足够多以来抵消参与者的感知开销并且激励其持续参与到感知活动中，同时也要足够少使得服务提供商可以在预算范围内维持感知活动的运行。参与者的开销是用户的私人信息，参与同一次感知活动不同参与者的开销不尽相同，而且参与者也有充分的动机高报开销来获取更高的报酬。所以参与式感知应用的激励机制要设计得使感知服务提供者经济上可行，同时激励参与者真实报价并保证一定的服务质量。

本文主要研究了参与式感知中通用的激励机制设计原则和已有的方案比较，并且根据论文依托的实验平台具体设计一种激励机制，仿真对比激励机制效果。同时还负责参与式感知实验平台的数据处理部分的设计与实现，包括需求分析、架构设计和技术实现，主要完成了平台对感知数据的接收、图片的分布式存储和高并发访问，以及平台开放的规范的REST接口的设计与实现。

## 论文结构

本文共分为6章：

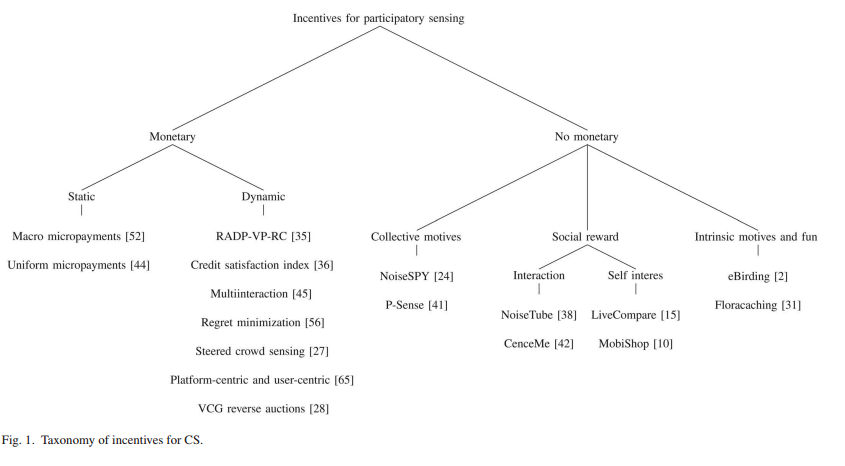
绪论介绍研究背景和意义、阐述研究内容和目标；第二章详细介绍激励机制研究现状，包括其要解决的问题和解决方案；第三章介绍实验平台，需求分析，设计激励（任务）方案；第四章是详细设计和实现；第五章仿真激励机制的效果。

第六章对本文的工作进行了总结，并对以后的工作进行了分析展望。

# 第二章 参与式感知中的激励机制

简要介绍激励机制的研究路线图鉴于激励机制对于参与式感知的重要作用，已有文献已对参与式感知中的激励机制做了广泛的研究，但是参与式感知应用的丰富多样也使得激励机制的设计多是具体适应于特定的应用场景，这一章将从分类学的角度梳理已有文献对激励机制的研究，并且总结激励机制设计通常需要考虑的特性，指出未来的研究方向，并在下文中为本论文依托的实验平台设计具体适配的激励机制奠定基础。

如图总结了参与式感知中激励机制的分类和相应代表应用










## 平台定价激励机制

平台定价激励机制是指参与者获得的报酬由平台端决定，参与者处于被动接受的地位，没有价格协商过程。为方便起见，下文中平台、服务器和服务提供者均指参与式感知应用中处于中心位置的感知任务组织者，系统中还存在参与者、服务订阅者两种角色。可以采用固定价格的方式，平台发布任务时在任务说明信息中指明完成任务参与者将获得的报酬，参与者根据报酬价格和心理预期价格决定是否参与感知任务，完成任务后平台兑现所允诺的价格；也可以采用动态价格方式，参与者成功完成感知任务后，平台根据用户贡献（数据的及时性、稀缺程度、准确性等指标）和可用预算，给予不同用户预期贡献成比例的报酬。很明显，动态价格是更高效、合理的激励方案，但固定价格方式也有其优点，明确的报酬对吸引更多用户参与有明显的的刺激效果。

对于服务器定价的激励机制，【6】中Reddy等提出两种激励方法：

1. MACRO: 提供给参与者固定金额的激励，参与者完成接下来的一段时期的感知任务（一次付清）；
2. MICRO: 提供给参与者固定金额的激励，每次感知数据的上传激励一次，每次的金额也可能不同，根据用户的参与程度决定激励金额（按次付清）。

[7]中Tridib Mukherjee等提出，用户报告城市中发生的事件后，获得奖励由固定的金额乘以一个加权系数来决定，加权系数代表这个报告的及时性、品质等因素。

