合肥工業大學
HEFEI UNIVERSITY OF TECHNOLOGY

大学生创新创业训练计划项目

结题材料

项 目 级 别： ☐国家级 ☑省级 ☐校级

项 目 编 号： S201910359345

项 目 名 称： 基于 Spark 的人工智能知识图谱构建

项目负责人： 文 华

项目组成员： 刘宏鑫、周 余

起 止 年 月： 2019 年 5 月至 2020 年 4 月

指 导 教 师： 罗月童

所 在 学 院： 计算机与信息系

创新创业教育中心

目 录

项目结题申请表 3
项目研究总结报告 4
1. 项目背景 4
2. 项目预期 4
3. 项目方案 5
4. 项目进度安排 7
5. 项目实践情况 7
6. 项目难点 8
7. 项目成果 9
8. 项目展望 10
个人总结报告—文华 11
个人总结报告—刘宏鑫 14
个人总结报告—周余 17
支撑材料 19

大学生创新创业训练计划项目结题申请表

<table>
<tr><td>项目级别</td><td colspan="3">☐国家级 ☑省级 ☐校级（请打“√”）</td><td>项目编号</td><td>S20191035934</td></tr>
<tr><td>项目名称</td><td colspan="5">基于 Spark 平台的人工智能</td></tr>
<tr><td>负责人</td><td>文华</td><td>学院</td><td>计算机与信息系</td><td>联系电话</td><td>18856302551</td></tr>
<tr><td>指导教师</td><td>罗月童</td><td>学院</td><td>计算机与信息学</td><td>职　称</td><td>教授</td></tr>
<tr><td>起止时间</td><td colspan="3">2019 年 5 月至 2020 年 4 月</td><td>资助经费</td><td>8000 元</td></tr>
<tr><td colspan="6">参加学生信息(包括负责人)</td></tr>
<tr><td>姓名</td><td>学号</td><td>学院</td><td>专业班级</td><td>项目分工</td><td>获取学分</td></tr>
<tr><td>文华</td><td>2017218007</td><td>计算机与信息系</td><td>物联网工程 17-2</td><td>数据处理</td><td>3</td></tr>
<tr><td>刘宏鑫</td><td>2017217989</td><td>计算机与信息系</td><td>物联网工程 17-2</td><td>可视化</td><td>3</td></tr>
<tr><td>周余</td><td>2017218005</td><td>计算机与信息系</td><td>物联网工程 17-2</td><td>数据爬取</td><td>3</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="6">随着计算机大数据的快速发展，可以借助于互联网平台的各种工具找到有价值内容，但海量数据給筛选、组织与评价带来极大困难。知识图谱具有强大的语义处理与开放互联能力，可以精确地表达概念及其相互关系所构成地语义网络，更好地为机器所理解；且能够帮助用户快速、准确地检索所需要地信息。本项目基于 Spark 平台构建了人工智能中的机器学习、自然语言处理与机器视觉等三个领域的知识图谱，完成相关知识的重整，取得了较好的实验效果。此外，撰写了一篇基于本项目所构建知识图谱的研究论文。本项目创新点：1）将人工智能的有关内容以知识图谱的形式展示出来；2）对人工智能知识内容实现了传统的思维导图等所达不到的可视化效果。</td></tr>
<tr><td colspan="6">参加竞赛、发表论文、申请专利情况
参加 2019 年 iCAN 国际创新创业大赛并荣获安徽省二等奖。</td></tr>
<tr><td colspan="6">项目验收意见

学院负责人（签字）：　　学　院（盖章）：

年　　月　　日</td></tr>
<tr><td colspan="6">学校意见

负责人（签字）：　　主管部门（盖章）：

年　　月　　日</td></tr>
</table>

基于 Spark 平台的人工智能知识图谱构建

项目研究总结报告

摘 要：大数据为知识图谱带来了新的思路和挑战。借助互联网平台的各种工具找到有价值内容，但海量数据給筛选、组织与评价带来极大困难。本项目利用实体识别、关系抽取、可视化分析等技术构建了人工智能领域的图谱，以期给广大学习者尤其是本科生提供有益的学习参考。

关键字：人工智能，Spark 平台，知识图谱

1. 项目背景

随着 Web 技术飞跃式发展，互联网先后经历了三个时代，它们分别具有不同的特征：文档互联的“Web 1.0”时代，数据互联为特征的“Web 2.0”时代以及当下正在发展的知识互联的崭新“Web 3.0”时代。知识互联为人们的学习与交流提供了极大便利，人类的知识交互达到了历史的新高峰。然而，互联网上的知识来源复杂、良莠不一，零散混乱、体系松散，尤其是在大数据的时代背景下，这給内容的筛选、组织与评价带来了极大挑战。知识图谱（Knowledge Graph）是人工智能（Artificial Intelligence，简称 AI）领域一项重要的技术分支，具有强大的语义处理能力与开放互联能力。值得注意的是，目前国内尚无针对人工智能这一领域的知识图谱工具。人工智能正处于快速发展阶段，了解、学习、掌握有关知识与技术是学生、工程师、科研人员所面临的一大挑战，优秀的知识架构可以帮助学习者达到事半功倍的效果。

目前，已经有许多大型知识图谱被构建出来，如 DBpedia 、Freebase 等，然而，当前的知识图谱工具普遍存在以下问题：1）通用知识图谱工具涉面较广，但知识冗余混乱、组织零散、系统性差，不利于用户的专业学习；2）垂直知识图谱工具种类少，成熟的应用仅限于某些领域，在一些具有较大应用需求的领域未获重视，前景广阔。

综上所述，本项目的目的是构建一个面向学习者尤其是本科生的人工智能领域的垂直知识图谱，意义在于通过 Spark 完成人工智能知识的重整，实现了一个学习者尤其是本科生适用的知识图谱工具。人工智能领域繁多，为消减技术流程的复杂度，我们选取机器学习（Machine Learning，ML）、自然语言处理（Natural Language Processing，NLP）与机器视觉（Machine Vision，MV）等三个领域作为代表。

2. 项目预期

本项目拟利用分布式爬虫并结合 PathFinder 算法或从现成学科数据库获取具体学科的知识内容。再借助于 Spark 框架优秀的并行化处理能力，过滤所获取内容中的无意义数据，并在其上应用知识抽取等相关算法，完成对文本的知识关系抽取。其次，采用垂直搜索引擎工具完成关键词与相关信息的联想。最终，通过数据可视化技术将取得的不同模块的内容，即知识图谱进行可视化展示。知识图谱的一般构建流程如图 2.1 所示。

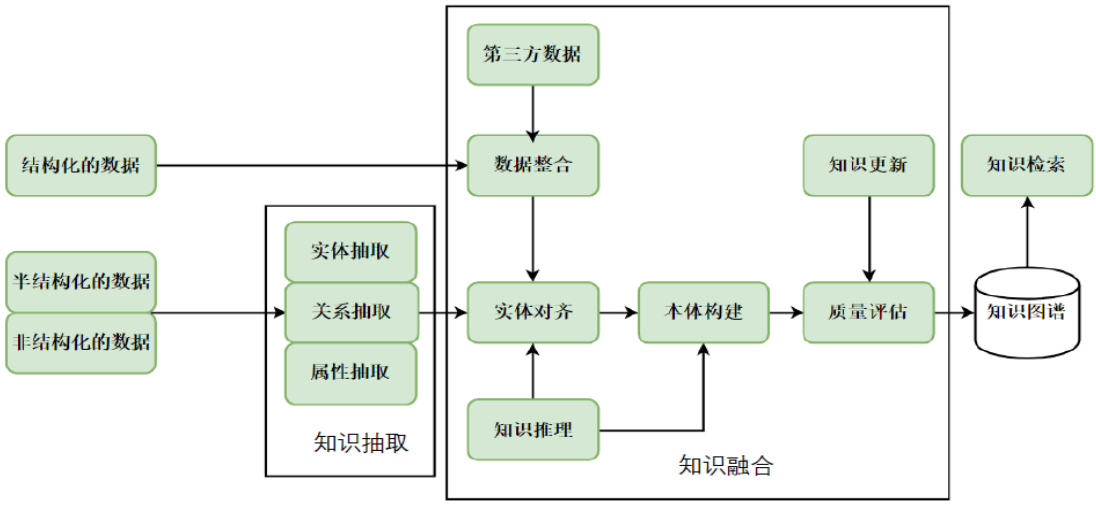


图 2.1 知识图谱的一般构建流程

3. 项目方案

3.1 原定技术方案

知识图谱构建流程大致如下：

①数据获取：借助分布式爬虫实现 PathFinder 算法获取学科知识数据，经 Spark 平台对数据进行初步过滤；

②知识抽取：本项目面临的知识抽取工作主要是术语抽取。

a.实体抽取：基于 VSM 描述文本进行实体抽取，最重要的是选择与分类相关的特征构造实例特征向量。因此，利用 Word 分词组件完成分词，将所得数据各个术语提取出来；

b.关系抽取：将关系抽取看作是一个分类问题，采用 x^2 统计完成文本特征选择并用贝叶斯（Bayes）方法对 a.所得分词，在训练语料的基础上构造分类器，进行文本分类；

c.属性抽取：将判别属性视作分类问题，在文本分类过程中完成此项工作。

③知识融合：

a.实体对齐：本项目将实体对齐定义为用户输入与知识库中的实例匹配；

b.知识推理：为了保持项目拟构建工具的稳定性，本次项目不对知识库中的内容进行知识推理，而借助基于贝叶斯统计推断的研究方法实现关键词联想；

c.本体构建：本体是用于描述一个领域的术语集合，本项目的本体构建暨术语的提取与分类在知识抽取阶段已完成；

d.质量评估：对给定输入所生成的知识图谱与现有知识体系进行对比、评价。

④创建完成：至此，知识图谱的创建工作基本结束。

3.2 实际技术方案

由于原定方案在实践中遇到诸多困难，结合实践我们做出如下修改：

① 原定使用分布式爬虫实现 PathFinder 算法获取学科知识数据，在实验中我们发现 PathFinder 算法对不确定源的网页文档爬取的效果较好，但在指定网址且被爬取网站对内容有原

始排序的情况下，是否使用 PathFinder 算法的效果差距较小。出于爬取效率、经济成本的因素考虑，我们改用主从分布式爬虫，其逻辑结构如图 3.1 所示。

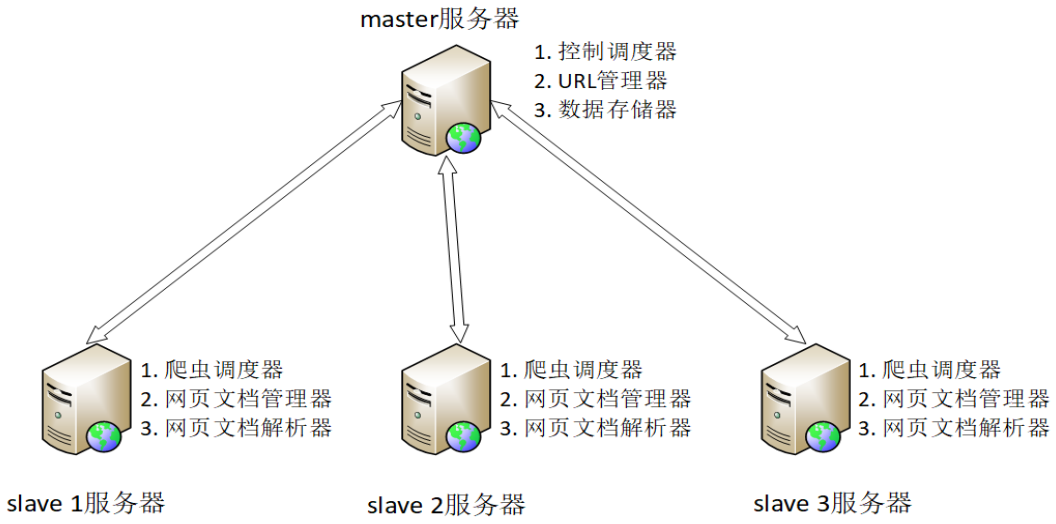


图 3.1 分布式爬虫逻辑结构

② 原方案将关系抽取看作是一个文本分类问题，在实验中我们发现文本分类所得知识关系并不理想，且时空复杂度较高、鲁棒性较差，而专门为关系抽取设计的工具，不论在准确度，还是时空消耗上都交文本分类方法低，因此我们改用 Jiagu 模型作为关系抽取的工具。

