**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：

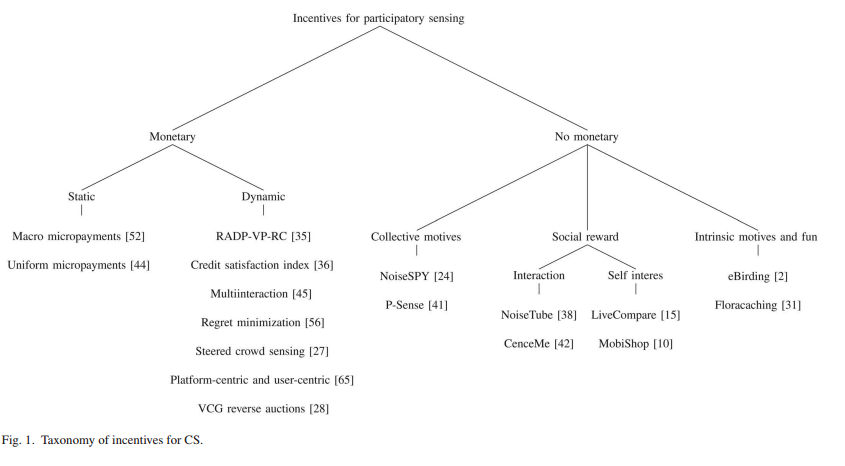
绪论介绍研究背景和意义、阐述研究内容和目标；第二章详细介绍激励机制研究现状，包括其要解决的问题和解决方案；第三章介绍实验平台，需求分析，设计激励（任务）方案；第四章是详细设计和实现；第五章仿真激励机制的效果。

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

。

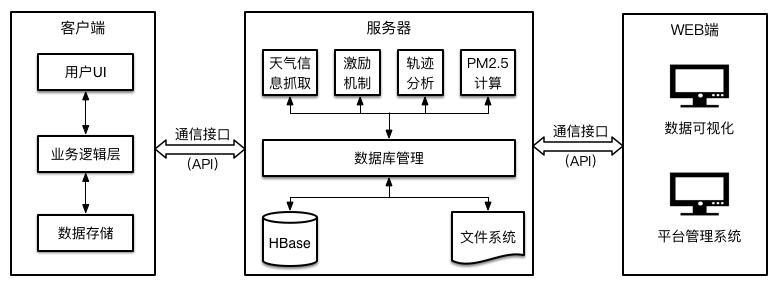
## 现有的总结和未来研究方向

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

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

### 数据获取实现中的难点

数据获取技术是主题知识库系统数据层实现中涉及的关键技术，能否高效、全面的获取到相关的数据，直接关系的知识库的详尽、准确程度。数据抓取一般包括主动抓取和被动获得两种方式，难点通常集中在主动抓取中，主动抓取一般通过网络爬虫来完成任务。主动抓取的难点则集中在如何从不同来源得到数据这个问题上。

不同于早期各种网络爬虫模拟Http协议抓取网页即可，随着 Web2.0 技术的发展，更多的WEB开发者为了提升用户体验，吸引更多的用户，在Web设计实现中使用了越来越多的技术，数据的抓取工作遇到了越来越多的挑战。尤其在近些年，JavaScript、Ajax、jQuery、登陆验证等技术越来越广泛的使用对数据获取造成了极大的影响[22]。网络爬虫通过简单的Http请求已经不能很好的完成抓取任务，越来越多的数据需要模拟用户的访问行为才可以得到。

另一方面，数据抓取的不全面则是由于越来越严重的Deep Web[23]问题。传统的Web页面一般通过页面上超链接方式来相互连接，然而Deep Web是指那些需要通过检索等方式才能访问到的Web页面，而且这些页面一般会针对不同的检索内容、访问者呈现不同的内容，这些页面一般只是其背后一个大的数据仓库的外显，例如Wikipedia、百度百科等等站点。由于网络爬虫一般通过入口地址及超链接对网页进行抓取，Deep Web特殊的方式也对传统网络爬虫数据的抓取造成了很大影响。

### 数据获取技术的实现

为了解决主题知识库实现过程中数据获取的问题，本节设计实现了一个用于主题知识库数据获取的主题网络爬虫。正如前文介绍，爬虫的抓取目标是视频领域的各种信息，包括优酷、爱奇艺等视频网站，以及百度百科、豆瓣等等站点之上关于视频的各种信息。针对4.1.1中分析的难点，该爬虫解决各种前段技术带来的数据抓取问题，包括动态数据、突破爬取限制，以及Deep Web数据的获取问题。



图4-1 数据获取模块的实现

爬虫在Linux平台上实现，采用MySQL[24]、Redis[25]、Django[26]、Scrapy、Webkit[27]等技术共同实现。在Scrapy的框架之下，根据需求定制了各种特有组件。