[8]中Yang D 等提出用户中心的激励机制模型，此模型在参与者之间分配预算，根据他们的参与程度不同每个用户获得总预算的不同比重。

服务器定价的优势是，省略平台和参与者价格协商过程，更高效，但面临的问题是没有考虑不同用户感知开销的不同，对于维持用户持续参与保障的不够。



## 基于逆向竞拍的激励机制

Juong-Sik Lee 首次在[9]中将逆向竞拍引入到参与式感知中。在逆向竞拍中，平台选出报价最低的N个用户作为选中者对其上传的行为给予奖励（普通竞拍（正向竞拍）中会选取报价最高的用户作为竞拍成功者）。在此模式下，平台将感知开销的定价负担转移给参与者，而且此模式可以适应市场环境的动态变化。从次开始，大量文献提出给予逆向竞拍的模型和方法，来从不同方面完善和提升其性能，如引入“VPC”来避免“开销爆炸”、[8]中构建用户中心模型以牺牲掉部分平台利润来获取报价的真实性、[10]提出在线的基于逆向竞拍的激励机制来实时地选择用户。[10]为了实时地选择用户，将会拒绝一些用户来进行采样，所以公平性减弱。[11]也是基于逆向竞拍，但是考虑了预算受限和位置因素等限制条件。



## 其他的重要分类中的代表方法

在线的激励机制

所用的算法总结

。

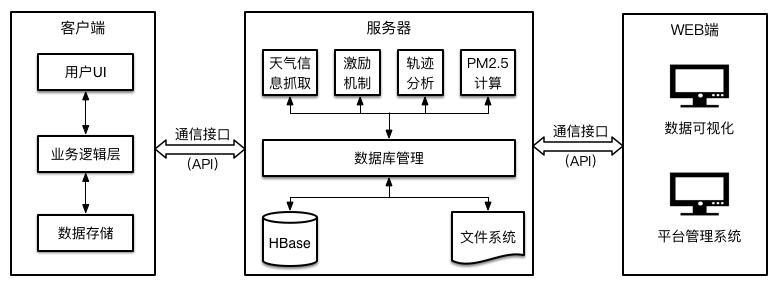
## 现有的总结和未来研究方向

本文的工作中，参考应用了许多优秀的开源软件，在使用的它们的时候，或者直接使用其某个功能，或者对其进行深度的定制来满足自己的需求。本节简要的对本文用到的一些开源软件和工具进行介绍。



# 第三章 实验平台的需求分析

平台介绍。

1. 

## 平台介绍

## 需求

主要负责感知数据的接收与存储、设计并实现适合于本平台的激励机制。

REST API的需求

iOS端的推送

负载均衡的需求

分布式图片服务器的需求

激励机制的需求

# 第四章 相关技术、设计与实现

主题知识库的设计实现中，涉及到较多的技术，例如如何获取相关数据、实体识别、实体特征发现、实体间关系的挖掘、面向垂直搜索的应用等。本章主要介绍实现第三章设计的主题知识库中涉及的一些关键技术，包括数据获取技术、实体关联、实体特征发现三部分的详细设计实现。



## 数据接收等api的设计

由第三章中的需求分析，客户端需要上传文本文件、二进制文件到平台，当检索资源时也需要下载文本文件和图片等二进制文件到客户端，根据网络分层模型，可以利用可靠地面向连接的TCP协议实现客户端和平台的通信。但是通过socket链接，客户端和服务器都需要做很多连接、通信过程中的琐碎工作，而现有的应用层协议，比如HTTP、FTP都可以完成我们的需求，并且简化工作量，客户端开发中iOS、Android平台也提供了丰富的SDK支持，所以我们不需要基于TCP定制应用层协议来开发。

采用http协议进行数据通信（还有其他选择吗？），http协议是啥？

Hypertext Transfer Protocal是无状态的应用层协议，应用于分布式的、协作的、超文本信息系统中。HTTP建立在可靠的传输层或会话层连接之上，通过交换消息来通信。HTTP客户端与HTTP服务器建立连接，发送一个或更多的HTTP请求。HTTP服务程序接受连接，通过发送HTTP响应来服务HTTP请求。

客户端通过请求消息给服务器发送HTTP请求，请求消息由请求行开始，包括请求方法、URI和协议版本，接下来是头部，头部由一个空行来结束，最后是消息体。

服务器通过一次或多次HTTP响应消息来回应客户端的请求。每个响应消息也包括三部分，开始行是状态行包括协议版本、成功或失败状态码，之后是用于说明状态码的文本，接着是响应头部（包括服务器信息、资源元数据和展示元数据），空行代表头部结束，之后是响应消息体。