③ 原方案拟使用 JavaScript 的 D3 函数库进行知识图谱可视化，在实验中我们发现 amCharts 4 的功能更为强大、操作更为便捷，因此我们改用 amCharts 4 作为知识图谱可视化的工具。图 3.2 所示为使用 amCharts 4 对数据库中的三元组进行可视化的过程。

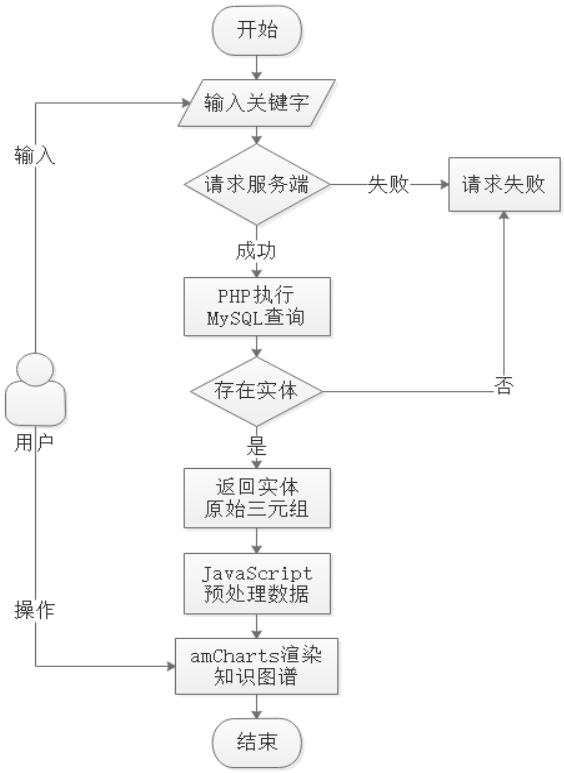


图 3.2 知识图谱可视化请求过程

4. 项目进度安排

如下所示为本项目的原进度安排，实践证明原进度安排是切实可行的，既给与了我们充足的时间完成既定计划，又使我们可以在项目进行到一定程度时使用成果参加相关的比赛。

表 4.1 项目进度安排表

时间	项目安排
2019.4	调研与准备阶段
2019.5-2019.7	收集数据,使用相关算法模型进行数据预处理
2019.8-2019.10	对已处理的数据文本分类，优化模型
2019.11-2020.12	项目平台的初步搭建,测试工具效果
2020.1-2020.2	准备中间检查,对项目的研究过程进行整理
2020.3	结题报告书和相关材料的撰写
2020.4	结题答辩

5. 项目实践情况

5.1 技术攻关

周余：周余主要负责元数据的获取。在项目开始前，我们选定了若干个国内知名度较高的网站作为数据爬取的对象，周余在实践中发现原定网站中大多数的网页存在大量的抄袭、质量不高等问题，而且她敏锐地感知到 PathFinder 算法对本项目所需数据的获取帮助不大。针对有关问题，周余进行了仔细的查证、调研与实验，最终确定了本项目爬取数据的方案。

文华：文华主要负责使用 Spark 平台对所爬取到的文本数据进行存储、清洗、过滤，并挑选文本进行标记，制作训练数据，使用 Jiagu 模型对标记过后的数据进行训练，得到基于本项目的文本数据模型，最终将模型用于文本知识关系，即实体组成的三元组的提取。此外，文华还主要负责论文的写作。

刘宏鑫：刘宏鑫主要负责使用三元组数据进行知识图谱的可视化。可视化设计前后端的一系列难题，刚加入项目组时刘宏鑫还自谦是一个技术新手，对这些问题一时难以下手，但是作为团队一员的责任担当，促使他奋发图强、力争上游。最终，刘宏鑫不仅完成了他负责的任务，还参与到了知识提取的建言献策与论文的写作当中。

5.2 团队协调

必须肯定的是，我们团队成员认真务实，具备较强的实践精神与合作意识。本团队三人都完成了本科阶段基本的知识储备，具备完成项目所需的数学推导、算法设计、编程实现等各方面知识，队员中两人有实验室的学习经历，三人旗鼓相当，队员对于硬件与软件平台的搭建，编码与算法优化游刃有余，团队成员或有过项目经历，或涉猎广泛。针对选题，团队成员已经做了大量

前期工作，譬如查阅了大量选题有关的资料，掌握了项目所需主体平台的搭建与一些算法。可见在项目开始前，我们就有充分的信心与能力完成本项目的既定任务。

我们也承认，在团队合作中也发生过一些摩擦与争论，但我们大多数时候都能顾全大局，为了团队的整体利益而尽量避免争执。原团队有 4 名成员，但其中一名队员在项目进行的过程中（中期检查以前），对参与项目的意义产生了怀疑，最终，因为其认为参与项目无法获取足够的“利益”，自愿选择了退出团队。整个过程全体成员见证共识而且本人在退队书中承认，这名队员的退队是其个人行为。

5.3 经费使用

本项目原申请了经费 8000 元人民币，实际使用了 5730 元，未超过申请的项目经费额度。其中 431 元用于购置内存条，为了提高电脑的性能，使之可以更好地在项目实践中发挥作用；5299 元用于购买显卡，原定将此显卡安装在实验室的主机上，增强其性能并用于本项目的数据训练、可视化实现等操作，但由于疫情而搁浅。

6. 项目难点

6.1 基于网络爬虫的数据获取

本项目目的在于构建人工智能知识的知识图谱，但目前并不存在有关内容的开源数据库或信息源，因此，利用分布式爬虫获取内容是唯一有效的方法。然而，传统的分布式爬虫虽然可以有选择地访问网页与相关链接并获取所需信息，但获取内容仍含有一定的无价值数据。在大数据环境下，分布式架构的分布式爬虫比单机多核的串行爬虫具有更高的效率与更新速度。爬取相关度更高的内容也是一个值得考虑的问题，为了解决这个问题，我们借助主从分布式爬虫，根据网站默认的排序所用权值，并设置阈值以获取内容。

另一个值得关注的问题是数据爬取源，恰当的数据源不仅可以更快速地得到所需内容，而且获取内容更“干净”、更接近直接在工程上应用。本项目拟实现所构建知识图谱的相关信息的联想，对信息热度、就业热度等进行统计分析，为学生的深入学习乃至就业择业提供参考。因此，数据源对最终结果的准确度、完整度至关重要。譬如：构建“编程语言”的知识图谱时，可选择“TIOBE 编程语言排行榜”作为信息热度的数据源；构建“机器学习”的知识图谱时，可选择“CSDN 博客”、“牛客网”、“LeetCode 中文官网”作为行业形势与就业热度的数据源。实验发现：只有从 CSDN 博客与博客园获取的数据质量较高。

6.2 数据处理与信息联想

文本预处理是将文本表示成一组特征项。将每个词作为文本的特征项是目前常用的处理方法，针对本项目的文本特征项主要是专有名词与术语，本项目在 Spark 平台下利用 Word 分词，实现分布式工作。Word 分词是用 Java 实现的，实现了多种分词算法，并利用 ngram 模型消除歧义，能有效对数量词、专有名词与人名进行识别。分词所得到词语组，主要用于信息联想，也就是在构建完成的知识图谱中检索与給定词语有关联的三元组。

如前文所述，我们选择 Jiagu 作为知识抽取的工具。 Jiagu 模型是一个国产的开源自然语言处理工具，以 BiLSTM 等模型为基础，使用大规模语料训练而成。Jiagu 模型提供中文分词、词性标注、命名实体识别、情感分析、知识图谱关系抽取、关键词抽取、文本摘要、新词发现、情感分析、文本聚类等常用自然语言处理功能，API 丰富，且操作便捷、稳定性高。本项目选择 Jiagu 模型作为知识抽取的工具，取得了十分理想的效果。

必须清醒地认识到，知识关系的抽取是一个十分复杂的工作，而且由于汉语作为一门分析语所具有的固有特征，对汉语文本进行知识抽取较之英语更加复杂，个人短时间内独立开发一套工具进行知识抽取很难，因此借助开源工具是很有必要的。目前开源的许多工具都是以英语作为语料和语义环境开发出来的，对汉语的兼容性较差，经过权衡我们最终选择了国产的开源自然语言处理工具 Jiagu。目前 Jiagu 的作者还在维护这个项目，可以预期今后的版本的功能将更加强大。

7. 项目成果

本项目选择“人工智能”作为学科知识图谱构建的出发点，解决用户在特定应用场景下的问题，高效、完整、准确地学习相关知识，同时借此论证本项目的方案在工程上的可行性，而这也是本项目的最终落脚点。经过近一年的努力，我们圆满完成了项目既定的目标，期间也对原项目方案进行了大范围的修改。本项目所取得的成果主要有如下几点：

① 开发了一个人工智能领域（以机器学习、自然语言处理、计算机视觉为例）的知识图谱，该图谱可以对用户输入的关键词进行快速检索、反馈，如图 7.1 与图 7.2 所示。