如图4-1所示，系统的组成及各部分的作用如下：

* 1. Core engine是整个数据获取系统的核心，完成抓取调度的策略、数据处理控制、模板控制等多种核心功能，是系统的核心，控制了数据抓取的整个流程。
  2. 在2.4.2节中给出过Scrapy的介绍，在Scrapy轻量级、可定制的特点之上，我们集成了Webkit浏览器内核插件，使用Scrapy + Webkit的方式来完成抓取下载。Webkit重要作为Scrapy下载器的插件，当遇到动态数据时取代默认的下载器。
  3. Redis是一款开源的优秀内存数据库，支持key-value等多种形式的数据存储结构，我们采用Redis作为链接库（即抓取队列）的存储。这部分作为链接库的存储模块，包含所有的链接信息，是抓取系统的核心。
  4. MySQL数据库作为网页信息的持久化存储。采用MyBatis配合完成分表分库的存储方案。用来存储抓取到的网页原始结果，以及针对页面抽取到的特定结果。利用MySQL的Dump功能可以完成周期性的页面备份。
  5. Django则用来实现爬虫管理面板。主要功能包括配置及更新抓取模板，监控系统的任务、资源状态等。在这里使用其Admin功能，主要是用来管理一个配置数据库，控制抓取爬虫的行为。

图4-1种用红色虚线圈出的部分就是我们针对4.1.1中爬虫所需解决问题设计的模块，这些模块配合来解决动态数据、爬取限制以及Deep Web问题，下文分别介绍几个问题的具体解决。

4.1.2.1 动态数据的抓取

在项目实现的过程中，由于我们需要对大量的视频网站上的数据进行抓取，例如优酷、爱奇艺、腾讯视频等等。而这些网站为了提升用户体验，用到了较多的Javascript、Ajax等很多高级技术，相当比例的数据都是动态数据，例如用户评论、相关视频推荐链接等等。所以，简单的Http请求已经无法完整的抓取所需的数据。此外，很多站点会针对不同的用户展示很多个性化的信息，一些额外的信息只有在登录之后才能抓取得到。而模拟用户交互也是动态数据采集的一大难点。

目前比较常见的动态数据采集方式有两种。一种是针对网站动态数据请求逻辑进行分析，并且用代码模拟整个访问逻辑，然后采集访问过程中的数据。这种方式执行效率较高，但开发成本较高，需要花费大量的人力分析网站的访问逻辑，且有时需要大量试验才能完整的重现访问过程。另一种方式是采用集成Webkit等浏览器内核来模拟用户访问网站时的交互事件，并采集数据。集成浏览器内核的方式效率相对较低，但适用性比较广、且开发成本较低。

由于视频领域存在大量的站点需要抓取，针对不同站点分析访问逻辑的方法费时多、成本大、扩展性差，所以我们选取了集成Webkit内核的方式来完成抓取任务。

如图4-1中红色虚线圈出的动态数据下载模块，我们对Scrapy的源码进行修改，集成Webkit创建一个新的下载器中间件。相关代码如下：

**from** scrapy.http **import** Request, FormRequest, HtmlResponse

**import** gtk

**import** webkit

**import** jswebkit

**import** settings

**class** WebkitDownloader( object ):

**def** process\_request( self, request, spider ):

**if** spider.name **in** settings.WEBKIT\_DOWNLOADER:

**if**( type(request) **is** **not** FormRequest ):

webview = webkit.WebView()

webview.connect( 'load-finished', **lambda** v,f: gtk.main\_quit() )

webview.load\_uri( request.url )

gtk.main()

js = jswebkit.JSContext( webview.get\_main\_frame().get\_global\_context()) renderedBody = str( js.EvaluateScript( 'document.body.innerHTML' ) ) **return** HtmlResponse( request.url, body=renderedBody )

然后，在scrapy的配置项settings.py中加入如下代码，这样在整个系统启动时候，实现了Webkit下载器插件的注册，当出现动态数据时，调用下载器完成任务：

*#which spider should use WEBKIT*

WEBKIT\_DOWNLOADER=['ccb']

DOWNLOADER\_MIDDLEWARES = {

'rate\_crawler.dowloader.WebkitDownloader': 543,

}

**import** os

os.environ["DISPLAY"] = ":0"

至此，基于Scrapy的数据获取系统已经具备了动态数据的采集功能。此外，为了提供效率，动态数据的识别是需要解决的另一个问题。数据获取系统可以根据程序指定的抓取策略及抓取脚本区分动态数据和普通数据。当遇到动态数据时，调用编写好的模拟点击脚本，来完成数据的采集。

动态数据的识别筛选模块流程如图4-2所示。链接库以（URL、抓取器）的格式来存储待抓取的数据。抓取从访问链接库开始，经过一下几个流程：

1. 访问链接库，得到下一个待抓取的链接，即拿到下一条（URL、抓取器）数据。调用每个URL的对应的抓取器集合中对应的抓取器，完成URL抓取。
2. 抓取到的页面持久化保存，存入页面库。以供数据处理器使用。
3. 普通链接的提取，依据统一的正则来完成，相关规则存在配置库中。提取Http、ftp等常见的超链接。
4. 动态链接的提取，主要是js、ajax请求。依据配置库中的正则规则、请求/相应规则，自动提取动态的数据访问链接。
5. 模拟点击。依据脚本库中的配置，模拟所有符合规则的点击按钮，将请求转化为链接。
6. 链接过滤，过滤无用的广告、推广，即与抓取不想关的链接以及已经抓取过的重复链接。将其余有用链接更新加入链接库。