Client request:

GET /hello.txt HTTP/1.1

User-Agent: curl/7.16.3 libcurl/7.16.3 OpenSSL/0.9.7l zlib/1.2.3

Host: www.example.com

Accept-Language: en, mi

Server response:

HTTP/1.1 200 OK

Date: Mon, 27 Jul 2009 12:28:53 GMT

Server: Apache

Last-Modified: Wed, 22 Jul 2009 19:15:56 GMT

ETag: "34aa387-d-1568eb00"

Accept-Ranges: bytes

Content-Length: 51

Vary: Accept-Encoding

Content-Type: text/plain

Hello World! My payload includes a trailing CRLF.

采集的相关传感器信息，可以构造成JSON格式的文本字符串发送给服务器，而二进制的文件（例如照片、音频等）可以通过二进制比特流放在消息体中来传输。

考虑到采集数据中有用户的位置等隐私信息，处于安全性考虑，我们采用HTTPS协议来通信。

如何通信，客户端和服务器需要做什么，具体采用什么开源框架、技术路线。有哪些设计标准，给出具体的设计方案

很多移动应用擦用HTTP协议和web服务器通信，因为HTTP简单、方便且被广泛支持。然后介绍HTTP协议

然后介绍HTTPS为啥更安全

同样的FTP服务器和socket的通信，最后为啥选择http

http服务器

REST介绍和接口确定

推送

## 激励的设计

### 4.2.1固定价格的任务

### 4.2.2逆向竞拍的机制

实时竞拍

封闭竞拍

### 4.2.3专利中的激励方案

## **负载均衡、高并发**

利用NGIX做负载均衡的方案

如何实现高并发支持

## 4.4 分布式的文件存储

有冗余，可用性增强；图片访问的优化

# 第五章 激励机制仿真实验对比

因为条件不允许，只能仿真对比激励机制的效果

给出三种方案的对比图



## 仿真场景

介绍实验方案，参数，伪代码

## 结果对比

# 第六章 总结与展望



## 工作总结

的必要性。

## 工作展望

# 参考文献

1. Statistics I T U. The world in 2014: ICT facts and figures[J]. 2015.
2. Burke J A, Estrin D, Hansen M, et al. Participatory sensing[J]. Center for Embedded Network Sensing, 2006.
3. Campbell A T, Eisenman S B, Lane N D, et al. People-centric urban sensing[C]//Proceedings of the 2nd annual international workshop on Wireless internet. ACM, 2006: 18.
4. Kanhere S S. Participatory sensing: Crowdsourcing data from mobile smartphones in urban spaces[C]//Mobile Data Management (MDM), 2011 12th IEEE International Conference on. IEEE, 2011, 2: 3-6.
5. 王慧贤. 社交网络媒体平台用户参与激励机制研究[D]. 北京邮电大学, 2013.
6. Reddy S, Estrin D, Hansen M, et al. Examining micro-payments for participatory sensing data collections[C]//Proceedings of the 12th ACM international conference on Ubiquitous computing. ACM, 2010: 33-36.
7. Mukherjee T, Chander D, Mondal A, et al. CityZen: A cost-effective city management system with incentive-driven resident engagement[C]//Mobile Data Management (MDM), 2014 IEEE 15th International Conference on. IEEE, 2014, 1: 289-296.
8. Yang D, Xue G, Fang X, et al. Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing[C]//Proceedings of the 18th annual international conference on Mobile computing and networking. ACM, 2012: 173-184.
9. Lee J S, Hoh B. Sell your experiences: a market mechanism based incentive for participatory sensing[C]//Pervasive Computing and Communications (PerCom), 2010 IEEE International Conference on. IEEE, 2010: 60-68.
10. Dong Zhao; Xiang-Yang Li; Huadong Ma, "How to crowdsource tasks truthfully without sacrificing utility: Online incentive mechanisms with budget constraint," Proceedings of INFOCOM'14, IEEE , 2014
11. Jaimes L G, Vergara-Laurens I, Labrador M A. A location-based incentive mechanism for participatory sensing systems with budget constraints[C]//Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on. IEEE, 2012: 103-108.

# 致谢

# 攻读学位期间发表的学术论文和科研情况

论文

宋涛，王洪波. 面向垂直检索的知识图谱系统研究与设计，中国科技论文在线. 2014-11-02. 论文编号201411-16.

科研项目

* + - 1. 横向研究项目：基于云计算的垂直搜索引擎系统开发
      2. 国家 863 项目：云计算中心网络大容量交换机关键技术与系统