② 撰写了 1 篇撰写基于本项目的研究论文，对构建知识图谱的步骤与所用方法进行了详细论述；

③ 本项目在 2019 年 iCAN 比赛中荣获安徽省二等奖。

图 7.1 项目搜索演示样例

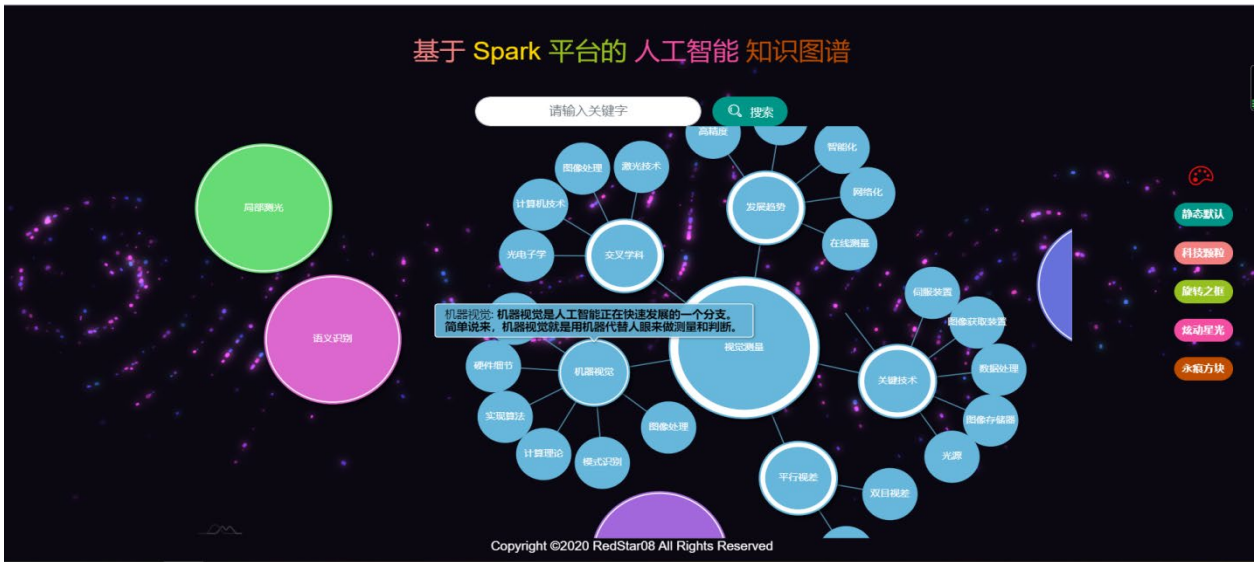


图 7.2 项目主题演示样例

8. 项目展望

本项目成功地构建了人工智能领域的知识图谱，首次将本科计算机类专业的课程内容知识以知识图谱的形式展示出来；可以帮助用户准确、快速地检索人工智能领域相关术语并提供解释，同时给出术语的联想结果，利于用户进一步学习；形象化地展示人工智能领域的脉络、历史沿革与发展趋势，为用户复习、深入学习提供参考。下一步的工作将从几个方面进行研究：采用知识联想等方法增加知识图谱中的知识实体规模，进一步优化知识关系抽取，改善知识融合等。

本项目大胆对目前热门的人工智能领域进行了知识图谱构建，初步探索出了相关图谱的构建步骤，得到了效果较为理想的实验结果。本文的构建方法可以应用于大多数针对特定学科或领域的垂直知识图谱的构建，以期在扩大训练语料的基础上得到较本文实验结果覆盖率更广的领域知识，即规模更为庞大的 RDF。值得一提的是，本文在构建图谱的过程中认识到：汉语作为一门分析语所具备的固有特点是构建汉语知识图谱的障碍之一，在后续工作中或可以考虑以英文语料为基础构建知识图谱，待完成后再行翻译。

本项目还以人工智能领域的机器学习、自然语言处理与机器视觉三个分支为例，介绍了构建相关垂直知识图谱的技术流程。以期能够抛砖引玉，使其他有志之士有所参考。

个人总结报告—文华

团队角色

我是我们团队的负责人，除了完成自己负责的技术任务以外，还需要协调其他两位组员，以期大家一起团家合作、戮力同心，创造出“1 + 1 ＞ 2”的效果。在近一年的合作中，我们团队有过争辩，甚至为了某个问题而发生争执，但大家都能够顾全大局、识大体，为了团队的整体利益而暂时搁置争议。总的来说，我和其他两位队友的合作是相当愉快的，队友都很給力——对于自己负责的任务都能够满额甚至超额完成，会主动帮助献言献策，这一点令我感动非常。

取得成就

（1） 知识与技能方面

前期我主要负责项目整体方案的设计和申报书的撰写，在和队友一起确定了选题之后，我随即着手查阅资料。在翻阅资料的过程中，我学会了如何有效地检索文献。在查找了相当的资料后，我确信自己已经对选题有了一定认识，随机开始设计项目整体的技术流程，由于没有和队友共同商讨方案的细节，导致后期在实施方案时遇到了很多疑难。虽然方案的设计与申报书的撰写走了不少弯路，但是我锻炼了自己独立搜集文献、设计技术流程与报告写作的能力。原定的技术方案有许多漏洞，具体如图 1 所示。

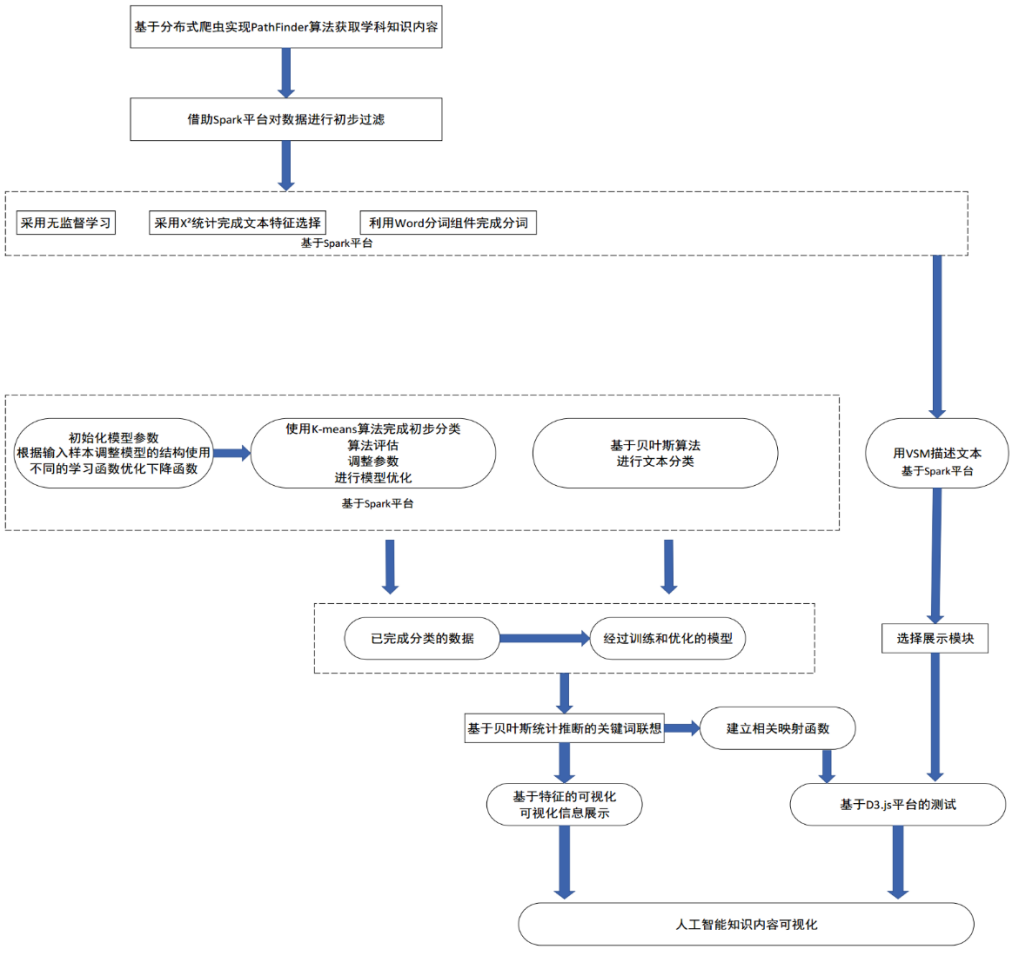


图 1 项目技术方案

在项目实施阶段，我和另外两个队友一起对整个项目进行攻关，我主要负责数据处理部分，即：数据存储、数据预处理（清洗、过滤等）、文本分词、标记文本、训练模型、提取知识关系等。许多原定的技术手段没有用武之地，有些之前未顾及的方面反而成为非常棘手的难题。例如：既定方案未将关系抽取看作是一个分类问题，采用 x^2 统计完成文本特征选择并用贝叶斯 Bayes）方法对文本所得到的分词，在训练语料的基础上构造分类器，进行文本分类；在实验中我发现这样操作带来了很大时空复杂性，且流程复杂，效果欠佳。后来遂改用了国产的自然语言处理工具 Jiagu 进行知识提取。此外，既定方案拟借助分布式爬虫实现 PathFinder 算法获取学科知识数据，但在实验中队友发现，在被爬取网站的内容经过原始排序时，所获得数据的质量与未使用 PathFinder 算法差距甚微。经过讨论，我们遂放弃了原定的数据爬取策略，而改用主从分布式爬虫。印象最深刻的莫过于知识图谱可视化工具的变更，原打算使用 JavaScript 的 D3 函数库完成数据的可视化，但实践发现 amChart 4 是更为理想的选择，其与 TypeScript、Angular、React、Vue 和纯 JavaScript(ES6)进行了原生集成[25]。为了能够操作 Spark 平台，我在几天内自学了一门以前完全陌生的语言：Scala。对 Scala 学习从初始的好奇，到后来的得心应手，我经历了一个计算机门外汉才会的生疏和紧张。

系统开发完成后，我负责论文和主要报告书的撰写，这让我初步体会到了论文写作的艰辛与不易，但也激发了我将来从事真正的科研事业的欲望。同时，在广泛查阅数据库、大数据、爬虫、可视化与 Linux 等资料的过程中，丰富了我的知识储备。

（2） 个人综合素质方面

除了知识与技能有所提高，我个人的综合素质也得到了很大提升。比如团队合作的重要性，大创任务繁重，任何个人想要独立完成都是十分困难的，换言之：任何工作都不可能是独立完成的，在学习过程中需要同学帮助，在工作中需要各部门合作，软件开发也是一样。我的耐心也得到了极大的考验，在编程的过程中经常遇到棘手的、长时间无法爬出的 bug，千言万语，最终我都予以一一克服。对于我自己做事情就应该有毅力，尤其是写代码，坚持到最后就是胜利。计算机是严谨的机器，当问题出现的时候，一定要有耐心去解决问题。态度决定一切。

心得体会

项目的每一个过程都需要队友间的协作与配合。项目确定初，我们三人一起查资料，再一起交流讨论，得到了方方面面的消息，了解了完成项目所需知识与项目的全貌，完善了项目的实施方案。