图4-2 动态数据筛选模块流程

4.1.2.2 Deep Web数据的抓取

构建视频主题知识库中需要大量的资料，这些资料不单来自各种视频网站，而且还包括互联网上海量的各种信息，例如百度百科中影视的介绍、豆瓣影评、用户博客评论等等。然而我们不可能对百度百科、豆瓣影评所有的页面都抓取一遍，然后再筛选出视频相关的页面，这种方式工作量太大，而且这些站点一般每有全部网页的链接地图，很多页面只能通过检索才能到达。所以采用传统的超链接遍历方式，是无法拿到这些网页的，这就是Deep Web问题。

为了解决Deep Web问题，如图4-1所示，我们在数据获取系统中设计了Deep Web处理模块，这个模块的主要功能就是利用视频领域词条库中的词条，结合Core engine给出的站点链接生成规则来直接生成网页URL，然后将URL更新至链接库等待抓取。



图4-3 Deep Web模块处理流程

图4-3是实现这一模块的流程图。大致流程如下，从视频领域词库中取出未处理的词条，然后取出站点检索链接生成规则，根据规则生成检索结果页URL，取出前N个较相关的检索结果。遍历这N个检索结果，直至结束。对每个检索结果判断其是否是视频领域相关的页面，是则存储链接库，否则丢弃。

4.1.2.3 数据获取系统类图

如图4-4，是基于Scrapy实现的本文所述的数据获取系统的类图。LinkExtractor是可以完成提取动态数据链接、模拟点击行为、生成DeepWeb链接等工作，分别由JSLinkExtractor、AjaxLinkExtractor、JSLinkExtractor、DeepWebLinkExtractor、NormalLinkExtractor等多种实现来完成；DownLoader可以完成各种类型数据的抓取，由HttpDownloader、HttpsDownloader、FtpDownloader、JsDownloader等类实现完成；Config完成配置的管理工作，包括RuleConfig、SpiderConfig等多种实现类用于规则及爬虫管理配置等等。

图4-4 数据获取系统接设计

## 激励的设计

### 实体关联要解决的问题

实体关联在主题知识库数据层处理子层和平台层的构建中都必不可少。本节设计实现的数据关联，主要解决两个问题，即实体与实体的关联和实体与资料的关联。

由于主题知识库的是以实体为和核心的组织结构，所以有统一的实体层次十分重要，现实中很多不同名字的东西其实都同指代一个实体，实体与实体的关联就是要解决这一问题，将实体按概念划分关联，而不是靠名字来区分。此外，数据采集系统一般会从多个来源得到很多资料，如何找出这些资料中一样的实体，也是实体与实体的关联问题。解决好实体的问题，是主题知识库知识体系建立的基础。

另一方面，将资料中的实体关联整理好，整个知识库知识体系由了点以后，如何将与这些点相关的资料各自找出，则是另一个需要解决的问题，即实体与资料的关联。实体的相关特征等知识信息，都需要从这些资料中得到，资料关联的准确性、全面性，直接关系到知识的准确性和丰富程度。

### 实体关联的实现

针对4.2.1种描述的问题，本节设计实现了一个用于实体关联的系统。类似于4.1节中数据获取技术的实现，实体关联系统的实现同样是面向视频领域。该系统致力于解决视频领域中实体与实体的关联和实体与资料的关联。

实体的关联不管在学术界还是产业界都是一直被研究和关注的问题。已经有分块法、排序法、索引法等一些比较常见的处理方法。基本的分块法（Basic Blocking）[28]在寻找相似记录对时，基本思路是将数据集分为多个彼此不相交的小块，通过两两比较分块内的记录来降低总体的比较次数。分块以Blocking Key作为依据，Key可以是记录的一个或多个属性值，这个值一般比较有代表性且在不同记录中区别较大，比如人名中的姓、年纪或邮政编码等。

排序法的基本思想则是将记录首先根据某一域值排序，排序法认为在排序后的序列中相似的记录会出现在临近的位置。排序法通常依据用户选定的若干个属性字符串作为键进行排序，之后采用固定区间大小的滑动窗口来找出相似的记录，在同一区间中的记录两两比较，来确定相似性。这种方法的缺点也比较明显：由于字符和单词位置对字符串排序的影响十分明显，所以相似的重复记录排在邻近的位置这一假设前提并不总是成立；另外当某一些值出现的次数很多时，即拥有相同或相似域值的记录数量超出了给定的窗口大小，会造成相似记录无法被比较，导致随后的聚类操作很难能将所有相似的重复记录完全识别出来，会造成结果的遗漏。

索引法[29]是Chaudhuri S提出一种在线的数据整合方法，是一种由代表性的数据清洗机制。索引法的处理需要有干净参照表的支持，其基本思路是基于干净的参照表数据生成一个ETI（Error Tolerance Index）索引，然后依靠这个索引来找到每一个在线输入的数据最匹配的干净记录，之后将二者的数据进行整合，从而完成对输入数据的整合和在线清洗。

针对传统几种方法的优缺点，本文提出了基于Lucene索引的分块关联法。该方法综合了分块法和索引法，分为两步关联。第一步粗粒度的筛选，利用基于Lucene的索引法找到迅速相似的实体或者资料，第二部在相似实体或者资料中用分块法确定真正关联的实体或者资料。

为了表述方便，我们将实体的资料统称为记录。当记录是实体时，它是一部视频及这个视频的描述信息，一般是我们利用4.1节中介绍的数据获取系统从视频网站上抓取的视频信息，这些视频信息一般有比较好的结构，通常包括名称、出产年份、演员、导演等比较完整的信息，可以理解为一堆词的集合；当记录是资料时，则一般是我们利用数据获取系统从百度百科、豆瓣影评及视频网站的用户评论区抓取到的文献，这些资料一般是一段一段的文字，可以理解为一堆句子的集合。在我们的设计中，将以句子形式组织的资料分词并处理后转变为与实体一样形式的记录，然后统一处理。



图4-5 数据关联模块的实现

如图4-5所示，关联需要经过转换器、索引、比较分类、审核、存入标准库几个阶段，各阶段的功能如下：

转换器。转换器是用于处理资料形式的记录的工具，目的是将以句子形式组织的资料转化为与实体形式相同的组织形式，以便于关联系统统一处理。该部分主要完成的功能是对资料进行切分，分为标题、概要、内容、评论几部分，然后对每部分的句子进行分词处理，形成以词构成的多个集合。这样，资料便转变成了与实体类似的形式。实体形式的记录本身便是以词组成的多个集合，这些集合可能包括视频名称集合、演员列表集合等等。经过转化器的处理，我们数据中每条记录都是一个集合的集合，其逻辑格式如下：

记录R：

{ 集合A：{词1、词2……}

集合B：{词1、词2……}

……

集合G：{词1、词2……}

}

Lucene索引器。关于Lucene本文研究在2.4.4中给出了基本介绍，这里不再赘述。如图4-3，在任务开始时，Lucene索引为空，每检测一条记录R，用该记录的描述信息，即多个词的集合关键域值搜索，然后将搜索结果中相关度最高的K条记录筛选出来，作为关联备选集，交由分类比较器进行准确分析。

分类比较。分类比较器在索引筛选后的小规模数据集上执行，在两个记录的描述集合之上用文本编辑距离的平均值来评判是否是一对关联记录，当高于成功阈值时匹配成功，成功匹配的记录统一计入标准信息库，而在成功阈值和失败阈值之间可能成功的记录会由人工审核确定，匹配值低于失败阈值的直接将这条记录加入索引里，这样确保了标准库的质量。

审核和存入标准库。可能匹配的记录对会交由人工审核。标准库的更新采用信息融合的方式，当信息冲突时候全部保留，确保数据的完整性，但美中不足的是可能会给标准库中带入错误信息。



图4-6 数据关联模块类图

如图4-6，是本文设计实现的数据关联模块的UML类图。LuceneIndexer封装了Lucene的索引、查询接口，提供统一的索引维护、查询接口；IndexLinker利用LuceneIndexer完成基于Lucene筛选的第一步关联；BlockMatcher用于计算两个个对应块之间的相似度；而BlockLinker利用BlockMatcher完成两个实体之间基于分块的匹配度计算；Linker中实现了两步的关联算法，DataLinker负责关联任务的调度、执行、结果统计。

采用本文设计实现的Lucene筛选加分块方式，能在效率和准确性上有较好的平衡。在筛选阶段，使用Lucene索引查找相似记录对有效地减少了记录比较的次数，时间复杂度接近于线性，同时也保证了容错率和召回率，因为是以多个域为依据查找记录，少数几个域相似程度低能通过其他相似程度高的域来弥补。而在分类比较，即分块阶段，则在校规模数据上进行准备计算，而且还加入了人工审核来弥补，对最终关联结果的准备性给出了保证。

## 技术选择（实现）

主题知识库中的实体在垂直检索中直观重要，大部分垂直检索的结果都以实体为单位来组织排列。然后，如何检索到用户真正需要的实体，同时提供关于实体更多的信息才是真正区别于传统搜索的地方。实体的特征在这些方面起到了至关重要的作用，实体特征的多少决定了一个展现给用户每个结果的详细程度，同时特征的多少还关系到实体间关系的准确性和多样性。所以实体特征发现是平台层中非常重要的一个组件。本节就描述我们在面向视频垂直检索的主题知识库中实体特征发现组件的实现。

### 实体特征发现的难点

利用本章前文介绍的数据获取技术和实体关联技术，我们已经可以将互联网上的资料划分组织成以实体及实体相关资料为单位的一个大集合。实体特征发现就是要从每一个每个实体相关资料中找到这个实体的代表性特征。

常见的特征发现一般以实体标签的方式体现，实现方法则包括人工提取及自动提取。人工提取的一般由专家根据资料或者自己的专业知识为相应实体打上数量较为有限的标签，类似于由作者选择的学术论文关键词的生成过程。这种方式生成的标签一般较为准确，但不够全面，而且成本极高，不适合大规模应用。自动提取方式则一般无需人工干预或者只需要很少的人工干预，通常综合利用规则、机器学习等多种方式进行自动化的标签提取。这种方式计算效率较高，但很多情况下准确率较差，资料数量的多少对结果的准确程度影响很大。实体特征发现的难点就是人工和自动两种方式的矛盾，如何能够协调好效率、准确性、适用范围，是能不能完成实体特征发现的关键。

### 实体特征发现模块的实现

本文采用的实现方案属于自动提取方式，基本思想是对实体相关文档中所有的词条进行重要性排序，重要性是指该词条是否能够体现改实体的特征。本文的实现方案用到了基于IKanalyzer的分词、TFIDF[30]思想、TextRank[31]、Word2Vec[32]等多种技术。