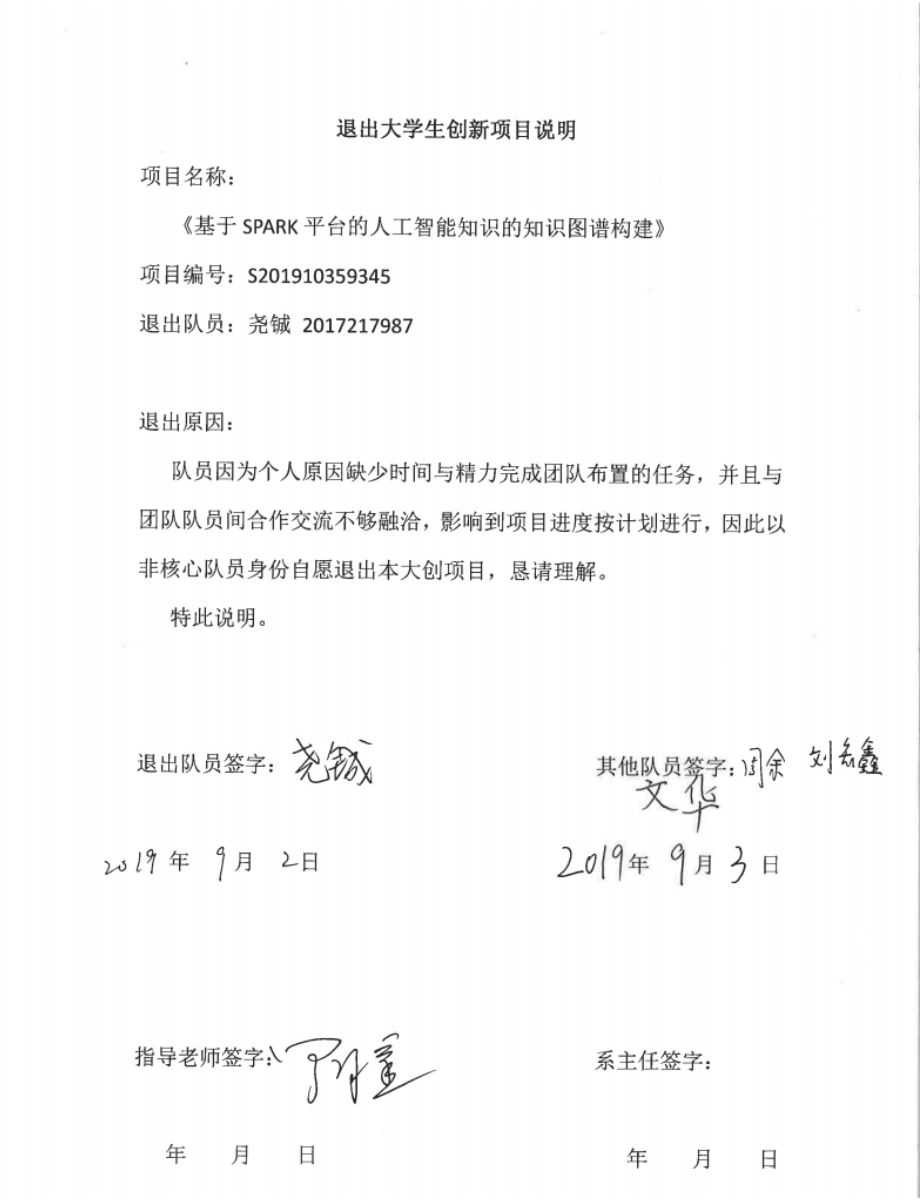
完成项目需要投入大量的时间与精力，这不是一个人就可以做到的，需要大家相互配合。每个人都要努力进自己所能，尽可能多承担一些工作。同时，团队当中的合作需要我们不断的磨合，学会倾听大家的意见和分享你的看法，做到尊重每一个你的组员，成员之间应互相帮助，高效快速的完成本项工作，以便尽快进行下一项工作。参与此次创业项目让我学会了合理安排时间，更加理解协作精神与团队意识的真谛，这对我的团结意识、协作意识、个人能力的培养提供了一个宝贵的机会。

项目开始之后，摆在面前的问题有很多，集体讨论的时间也不应该太长，不然就会变得斗志

消沉，所以，将问题统计出来，解决了得就先解决，解决不了的就要大化小，分解去解决，几个人讨论好对策，集思广益，解决问题的能力开始变快，掌握新知识的速度也就加快了。

不得不提的是，就是要考虑不同团队成员的诉求。有些同学参赛是纯属兴趣，有些同学是为了学分，而有些同学为了刷简历，甚至不排除有个别人有其特殊的“利益”需求。我们项目进行途中发生了队员退队的情况，作为负责人的我极力求得双赢，但还是造成了比较尴尬的局面。然扪心自问，其他队员有目共睹：那位队员的退队完全是其“个人原因”，如图 2 所示为其所写的指导老师签字的个人退队说明书截图（说明书在指导老师签字后即由我转交給这名队员，其拿給有关领导签字后直接交給了系里面主管大创的工作人员）。

历时一年的大创终于完成了！



退出大学生创新项目说明

项目名称：

《基于 SPARK 平台的人工智能知识的知识图谱构建》

项目编号：S201910359345

退出队员：尧铖 2017217987

退出原因：

队员因为个人原因缺少时间与精力完成团队布置的任务，并且与团队队员间合作交流不够融洽，影响到项目进度按计划进行，因此以非核心队员身份自愿退出本大创项目，恳请理解。

特此说明。

退出队员签字：尧铖　　其他队员签字：周余 刘嘉鑫 文华

2019 年 9 月 2日　　2019年 9月 3 日

指导老师签字：　　系主任签字：

年　月　日　　年　月　日

注：尧铖同学退队申请书

个人总结报告—刘宏鑫

在过去的一年多时间里，自己和团队的两个小伙伴一直专心于本项目的研究。我现在还依稀记得在去年 4 月份，我们组申报大学生创新创业项目的场景，那时候的我们什么都不懂。可能是出于好奇心，我个人对大学生创新创业项目特别感兴趣，尤其是觉得自己能动手做项目，真正的把所学的专业知识变成实际的应用当中。基于这个出发点，我们组成了本次大学生创新创业项目的小团队。团队里面的小伙伴对于专业课程掌握知识比较良好，动手能力也特别强，加之一系列的机缘巧合，最终让我们走到了一起。从刚开始我们对于本项目的一个非常具有创新的想法开始，虽然那时候我们什么都不懂。只能通过校图书馆、中国知网等一些具有权威性的资料宝库寻找我们想要的资料。包括项目背景调研、国内外的状况、可行性分析、成本以及最重要的技术途径。当时的我们只是单纯的想把自己的想法通过大学生创新创业项目的形式展现，并且做出实质性的东西出来。经过这一年多的工作，以来体会到了非常多的感受。下面我就个人的工作状况，做出个人的工作总结。

项目刚开始的时候，我对于本项目的了解可以说是几乎为 0，自己都是靠着我们的组长（项目主持人）通过技术，可行性等分析，把我们整个项目的流程和大体框架梳理了一遍。我们通过一次次的内部会议逐渐把项目的大体框架和实现的技术方案给摸清楚。首先是确定项目的流程，本项目的流程简化而言可以分为以下三个阶段：数据获取、数据处理、数据可视化。而我个人在本次项目中承担的是第三阶段------数据可视化的相关工作，其余的两个最为重要的阶段由我们组的其他两名队员完成。除此之外我个人在项目中还承担以下工作：资料的收集、项目技术方案的分析、项目技术学习途径、项目可视化方案和框架的设计、项目相关的文档整理等等。由于我主要是负责项目可视化方案和框架的设计，所以在这方面自己的体会最为深刻。首先，我们小组通过内部会议，确定了我们的技术路线的最后一步------可视化平台的搭建。主要有 C/S 和 B/S 架构两个可行方案，由于我们的目标是做出具有通用性的知识图谱工具，加之现在 Web 3.0 的兴起，Web 平台广泛用于大数据、人工智能等可视化的应用，所以我们选择了 B/S 架构。B/S 架构的产品明显体现着更为方便的特性，便于维护和升级。我们构想的是首先用 HTML/CSS/JavaScript 先搭起一个基本的可视化框架，主要先达到我们对于数据可视化的预期效果。由于我们所学的专业课程知识不够，于是我们充分利用暑假 1 个半月的时间，集中的系统性学习自己负责工作任务的技术。由于我是负责可视化方案和框架的设计，因此我在暑假期间通过 W3 School、中国大学 MOOC、CSDN、博客园等学习平台，把 HTML/CSS/JavaScript 的知识学习了一遍，并且利用一些简单的样例，快速上手，完成并实现可视化方案和框架的设计。就这样我设计出了第 1 代可视化方案，如下图 1 所示。

图 1 初代可视化方案设计

框架主要分为用户交互搜索区和知识图谱主体区两个部分，由于初代方案设计出来时，我们的知识图谱数据处理阶段的任务还没有完全完成，所以缺少大量数据的测试，但这也恰巧为我们后续的改进埋下了伏笔。

在项目有了进一步的进展之后，自己的技术水平也有了充分的提高。此时，队友们的技术也已经基本成型。在确定了首批测试数据之后，我们对原先的方案进行了非常大的改进，同时增加了许多新的元素和板块，为例实现更贴近浏览器用户的体验。我们升级了相关的 Web 技术，将模块化的思想和 HTML5 和 CSS3 等新技术联合，设计出了第 2 代的可视化方案，如下图 2 所示。

图 2 改进的可视化方案设计

由上图可以看出，在初代可视化方案设计的基础上，我们首先优化了界面和用户的体验。与此同时，我们增加了，主题筛选的功能，可以让用户选择丰富的主题，增加了可操作性。

在数据获取、数据处理两个阶段的任务全部完成后，我们对于数据的可视化方案进行了最终的设计，为了体现出与匹配人工智能和大数据相关度，我们将之前设计好的框架搭建在服务器平台上，同时将一部分的数据处理工作和知识实体的统计工作交给服务器完成，大大节省了本地主机的资源，增加了通用性。最终实现的可视化方案，如下图 3 所示。

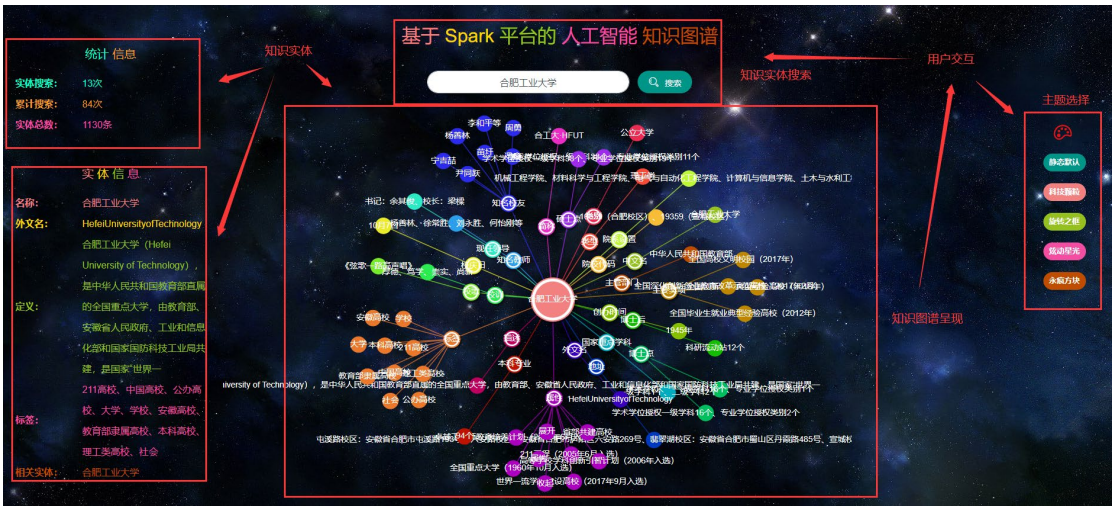


图 3 最终的可视化方案设计

由上图可以看出我们在第 2 代方案的基础上，增加了统计信息和实体信息等模块，增加了本项目的用户体验性。而新增加的部分也恰巧将知识图谱带来的强大语义发挥到了极致。

创新驱动发展是当今社会的主旋律，随着各行各业对创新能力、对科技的偏好，创新能力培养已经成为大学教育最值得重视的内容。总得来说，自己在团队的工作虽然不如其他两位小伙伴艰巨，但是自己对于团队的贡献也是非常大的。在本次项目中，自己学到的不单单是自己在做项目的技术和自己学习的课外知识动手能力等，我觉得更多的是，自己在整个项目的流程下所学到的经验，以及自己碰到问题，独立解决问题的能力。对于非常困难的问题的挑战。在整个项目的立项、设计方案、学习技术、搭建环境、实现项目等等的过程中，自己学到的东西非常非常的多。首先是项目对于我个人能力的培养，在这一年多的时间里，我先是从一个什么都不懂的小白，通过自己查资料，课外学习，一步步培养自己快速学习和综合能力提升的能力。让自己跳出舒适区，在不断的思考和进取中提升自己的综合素质，也许这就是创新创业项目给我带来的最大的收获吧。

个人总结报告—周余

这次能够参与大创是在机缘巧合下促成的。当时正在台湾作交换生，如果参加大创的话，没办法和队友及时交流，因此也没想到会得到负责人的邀请，对此真的非常感谢。同时因为加入的时间比较晚，初次审核的内容基本上团队成员都已经准备好，匆匆忙忙的看了材料以及在一些很小的方面，如论文修改等提出自己的建议。初次答辩，队友也准备的十分充分，顺利通过。在暑假之前，我们团队负责人初步安排了每位成员的大创学习任务。我们在完成学校规定的学习课程之外，也在为大创做着准备。并且把各自的学习进度同步到了 git 仓库。

此次大创，我主要负责的是编写程序对 CSDN、博客园中关于人工智能的三个领域：机器学习、自然语言处理、机器视觉进行爬取以及图谱可视化的 PHP 操作和少部分 MYSQL。暑假时负责人将我们聚集在了一起，互相鼓励学习以及开始各自的任务。

在选择编写爬虫的语言方面，我选择了 python。python 不单用于抓取网页文档的接口简洁，同时其访问网页文档的 API 也相当完整(如 urllib2 库)；除此之外，很多网站会封杀爬虫。这时我们需要构造合适的请求来模拟浏览器的行为，而 python 提供了非常优秀的第三方包如 Requests，mechanize。在抓取了网页之后，还需进一步的处理（比如过滤 html 标签，提取文本等）。而 python 的 beautifulsoap 库等能让我们编写非常简洁的代码即可完成大部分文档的处理。虽然一些其它语言或工具来实现以上功能，但是 python 是最快速简洁的语言。

在编写爬虫的时候也遇到各种各样的问题，记录如下:

1.robots.txt

robots.txt 通常存放于网站根目录下，负责告知我们哪些内容是不能被获取的，哪些是能被获取的。其实际上并不是一个必须遵守的规范。如果爬虫不遵守，那么网站的隐私是不能通过这种方式来防止泄露的。

2.头文件 headers

浏览器和爬虫程序，通常会向服务器发一个头文件：headers 以发送网络请求。这里面的大多数的字段都是浏览器对服务器的标识，因此大多数时候我们编写的爬虫，都需要伪造 headers 以掩盖爬虫的身份，向服务器发送请求。而符合该网站要求的 headers，可打开浏览器开发者工具->选择 Network->刷新页面->选择最上方的文件->找到 request headers 即可

3.验证码

部分网站会要求输入验证码，应使用 python 的 selenium 库，模拟浏览器，网站会要求扫二维码进行验证，扫二维码即可解决，但这个策略并不总是有效，因而验证码一直是反爬虫一个非常有效的手段，

4.IP 限制

封 ip 和封账号也是非常有效的反爬虫策略封掉了 IP，网页返回的状态码为 302，解决方法:

(1)限制程序抓取频率，每隔几秒登录一次（这种方法工作量巨大）

(2)用多个账号、或者多台机器抓取，既处理了爬虫问题，也等同于做了分流处理，可以降低单台机器带宽压力（此方法在此次项目实现有点困难，不现实，仅供参考）

(3)使用代理 IP，先写一个爬虫程序对免费代理 IP 的相关网站进行爬取。此爬虫程序较为复杂，这次选择的是使用在 github 上开源的程序(github 上有详细的使用说明，以及需要的环境，非常方便，有效的解决了此次项目的问题)，虽然免费网站的代理 IP，每一天能用的数量少，质量不高，爬取速度非常慢。但是也足以解决这次项目中 IP 被限制的问题

5.cookie

通过 cookie 限制抓取信息。 此问题在抓取博客园，csdn 时均遇到，我们选择的是在伪造请求头时，因为 headers 里包含 cookie，同时伪造 cookie 即可，通过浏览器的 F12 查看器，然后逐个尝试在整个过程都请求的 URL(主要包括 HTML、JS、XHR)，直到成功为止，一般情况下是最上方的 HTML 包含了我们所需要的 cookie。

6.JS 渲染

采用 JS 渲染页面。 有些时候返回的页面是由 JS 操作 DOM 得到，我们实际上拿不到正确数据。这个也是非常有效的反爬虫策略。我们选择使用 selenium 库模拟浏览器请求返回渲染完 JS 的页面。

在 PHP 方面，我是负责用 MYSQL 语言编写存储过程与触发器，并让 PHP 与 MYSQL 交互，从而使得形成前端->PHP->MYSQL 动态的数据交互，完成可视化操作。

在这次完成大创项目的过程中，我熟悉了 Python，PHP，MYSQL 等三种语言。同时因为在对数据进行爬取后会大概的浏览一下内容,我对人工智能相关知识有了进一步的认识，同时也看到了一些比较优秀的文章。最难忘的便是暑假的时候，负责人把我们聚在一起。这是第一次体会到团队的感觉，大家在一起学习，互相交流思想，互相提醒监督，一起成长。非常感谢团队里的其它成员，尤其是负责人。规划了我们这个项目的基本方向，以及及时的与成员交流，推动项目的进行。在暑假里，大家都借了许多书籍，我个人也在负责人的帮助下，对 scala 以及 spark 有了一定的了解。而在暑假过后的那一学期，我们某项课程的实验便要求用 Python 来完成，得益于在大创的时候已经对 Python 使用的比较熟练了，实验完成的十分顺利。与此同时，我们凭借着这个项目参加了安徽省 ican 创新创业大赛，并最终获得了省二等奖。

这次大创不仅仅收获了素质方面的提高，能力的培养，更重要的是明白了团队协作的重要性。以及在一个团队里，每一个人都要付出，每个人都需要尽力为团队做出贡献。我们的团队只有三个人，很多另外的项目团队人数超过我们许多。但是我们还是最终完成了这个大创项目。非常感谢，以及非常荣幸能和两位同专业的同学成为队友。我从他们身上学习到了很多优秀的品质。无论是面对难题时的冷静思考，还是答辩场上的镇定自若。他们都在闪闪发光。

支撑材料

iCAN 国际创新创业大赛

组委会发〔2019〕31 号

关于公布 2019 年第十三届 iCAN 国际创新创业大赛安徽赛区比赛结果的通知

2019 年第十三届 iCAN 国际创新创业大赛安徽赛区比赛圆满落幕，以下为安徽赛区比赛中获奖的参赛队伍（同等奖项中排名不分先后），其中，一等奖获奖团队将代表安徽赛区参加 2019 年 iCAN 国际创新创业大赛中国总决赛的比拼！

一等奖		
作品名称	队长学校	队长
智能旋转单车之家	合肥工业大学（宣城校区）	徐峥岩
”心晴”音乐小熊-青少年可视化编程的新型教具	合肥工业大学（宣城校区）	吕雯昕
TDG-W 动态称重智能控制系统	合肥工业大学（宣城校区）	黄文君
一种基于 APM 飞控的水下无人机设计	合肥工业大学	李恒
基于 openMV 的多功能智能护眼台灯	合肥工业大学	许乾
Cloud 智测-支持云服务的汽车智能测试系统	合肥工业大学	潘斌
下水道气体检测四轮机器人	合肥工业大学	梁燚
球星	合肥工业大学	陈宇轩
鲜度呼吸 Fresh Breath	合肥工业大学	郑雨
可视化智能外卖骑手头盔	合肥工业大学	蒋光睿
印吧——全国领先的云印生态链	安徽大学	秦超凡
购物精灵	合肥工业大学	孙东武
基于机器学习的小麦赤霉病智能诊断装置	安徽大学	杜志强
基于声波发电原理的新型城市公路噪声发电装置	合肥工业大学	卢思婷
安全驾驶智能助手	中国人民解放军陆军军官学院	杨建新

基于 turtlebot3 的安全检测机器人	合肥工业大学	庄树理
海参养殖中基于低功耗 LoRa 传输技术的远程智能测控系统	安徽大学	潘显华
智能户外晾衣架	河海大学文天学院	石恒鹏
基于视觉感知的智能避障小车	合肥工业大学	朱月婷
基于 ESP8266 芯片云控制的小车自动巡航系统	合肥工业大学（宣城校区）	许人航
e-packing 一体化自动包装机	合肥工业大学（宣城校区）	汪健
二等奖		
作品名称	队长学校	队长
基于 APM 飞控的水下无人机	合肥工业大学	何佳钟
多功能水上移动平台	合肥工业大学	张皓源
果核智能——末端执行器的设计与应用	安徽工业大学	余林凤
高性能柔性锌离子微型电池	安徽大学	李子彬
水质及鱼群监测---仿生机械鱼	合肥工业大学	王瑞坤
I-Guider 多功能导盲杖	安徽大学	王璐琳
智能太阳能地源热泵复合热泵系统	安徽理工大学	杨腾飞
基于计算机视觉的小区智能监控系统	合肥工业大学	钱洋洋
智谷局部高品质空气供应系统	安徽工业大学	汪稼钰
水果「别」送	合肥工业大学	滕娴
无人机树障电力巡线系统	合肥工业大学	沈涵
智能球形机器人	合肥工业大学	王鹏
翔宇高楼逃生装置	安徽工业大学	潘亚洲
基于声波检测原理的家用水管检漏出设备	合肥工业大学	王天霁
远程控制探伤爬墙车	合肥工业大学（宣城校区）	张子剑
基于麦克纳姆轮和 LPC54606 芯片控制的智能轮椅	合肥工业大学（宣城校区）	张倩倩
“Freedom”智能购物车	合肥工业大学（宣城校区）	闫国祚
一种结构可调式褶形滤袋框架	安徽工业大学	曹博文
基于 Mask R-CNN 的智能交通检测与控制系统	安徽师范大学	许晨晨
二维码智能门禁访客系统	合肥工业大学	葛晓露
基于 RFID 技术的智能结算购物车	合肥工业大学	张明伟
基于 Spark 平台的人工智能知识的知识图谱构建	合肥工业大学（宣城校区）	文华
智能图书污损检测及分类装置	合肥工业大学	尹德强
基于 3D 打印原理的咖啡拉花机	合肥工业大学	文智贤

基于 Spark 的人工智能知识图谱构建

文华，刘宏鑫，周余

摘　要：随着计算机大数据的快速发展，可以借助于互联网平台的各种工具找到有价值内容，但海量数据給筛选、组织与评价带来极大困难。知识图谱具有强大的语义处理与开放互联能力，可以精确地表达概念及其相互关系所构成地语义网络，更好地为机器所理解；且能够帮助用户快速、准确地检索所需要地信息。本文基于 Spark 平台构建了人工智能中的机器学习、自然语言处理与机器视觉等三个领域的知识图谱，完成相关知识的重整，取得了较好的实验效果。

关键词：知识图谱；Spark；可视化

Abstract:

With the rapid development of computer data, it has been developed into reality, finding valuable content with the help of various tools of Interne. However, the massive data has brought great hardships to screening, organization and evaluation. Knowledge map owns a powerful semantic processing and opens interconnection ability, which can accurately express the semantic network formed by concepts and their mutual relations that can be understood by the machine better. Furthermore, it can help users retrieve the required information quickly and accurately. Based on the spark platform, this paper constructs a knowledge map of machine learning, natural language processing and machine vision that are related to Artificial Intelligence, which completes the reorganization of relevant knowledge, and achieves good experimental results.