本文在2.4.3节中已经对IKanalyzer进行了简要的介绍，这里不再赘述其基本原理。如图4-7，本文实现的实体特征发现模块中分词功能采用了基于IKanalyzer的实现，实现中保留了IKanalyzer的核心分词算法，但为了满足需求，对IKanalyzer的词库管理进行了扩展修改，并且增加了词来源功能。在分词过程中，我们对字典的管理进行了类别区分，利用抓取时得到的结构化信息分为了视频名称、演员名、导演名称等。在子分词器进行分词时候，不光给出分词结果，还附带该次所属的类别。之后统一交由决策器进行歧义处理、合并结果，返回最终的结果即由词、词对应来源为单位组成。



图4-7 基于IKanalyzer的分词实现

TFIDF是一种已经被广泛应用的传统的词项加权方法，核心目标在于评价一个词对于一个文件集或一个语料库中一份文档的重要程度，即该词能否代表该文档。一般字词的重要程度随着它在文档中出现的次数成正比增加，但同时会随着它在文档集合或者语料库中总的出现的频率成反比下降。

TextRank算法是Mihalcea于2004年提出的一种无监督的标签提取方法，其基本思想类似于PageRank，首先根据词之间的共现关系构建一个词项图，认为词与词之问共现代表一种推荐关系，与重要词共现的词也重要，在这个图上经过迭代可以得到词项的重要度排序[33]。

Word2Vec是Google在2013年开源的一款将词表征为实数值向量高效工具，采用了CBOW（Continuous Bag-Of-Words，即连续的词袋模型）以及Skip-Gram两种模型。Word2vec是一个基于深度学习的训练模型，通过训练，可以把对文本内容的处理转变为在K维向量空间中的运算，通过计算向量空间上的向量的相似度来表示文本的语义特征。因此，Word2vec计算输出的空间词向量被用于多种自然语言处理相关的任务，比如聚类、找同义词性分析等。此外，Word2Vec的处理效率也非常高，Mikolov在他的论文中指出一个优化的单机版本一天可训练上千亿词。

如图4-8所示，是本文实现的实体特征发现系统，包括分词、分类、多维度特征提取、评分整合、类别重分配几个过程，最终生成实体的特征集合。各部分作用如下：

1. 分词。采用基于IKanalyzer定制的中文分词器，词典由定向从各大视频站点抓取的结构化数据生成以及通用词典生成。分词将实体相关的资料文档分为词的集合，以便于下一步处理。
2. 类别信息生成。采用Word2Vec工具对所有的文档集进行词向量分类，在实际工作中我们将此划分了大约800个类别，词结果以（词，词类别）的形式存储与文件中。在系统加载时，以key-value的形式存储于内存中以供查询。
3. 基于TFIDF及规则维度的特征生成。TFIDF维度，计算某一实体相关文档集中所有词的TFIDF权值算出所有词的排名。而规则主要起到过滤作用，基于词性、黑名单的规则过滤掉无用词。
4. 基于TextRank的特征生成。利用前文介绍的TextRank算法，计算与同一实体相关的所有词在文档集中的评分及排序。同样需要在结果中对无用词进行过滤。
5. 评分整合。本过程主要作用是用3）、4）两部分生成的特征集合融合生成特征集，在这一部设定两部分的权重，循环叠加计算，得到所有标签的综合排名及分值。
6. 类别重分配。根据2）种Word2Vec计算得到的类别信息对特征集合进行调整，基本思想是避免同类词过多而降低特征集合表示意义的丰富程度。计算规则是取前N个特征，分类后计算各部分比重，然后按比重排名重新分配各类别标签的数量。



图4-8 实体特征发现模块的实现

图4-9实体特征发现模块类图

图4-9是上文所述的实体特征发现模块的类图。FeatureDetector负责整个特征发现任务的调度、统计；TfIdfDetector采用TF-IDF的方法计算实体特征，TextRankDetector采用TextRank方法计算实体特征，Word2VecUtil封装了word2vec组件，提供词向量的分类及类别查询功能；Detector由将TfIdfDetector和TextRankDetector的计算结果进行合并操作，即merge；然后依据Filters将杂质特征过滤，最后依据Word2VecUtil提供的类别信息对结果进行调整，得到最终结果。在这里，Filters是一系列实现Filter接口的自定义过滤器。

## 测试

# 第五章 激励机制仿真实验对比

在前两章，本文介绍了主题知识库的设计及一些关键技术的实现。主题知识库的成功构建并不是结束，只有将知识库中的知识体系应用于垂直搜索之中，才能真正的体现主题知识库的价值。主题知识库在垂直检索中有着众多的应用场景，本章结合笔者实验室的一个视频垂直搜索项目，介绍了视频主题知识库在视频垂直检索中的应用。具体而言，针对推荐及检索排序两个场景做了介绍。



## 仿真场景

本节介绍视频主题知识库与视频垂直搜索引擎的集成。集成后的系统是本章所有测试的基本环境，所有的测试数据也由这个系统得出。主题知识库与垂直搜索引擎的集成主要通过开放接口的形式完成，即知识库开放知识，搜索系统利用这些知识来完善搜索中的各个环节。