Keywords: Knowledge Graph; Spark; Visualization

0. 引言

人工智能（Artificial Intelligence，简称 AI），是当前最热门研究领域之一，甚至被誉作世界三大尖端技术之一[1]，近年来我国甚至将其上升到国家战略的高度：2017、2018 与 2019 年的政府工作报告中均被提及[2-4]。可见，人工智能在现代科学技术与经济社会中有着不可替代的地位，随着 5G 时代的到来，人工智能必将展现更广阔的应用前景。与此同时，人工智能相关方向的人才匮乏也正越来越成为（市场）关注的议题[5]，而在培养人才时，如何准确把握所授相关领域知识的准确性、全面性与前沿性成了一个难题，知识图谱（Knowledge Graph）是解决这一难题的有效工具。知识图谱是人工智能领域重要的一个技术分支，其目的是将现有的人类知识构建为一个结构化的知识库。目前，已经有许多大型知识图谱被构建出来，如 DBpedia 、Freebase 等，然而，当前的知识图谱工具普遍存在以下问题：1）通用知识图谱工具涉面较广，但知识冗余混乱、组织零散、系统性差，不利于用户的专业学习；2）垂直知识图谱工具种类少，成熟的应用仅限于某些领域，在一些具有较大应用需求的领域未获重视，前景广阔。

综上所述，本文的目的是构建一个面向学习者尤其是本科生的人工智能领域的垂直知识图谱。人工智能领域繁多，我们选取机器学习（Machine Learning，ML）、自然语言处理（Natural Language Processing，NLP）与机器视觉（Machine Vision，MV）等三个领域作为代表。

1. 相关工作

知识图谱的构建技术仍在持续发展中，目前存在多种流派，每一种技术手段途径各异、效果良莠不齐随着相关技术的不断演变与发展，新的知识图谱构建方法被不断推出，有些研究也在尝试使用经典的方法在新的应用领域构建相应的垂直知识图谱，均取得了一定效果。构建知识图谱的一般技术流程如图 1.1 所示。

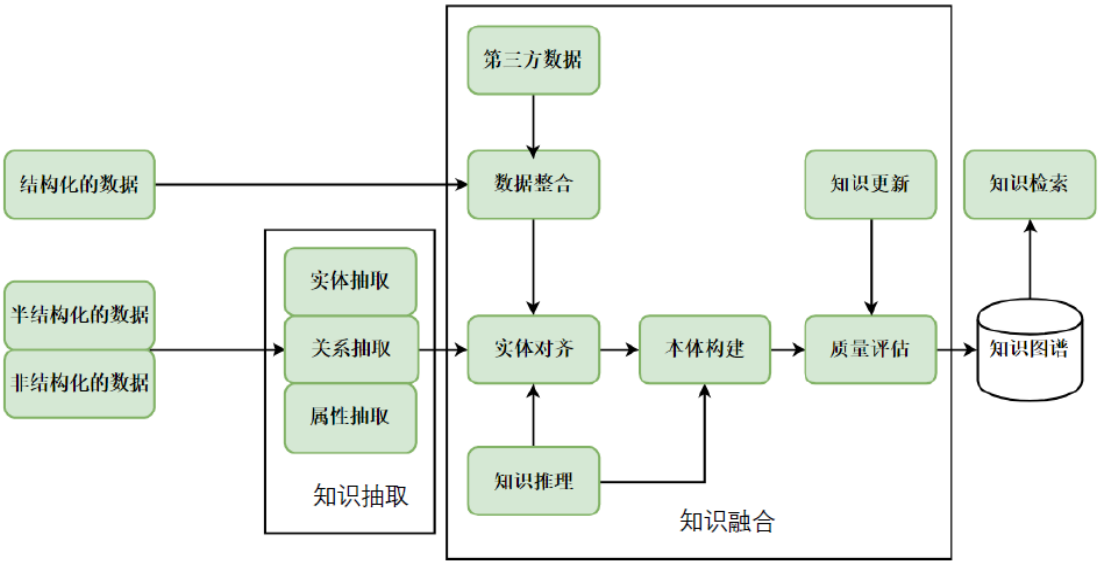


图 1.1 知识图谱构建流程

金婧等[6]侧重于知识图谱表示学习方法，在 TransE[7]模型的基础上提出了一种融合实体类别信息的知识表示学习模型（TEKRL），实验表明该模型在各项评价指标上得到了提升；杨玉基等[8]在对领域知识图谱的系统研究上，提出了一种构建领域知识图谱的“四步法”，该方法可以在较短时间内构建准确率较高的学科知识图谱；孙昊天等[9]实现了一种基于带权三元组（Unit Triplet）构建时政类知识图谱的方法，该方法在参数设置得当的情况下可以得到较为理想的以亲密程度为关系的知识图谱；董永强等[10]提出了一种基于 YANG[11]模型由数据模型驱动（Data-Model-Driven）的网络领域知识图谱构建方法，通过该方法构建的知识图谱可为网络维护大数据（Big Data）提供支持，降低了人工成本。

而在通过经典方法构建垂直知识图谱上，熊晶等[12] 基于多源异构数据源构建了甲骨学融合知识图谱，所得的知识图谱节点较多，可以满足甲骨学研究的基本需求；刘燕等[13]利用相关技术构建了医学知识图谱，在医药卫生知识服务系统平台取得了理想的效果；白如江等[14]提出科学事件（Scientific Events）的概念，并利用 LTP[15]语言云根据所谓科学事件模型构建了图情（Library Information）领域的知识图谱，实验结果差强人意；陈成等[16] 提出了意图知识图谱的定义并完成了构建，通过有关范例说明了该图谱可以作为政府治理的一种依据。

有鉴于新兴理论与技术在构建知识图谱，以及使用经典方法在新的应用领域构建有关垂直知识图谱所取得的成功与不足，本文基于大数据处理平台 Spark，并借助 Jiagu 模型出色的知识关系提取能力，并使用从国内两大流行的技术博客平台 CSDN 与博客园爬取到的元数据，构建了一个学习者尤其是本科生适用的人工智能领域的知识图谱。

2. 数据来源

2.1. 爬取工具的选择

本文选择 CSDN 与博客园作为主要的元数据（Metadata）获取平台，因其主要数据采用网页来展现，所以本文选择 Python 作为爬取工具。python 不但用于抓取网页文档的接口简洁，同时其访问网页文档的 API 也相当完整。

值得一提的是，抓取网页有时需将爬虫（Crawler）程序伪装成普通的浏览器。因为许多网站都采取了防爬措施，单纯的爬取操作极容易被网站检测出来并封杀。Python 提供了许多鲁棒的第三方包如 requests、mechanize、selenium，可以帮助爬虫轻松地越过网站的防爬策略。

在抓取了网页之后，仍需进一步的处理，如过滤 html 标签，提取文本等，而 python 的 beautifulsoap 库等使编写非常简洁的代码即可完成大部分文档的处理成为可能。

2.2. 提高爬取效率的方法

传统的网络爬虫是运行在本地，稍优化的策略是采取“单机多核”的方式。为了更有效地解决爬取效率过低的问题，同时结合实际的实验条件，本文采用主从分布式爬虫（Master-Slave Distributed Crawler）[17]。

本文将一台阿里云服务器作为 master 服务器，用于分发所需爬取内容的 URL，同时维护存储在 redis 中待爬取 URL 的列表。由三台本地的笔记本电脑组成 slave 服务器组，用于对各自从 master 服务器所获得的 URL 执行网页爬取任务；若 slave 在爬取过程

中遇到新的 URL，一律将其返回 master 服务器由 master 解析处理，slave 服务器间不进行通信。本文所用 master 服务器与 slave 服务器组的性能配置如表 2.1 所示，主从分布式爬虫的逻辑结构如图 2.1 所示，爬虫的类图结构如图 2.2 所示。

表 2.1 master 服务器与 slave 服务器组性能配置

Server	Processor	RAM/GB	Storage/GB	CPU core(s)
master	Intel(R) Xeon(R) CPU E5-2682 v4 @ 2.50GHz	2	40	1
slave 1	Intel(R) Core(TM) i5-8300H CPU @ 2.30GHz 2.30 GHz	16	128(SSD) + 1024	4
slave 2	Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz 1.99 GHz	16	128(SSD) + 1024	4
slave 3	Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz 2.21 GHz	16	128(SSD) + 1024	6