图5-1 集成知识库的视频垂直检索系统

如图5-1所示，集成了主题知识库的视频垂直检索系统相对与常见的检索系统主要有以下几个不同：

1. 检索数据的来源。系统可以检索的数据来源包括两部分，一部分是互联网上海量的文档，与传统搜索引擎没有区别；另一部分是知识库应用层提供的知识体系。此外，资料的获取也依托于知识库系统的数据层来完成。
2. 知识库应用层对垂直检索引擎的支持。知识库应用层提供的接口可以帮助解析器、查询器、数据获取模块等垂直搜索的功能模块改进。对数据的来源、处理及分析提供了多种支持。
3. 扩展业务的支持。针对推荐、个性化检索等模块，知识库的应用层可以提信息补充支持，改进这些与检索相关的功能效果。

## 三种机制的介绍

### 推荐系统设计

图5-2 推荐系统模型

主题知识库对推荐系统设计的改进如图5-2所示。本节通过视频检索中的推荐系统来说明主题知识库的应用效果。视频检索的推荐系统基于协同过滤[34]实现，将实体的特征和关系加入了协同过滤中，构成了用户、视频（即实体）、视频特征（即实体特征）三种节点，用户的点击行为和视频间的关系（即实体关系）则构成了节点间的边。相对原始的协同过滤数据流图多了图书特征一种节点和图书关系一种边，丰富了数据流图。

### 数据验证与测试分析

5.2.2.1 测试数据集

本文选取了视频检索系统推荐模块中人工标注的200个推荐结果作为对比依据。每个人工标注的数据包含10部相关视频推荐。这些测试数据包含电影、电视剧、动漫、综艺四个类别，具体数量及比例如表5-1所示。在每次推荐之中，我们以视频名称为搜索词搜索，并记录结果页中的相关推荐结果，然后依据测试指标计算每个测试数据的结果好坏。另外，在转化率的评测方面，我们采用分析日志，测试数据集即是系统所有用户的点击日志。

表5-1 推荐测试数据

|  |  |  |
| --- | --- | --- |
| 类别 项目 | 数量（部） | 所占比例（%） |
| 电影 | 80 | 40 |
| 电视剧 | 70 | 35 |
| 动漫 | 30 | 15 |
| 综艺 | 20 | 10 |

5.2.2.2 测试指标

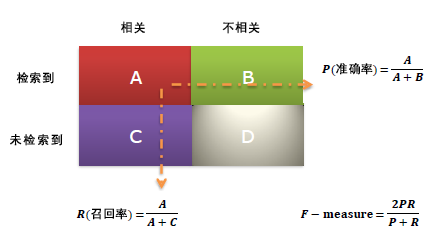


图5-2 准确率、召回率及F1值

测试指标采用通用的召回率、准确率、F1值以及转化率来评价，图5-2表示了召回率、准确率、F1值的计算规则。

召回率：指推荐结果靠前的列表中好的推荐占整个好的推荐的比例。召回率通常反映系统的全面性，即能否找到所有合理的结果。

准确率：指在位于推荐结果靠前的列表中好的推荐所占的比例。准确率通常反应系统的准确性，即能否将合理的结果排在合理的位置，即靠前的位置。

F1值：通常准确率和召回率是相互影响的，理想情况下是做到二者的均衡，即双赢。但是一般情况下，准确率高了召回率就相应的较低，而准确率低了召回率就相对较高。F1值是综合了召回率和准确率的一个评价指标，一般可以认为F1值越高，测试的结果越好。

转化率：转化率指用户点击推荐结果的比率，其评测较为特殊，我们采用系统上线后一段时间内的日志来进行评测，统计整个推荐系统中用户点击转化率的变化规律。转化率从用户的角度真实的反映了结果是否如用户期望的一致。

5.2.2.3 测试结果及分析

1. 准确率、召回率、F1值的评测

如图5-3所示，是视频搜索系统中推荐系统在利用知识库优化前后的评测对比，蓝色代表召回率、绿色代表准确率、红色代表F1值，其中较浅的颜色是未使用主题知识库的推荐，深色是使用了主题知识库之后的推荐。此外，图中包括多个类别独自的测试数据以及综合的数据。

图5-2 系统准确率、召回率及F1值对比

从图中可以看出，使用主题知识库之后，推荐系统在准确率、召回率、F1值方面的表现均好于之前。可以看出在电影及综艺两个类目，数据效果提升最为明显。对于电影，造成这种差异的原因可能是电影相关的网络资料较多，影评、剧情分析之类的数据为电影类目的实体提供了大量的视频信息之外的知识信息，这些信息能更完善的描述电影类实体，完善的信息在推荐时起到了较多的良性影响。而对于综艺节目，推荐效果的提升应该是由于综艺视频包含信息较为杂乱，涵盖内容广泛，单靠视频网站的视频描述信息并不能很好的对综艺节目进行描述，有了知识库的支持信息之后，同样扩展了综艺类实体的描述信息，在推荐时起到了作用。

此外，电视剧及动漫类目也有一定提示，但效果不如其他两个类目明显，应该是一部电视剧或者动漫包含集数较多，视频网站对每一集都已经有一定数量的描述，多集之后信息相对电影或者综艺完善程度较高。所以推荐方面提升不如另外两个类目明显，但主题知识库的引入，仍然带来了推荐效果的提升。

1. 转化率

推荐系统是区别于传统导航或者搜索的另一种浏览方式，目的是简化用户的访问代价。同时，站点通过推荐把用户尽最大可能的留在了自己站内。因此，转化率的测试就是计算运行时期用户在推荐板块的点击量与推荐信息总量的比值，该比值可以在一定程度上反映推荐结果的有效性。我们选取了系统试运行之前的10天到运行3个月后这段时间内的日志进行了分析统计，得到的推荐转化率如图5-3所示。

图5-3 系统上线后转换率曲线

从图5-3我们可以清晰的看到，系统试运行期间，转化率在不断地上升后稳定，说明系统对用户的吸引度以及用户对推荐的信任度在逐渐增强。因为在优化后的系统不止用到了知识库信息，还用到了通过用户的点击行为，所以在系统上线的前一个多月之内，随着不断增多的收集到的用户新的观看行为，系统的转换率不断提升。在将近两个月之后，转换率已经比较稳定。从图中可以看出，最终系统的转换率大概提示不到一倍，在推荐方面，主题知识库的应用效果已经十分明显。

## 结果对比

### 应用方案设计

本节通过视频垂直检索中的检索系统来说明主题知识库的应用效果。视频垂直搜索的检索系统是采用基于Lucene的Solr[35]来实现的。主题知识库通过丰富检索的索引内容和扩展改进检索的评分因子项来改善系统的搜索结果。



图5-5 检索及排序系统流程

主题知识库对检索系统设计的改进如图5-5所示，并未做过多修改，只是在索引阶段将知识库中的知识数据加入了索引。在索引中，知识库数据与视频信息标准库的数据分别存储，在检索时候权重处理有所不同。图中红色箭头代表搜索过程，虚线箭头代表索引过程，知识库数据和标准库数据两部分作为被索引的资料，是用户搜索结果的两大来源。

### 数据验证与测试分析

本文选取了视频垂直搜索系统的300次检索结果进行评测，分别针对结合了主题知识库知识的检索以及未利用主题知识库知识的检索结果进行对比评测。检索测试的300个检索结果来自从系统检索日志提取的真实的用户检索关键词记录。

类似针对检索和排序的评测，我们从准确性、多样性以及搜索的点击转化率三个维度进行评测。准确性只需满足排序前两个的结果中有正确结果即可；多样性是指检索结果中数据的区分度，结果中区别越大则多样性越好。其他标准与推荐系统测试中含义一样，可以参看5.1.2节，此处不再赘述。评测结果由人工判定，取多个人工结果的平均值作为参照标准。此外，由于视频垂直检索系统准确、丰富的特性，每个结果页显示结果较少但信息较多，我们只取前8个结果作为检索结果集来评价检索效果。

图5-6 检索及排序系统优化前后对比

如图5-6所示，未集成知识库知识的检索和集成之后的检索对比，在准确性方面并没有提示，反而有微小的下降，而多样性的评测结果则显著提升。这是由于主题知识库中带来了很多视频信息标准库中没有的资料，当用户检索时显然便可以检索到更多的内容。同时，随着检索内容的增多，准确性方面有所下降，但是由于垂直检索的特殊性，即只要实体关键信息完整，检索很难出错，所以准确性的下降并不十分明显，在接受范围之内。然后多样性效果的提示显然会给用户带来更好的体验。

如图5-7所示，是从检索系统更新前二十天到系统更新60天这段时间内的系统检索日志中分析计算，得到的用户搜索结果点击转换率的数据。可以看出，用户的搜索转换率随着知识库数据的加入，迅速提升到了97%以上。可见，主题知识库对于垂直领域的信息丰富程度提升有极大的帮助。

此外，还应注意的一点是，与推荐系统中应用效果缓慢提示然后稳定的情况不同，主题知识库在检索系统中的应用能够迅速的对系统产生影响，这是由于推荐系统中除主题知识库数据外还用到了用户点击数据，系统需要积累较多的用户点击数据才能有比较稳定的推荐结果，所以结果又一定的延迟性。而检索直接利用知识库信息，无需用户反馈，效果可以迅速体现。

图5-7检索系统优化前后用户点击率

# 第六章 总结与展望



## 工作总结

本文主要对面向垂直搜索引擎的主题知识库中的关键技术进行了全面、深入的研究，在研究基础上给出了一个主题知识库设计的例子。同时，以视频垂直检索为背景，给出了主题知识库构建中几个关键技术的实现，以及主题知识库在视频垂直检索中的几个应用场景。

随着互联网各领域垂直化服务的广泛开展，主题知识库的应用越来越广泛，关于知识库的研究受到了越来越多人的关注。本文从主题知识库构建的相关技术研究开始，研究介绍了主题知识库的设计、关键技术的实现以及主题知识库在垂直检索中的应用，对主题知识库进行了全面、深入的研究介绍。本文主要工作及成果如下：

1. 本文较为全面的介绍了知识库领域的相关信息，分析了当前主题知识库研究的进展及重要性。并针对主题知识库设计实现中涉及到的相关技术进行了研究分析，包括垂直检索、自然语言处理、知识库系统等等。