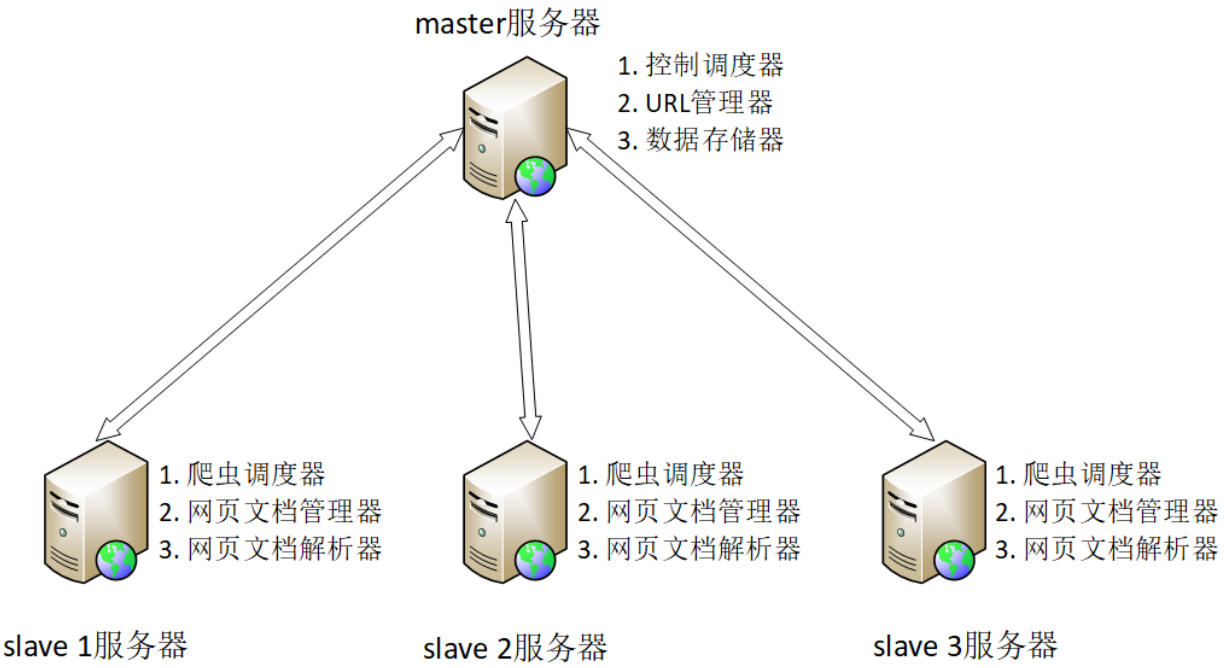


图 2.1 主从式分布爬虫逻辑结构

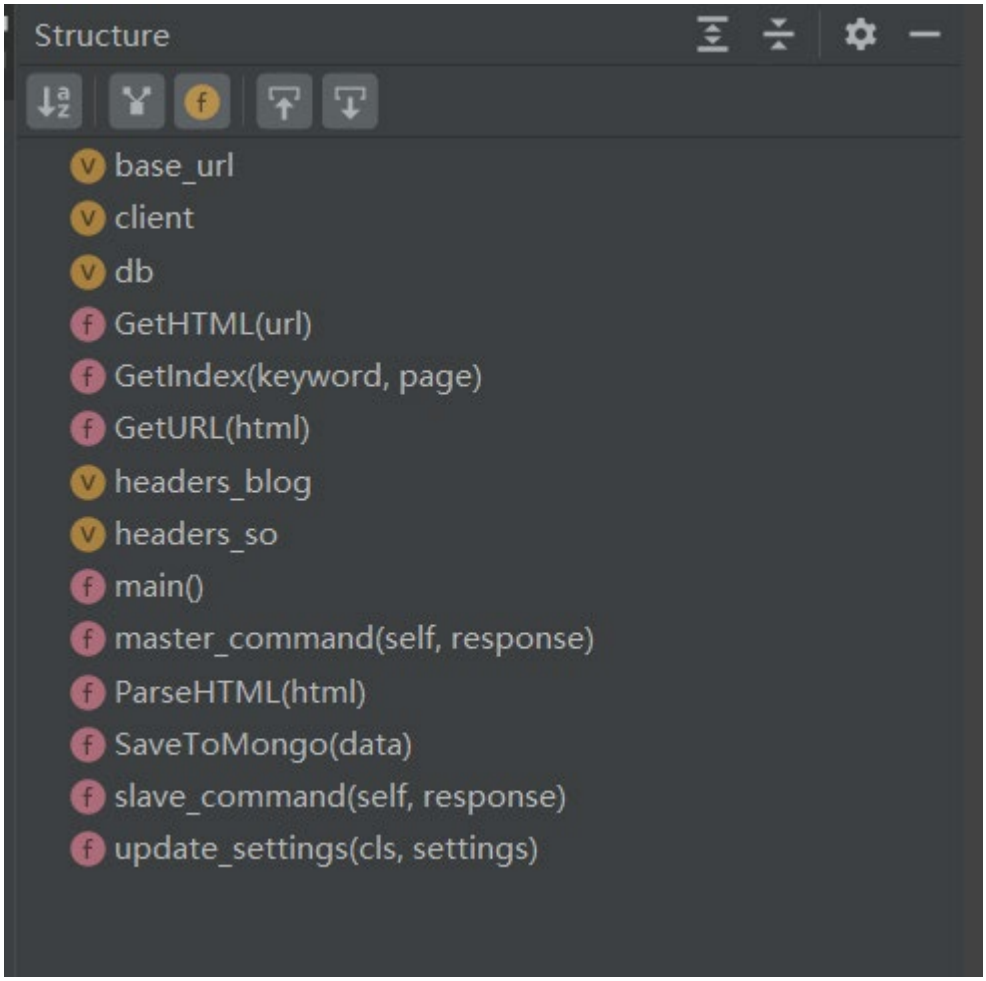


图 2.2 爬虫程序的类图结构

此外，为了防止网站服务器锁定爬虫的 IP，本文所使用的爬虫程序对爬取频率进行了限制，以及使用代理 IP 池。

3. Spark 与 Jiagu 模型

3.1. Spark 与 hive 平台

Spark[18]是 基于内存计算的大数据并行计算框架，因为它基于内存计算，所以提高了在大数据环境下数据处理的实时性，同时保证了高容错性和高可伸缩性，允许用户将

Spark 部署在大量廉价硬件之上，形成集群。hive[19]是一个基于 Hadoop 的数据仓库平台，通过 hive 我们可以快速地对存储在数据库中数据进行抽取、加载与转换（，Extract， Transform， Load， ETL）等操作。hive 定义了一个类似于 SQL 的查询语言：HQL，能够将用户编写的查询语句转化为相应的 Mapreduce 程序并基于 Hadoop 执行。需要注意的是，hive 本身并不存储数据，因而用户需要选择一个传统的数据库进行数据存储，基于可操作性与成本等角度考虑，本文采用 MySQL。

本文将使用 Spark 平台的相关工具进行数据预处理。

3.2. 数据预处理

第 2 节所爬取到的元数据杂源异质，散乱冗余，并且由于网页文本本身的结构导致数据中存在大量标签，无法直接用于下一步操作。因此本文借助 Spark 平台快速的数据处理能力及 hive 对数据库高效的 ETL 操作，对文本进行预处理。

首先，在 spark-shell 上将数据成功加载到 hive 中，为后续存取提供了数据来源。其次，在 hive 上创建了数据库，在 spark-shell 上依次将爬虫爬取的 json 文件导入成表。而后，在 IDEA 上编程对数据去重，这里主要使用了 Spark 的几个 API，如：duplicate、filter、regexp_replace、regexp_extract 等。完成数据的存储、去重和标签过滤后，借助于 github 上开源的敏感词汇库[20]，对表数据进行敏感词（Sensitive Word）过滤，以此得到更干净的数据。本文所用部分 spark-shell 处理命令如图 3.1，数据预处理的程序类图如图 3.2 所示，预处理后的部分数据如图 3.3 所示。

```
scala> val dataDF1 =
spark.read.format("json").load("file:///home/hadoop001/hadoop/data/Spider-
Data/cnblog_computer_version.json")
scala> dataDF1.select(dataDF1.col("author"),
dataDF1.col("content"),dataDF1.col("date"),
dataDF1.col("title")).write.saveAsTable("dachuangppreprocessingdata.cnblog_computer
_version")
```

图 3.1 spark-shell 处理命令

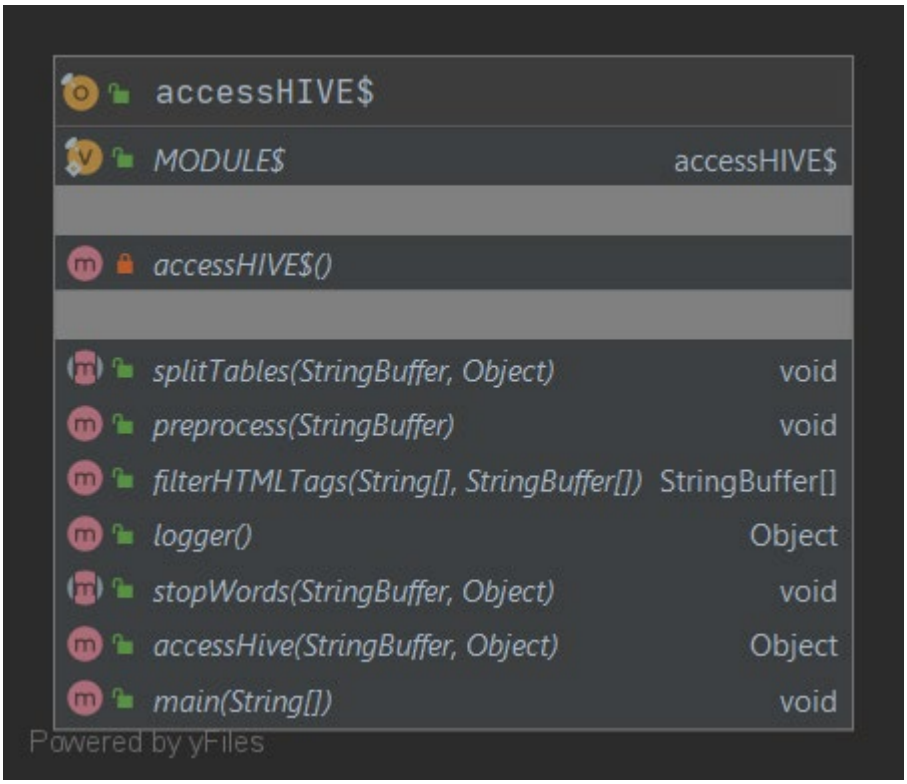


图 3.2 数据预处理程序的类图

```
提到 人工智能 ， 大多数 人 的 第一 反应 就是 距离 我们 太远 了 。 智能 机器人 、 无人驾驶 ， 这些 好像 都 是 未来式 。 我们
比如 ， 应用 最广 的 美颜 自拍 ， 更 准确 的 说 ， 是 人 像 处理 。
现在 的 人 像 软件 之所以 能 帮助 人们 从 繁琐 的 PS 中 解放出来 ， 就是 因为 利用 了 大量 计算机视觉 技术 ， 像是 人脸 定
今天 就 以 天天 P 图为例 ， 看看 是 什么 让 他们 成 了 一款 AI 软件 。
AI 赋能 ， 让 手机 如何 读懂 你 的 脸
人像 处理软件 之所以 能 成为 AI 产品 ， 是因为 有 了 大量 图片 数据 ， 尤其 是 人脸 数据 的 累积 。 而 通过 大量 图片 数据
以 天天 P 图 的 自动 美颜 功能 为例 ， 软件 之所以 能 放大 眼睛 、 添加 贴图 动效 ， 是因为 准确 的 找到 了 人脸 和 五官 的
在 每 帧 图像 中 准确 的 找到 人脸 和 五官 后 ， 就 可以 “ 加 特效 ” 了 — — 增加 美妆 、 萌宠 贴图 ， 自然美 妆 。
除了 对 人脸 的 识别 和 处理 ， 为了 给 用户 提供 更 多 丰富 智能 的 玩法 ， P 图 团队 还 联合 优图 团队 对 视频流 进行 了
背景 分割 也 是 另 一项 基于 AI 的 创造性 玩法 ， 通过 深度 优化 加速 后 的 神经网络 ， 使得 P 图 可以 在 移动 端 实现 对
打造 图像处理 云 ， 美颜 AI
能 做到 的 不仅仅 是 变脸
美颜 AI 能 做到 的 不仅仅 是 变脸 。
在 大多数 人 的 印象 中 ， 天天 P 图 这 类人 像 处理软件 即使 有 AI 技术 ， 基本 也 是 应用 于 自己 的 产品 之中 ， 缺乏
细心 的 人会 发现 ， 军装 H5 并非 在 终端 上 进行 运算 ， 而是 通过 H5 上 传到 云端 处理 。 基于 云端 的 人脸识别 ， 五官
除了 家喻户晓 的 军装 照 ， 天天 P 图 最近 推出 的 萌偶 功能 也 利用 了 AI 图像处理 云 。
通过 在 云端 的 神经网络 ， 找到 与 用户 五官 相似 度 最高 的 卡通 素材 。 建立 标准 人脸 和 标准 卡通 人脸 间 的 映射 关系
这些 能够 提供 丰富 玩法 的 AI 图像处理 云 ， 也 解决 了 深度 神经网络 模型 可能 过大 ， 无法 在 终端 运行 的 问题 。 天天
强大 的 分布式 部署 能力 降低 了 客户端 的 门槛 ， 使得 算法 可以 配适 各种 环境 ： 手机 、 电脑 、 电视 、 App 、 H5 … …
作为 用户 可能 很难 明确 感受 到 图像处理 云 的 存在 ， 但 这项 能力 却 为 天天 P 图 打开 了 更 多 依靠 AI 创造 营收 的 路
从 隐性 到 显性 ， 人像 处理 AI
```

图 3.3 预处理后的部分数据

3.3. Jiagu 模型

Jiagu 模型[21]是一个国产的开源自然语言处理工具，以 BiLSTM 等模型为基础，使用大规模语料训练而成。Jiagu 模型提供中文分词、词性标注、命名实体识别、情感分析、知识图谱关系抽取、关键词抽取、文本摘要、新词发现、情感分析、文本聚类等常用自然语言处理功能，API 丰富，且操作便捷、稳定性高。本文选择 Jiagu 模型作为知识抽取的工具，取得了十分理想的效果。