2. 本文给出了一个详细的主题知识库的设计范例，并针对设计，给出了主题知识库构建中较为核心的几项技术的实现和分析，是知识库设计实现工作的一个很好参考。知识库的设计采用了分层结构的适用于垂直搜索的原型。设计原型将知识库分为数据层、平台层、应用层以及监控及报表系统四大模块，详细分析了各模块需要完成的任务及设计中的要点难点。在几个关键技术的实现中，针对数据获取方面，主要解决了动态数据、爬取限制、Deep Web等问题；在实体关联方面，给出了实体与实体关联和实体与资料关联的统一解决方案；在实体特征发现发面，提出了基于Lucene索引加分块法的二步关联法。这些技术都是实现主题知识库中必须解决的问题，直接关系到主题知识库的准确性、全面性等问题。

3. 本文给出了一个主题知识库在垂直搜索中的应用场景及效果分析，论证了主题知识库对垂直搜索引擎功能上的积极作用。在垂直搜索系统中的推荐及检索排序两个方面进行了知识库应用的研究分析，得出了很好的结论，验证了本文设计的主题知识库系统的有效性。这也从侧面印证了主题知识库系统研究的必要性。

## 工作展望

随着近些年移动互联网热潮的兴起，各种各样的垂直化服务越来越多的出现在人们的眼中。围绕“吃喝玩乐”，专注于一个领域的互联网服务越来越智能化、个性化。在这种背景之下，主题知识库的研究和应用越来越广泛，越来越受到学术界和产业界的重视。特别是云计算、大数据技术的蓬勃发展，更是使主题知识库的构建和应用成为了可能。本文设计的主题知识库在多个方面还可以进行扩展和深入优化：

1. 围绕实体的知识体系构建。随着互联网上信息爆发式的增长，越来越多的信息可以纳入主题知识库的范围。在Deep Web中，还有大量的信息需要去挖掘获取。
2. 在知识库知识体系迭代中，如何基于各种用户的反馈数据对知识库体系进行改进也是一个有待深入研究的问题。
3. 主题知识库的构建是一个复杂的、代价高的任务，如何利用云计算、大数据技术对主题知识库的构建提供有效帮助，也是十分值得研究的一个问题。

# 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.
12. http://www.google.com/intl/zh-CN/insidesearch/features/search/knowledge.html
13. http://nlp.stanford.edu/software/
14. http://scrapy.org/
15. https://code.google.com/p/ik-analyzer/
16. http://lucene.apache.org/
17. http://hbase.apache.org/
18. http://baike.baidu.com/view/3061630.htm
19. http://hadoop.apache.org
20. https://storm.apache.org
21. G Pant and F Menczer. Myspiders: Evolve your own intelligent Web crawlers. Autonomous Agents and Multi-Agent Systems, 5(2):221–229, 2002.
22. Yeye He, Dong Xin, Venkatesh Ganti, et al. 2013. Crawling deep web entity pages. In Proceedings of the sixth ACM international conference on Web search and data mining (WSDM '13). ACM, New York, NY, USA, 355-364.
23. https://www.mysql.com/
24. http://redis.io/
25. https://www.djangoproject.com/
26. https://www.webkit.org/
27. Jaro M A. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida [J]. Journal of the American Statistical Association, 1989, 84(406): 414-420.
28. Chaudhuri S, Ganjam K, Ganti V, et al. Robust and efficient fuzzy match for online data cleaning[C]. Proceedings of the 2003 ACM SIGMOD international conference on Management of data. ACM, 2003: 313-324.
29. http://en.wikipedia.org/wiki/Tf–idf
30. Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]. Association for Computational Linguistics, 2004.
31. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]，Advances in Neural Information Processing Systems. 2013: 3111-3119.
32. 李鹏，王斌，石志伟. Tag-TextRank：一种基于 Tag 的网页关键词抽取方法[J]. 计算机研究与发展，2012，49（11）：2344-2351
33. 项亮．推荐系统实践[M]．北京：人民邮电出版社，2012年.
34. http://lucene.apache.org/solr/

# 致谢

时光飞逝，转眼间研究生的生活就要结束了。两年半的研究生生活对我的人生具有重要的意义。在网络技术研究院宽带中心这两年多的时间里，我的学术水平和实践能力都得到了很大的提高，同时我也结识了很多非常优秀的同学，学习到了很多珍贵的东西。最重要的是这两年的生活学习中受到的导师的谆谆教诲，这对于我来说都是巨大的收获，我将深切感激。

首先，我要感谢导师程时端教授，是您给予了我一个难得的学习和深造的机会。您渊博的知识和严谨的科研精神是我耳濡目染，使我发自内心的感到敬佩。您的学术造诣、谦虚的精神将成为我人生的标杆，不断激励我进步。

感谢项目组王洪波副教授，您在日常的学习工作中给予了我极大的帮助和指导，带领我完成多个科研项目，在论文的撰写中给我提了很多宝贵建议，在就业选择上给我提出了很多真诚可贵的建议。您对工作的认真的态度是我以后工作的榜样。

感谢项目组成员在本文工作中给予的帮助，同时感谢所有在科研工作中曾给予我无私帮助和关怀的老师和同学。

感谢宽带网研究中心全体老师、同学和工作人员。我相信在大家的不断努力下，宽带网中心的发展会越来越好，能够结识大家我感到非常的荣幸，祝愿大家工作顺利，前程似锦。

最后，感谢各位专家、教授、评委老师在百忙中审阅本文，也希望能得到各位专家老师的宝贵意见和建议。

# 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