3.4. 知识抽取

在知识图谱中，知识一般以三元组(p, r, q)的形式来表示，其中 p 与 q 分别代表前后两个实体，r 代表前后实体之间的关系[22]。显然三元组是构建知识图谱的重要基础，三元组中实体间的关系是否准确、完整等也是知识图谱的构建成功与否的重要判据。

本文采用 BIO 方式[23]对待训练文本进行实体命名标记，每行一个字符，并按 19:5 的比例分别设置训练数据与验证数据，且为测试训练所得模型的准确程度设置了较训练数据 75%的测试数据，详细信息如表 3.1 所示。在分别调节学习率（Learning Rate）、迭代次数（Iterations）、阻尼系数（Damping Coefficient）等参数后对标记文本进行训练，参数详情如表 3.2 所示。实验结果用 held-out 方法[24]进行评估，即统计知识图谱中已有的实体被 Jiagu 模型检测出的数量，正确的实体被排序靠前的数量愈多，则在准确率/召回率曲线上，随着召回率（Recall Rate）的增长准确率（Accuracy Rating）就下降得越慢，也即知识抽取的质量愈高。实验结果的准确率/召回率曲线如图 3.4 所示，所得部分三元组如图 3.5 所示。

表 3.1 数据集的统计信息*[1]

数据集	关系数量	语料行数
训练集	10	2435796

* 在训练文本中，每行一个字符。

验证集		634547
测试集		1849620

表 3.2 所用训练参数

Learning rate	Iterations	Damping coefficient
0.001	50000	0.85

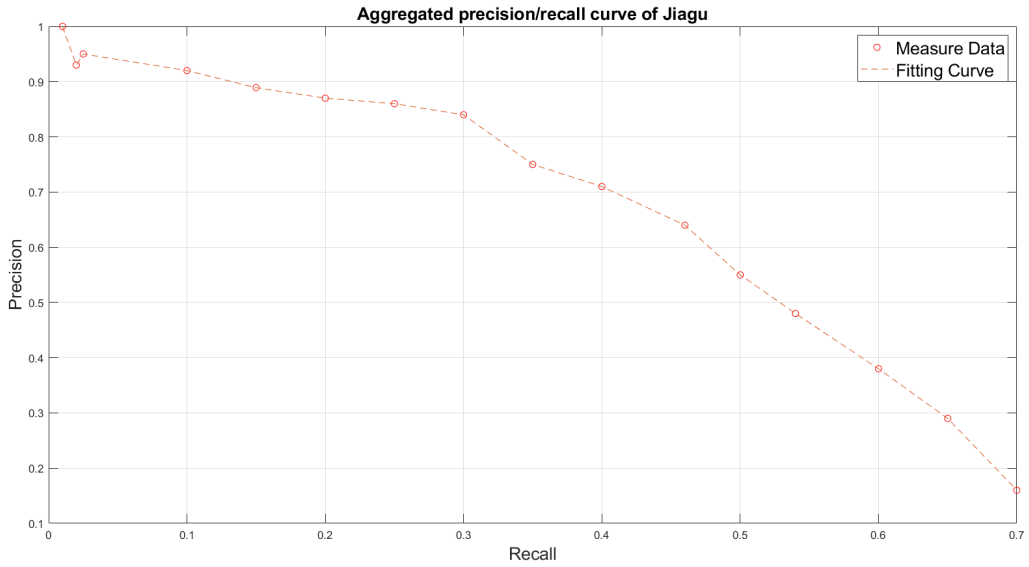


图 3.4 准确率/召回率

```
逻辑回归,优势,处理非线性效应
逻辑回归,缺点,仅用于二进制分类
随机森林,优势,防止过拟合
随机森林,用于1,回归
随机森林,用于2,分类
随机森林,缺点1,容易生长
随机森林,缺点2,随机子集高
评价矩阵,术语1,真阳性(TP)
评价矩阵,术语2,真阴性(TN)
评价矩阵,术语3,假阳性(FP)(I型错误)
评价矩阵,术语4,假阴性(FN)(II型错误)
特征选择,也称为1,变量选择
特征选择,也称为2,属性选择
特征选择,也称为3,变量子集
特征选择,选择,最佳相关特征
特征选择,帮助1,简化ML模型
特征选择,帮助2,提高ML模型的准确性
特征选择,有助于,更快地训练
特征选择,防止,过拟合
```

图 3.5 三元组数据

4. 知识图谱的可视化

4.1. 三元组的转化

本文所选可视化工具为基于 TypeScript 开源的可视化框架 amCharts 4，其与 TypeScript、Angular、React、Vue 和纯 JavaScript(ES6)进行了原生集成[25]。由于用户通过某个关键字请求实体的三元组信息时，其数据量可能是非常大的。此外，amCharts 4 要求数据以特定的 json 格式存储，显然 3.4 节所得的三元组无法直接用于可视化（Visualization）。出于存取效率、数据可拓展性等因素考虑，本文将三元组数据预先导

入 MySQL 数据库，当前端发出数据请求时，通过 PHP 编程实现从服务器端查找相应的原始三元组数据并使用相应 API 转换为 json 格式返回给前端。前端在接收到 PHP 返回的原始三元组数据后，需要对原始三元组数据进行预处理，将原始的 json 数据转化为 amCharts 可识别的特定格式 json 数组，并最终作为 amCharts 的数据源加载，渲染（Render）到指定的 SVG 画布上，最终形成可操作的力导向图谱。具体交互的流程如图 4.1 所示。

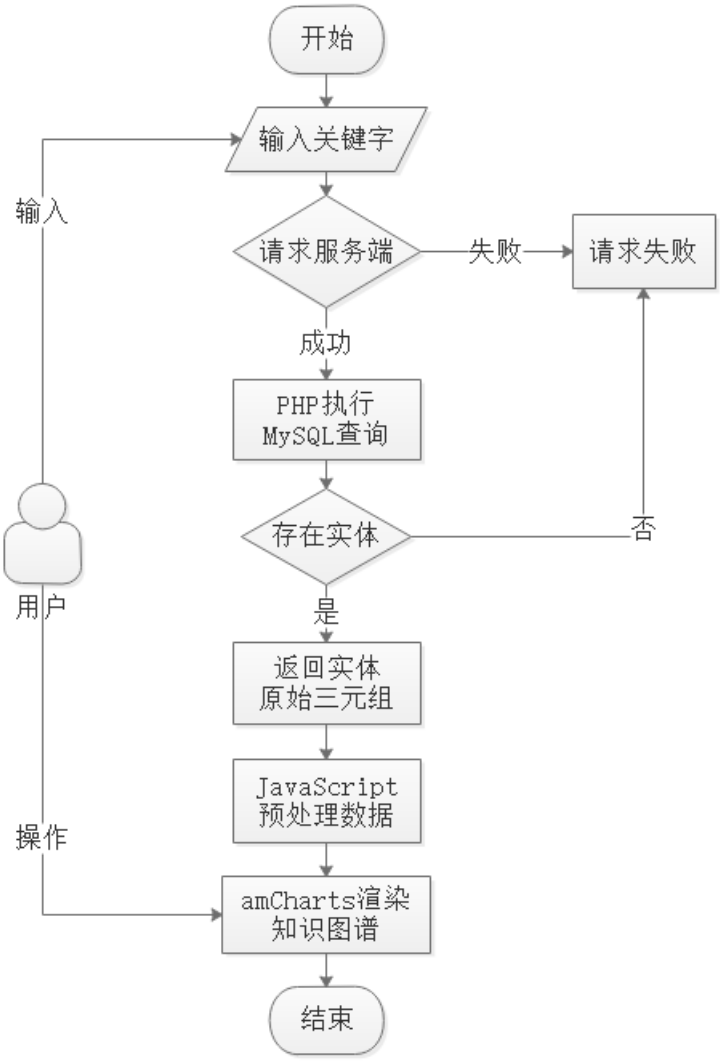
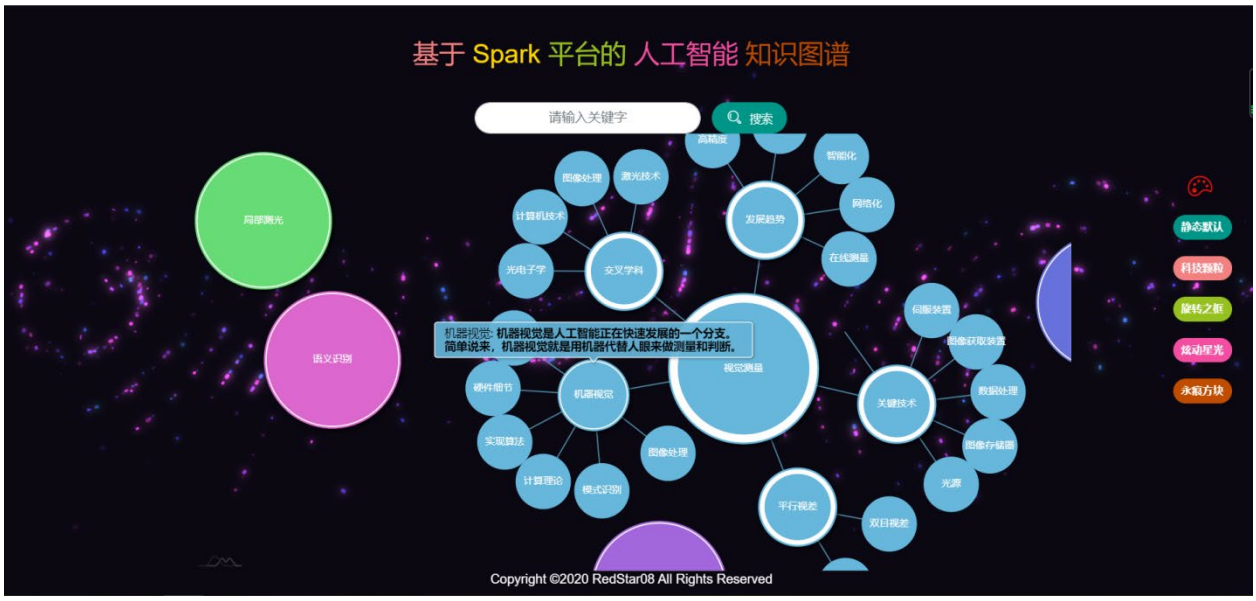


图 4.1 知识图谱可视化流程图

4. 2. 图谱可视化

amCharts 4 是一个基于 TypeScript 开源的可视化框架，具有图表种类丰富、图形效果炫丽、动画或静态呈现、与平台无关等特点，适用于各个行业的可视化需求场景，因此本文将其作为知识图谱的可视化工具。本文使用 HTML/CSS/JavaScript 设计页面元素及基本布局，并采用力导向图作为图谱的呈现形式。当用户在搜索框键入查询关键词时，通过 GET 请求关键字，后台通过 PHP 查询数据库并返回请求的数据。前端得到请求的数据后，通过 JavaScript 进行预处理并借助 amCharts 进行可视化展示。本文所构建知识图谱的可视化结果示例如图 4.2 所示，此外，本文所采用的图谱可视化工具支持多种主题背景的选择，如图 4.3 所示。

a)



b)

图 4.2 知识图谱可视化结果

图 4.3 丰富的主题选择

5. 结果与分析

本文成功地构建了人工智能领域的知识图谱，首次将本科计算机类专业的课程内容知识以知识图谱的形式展示出来；可以帮助用户准确、快速地检索人工智能领域相关术语并提供解释，同时给出术语的联想结果，利于用户进一步学习；形象化地展示人工智能领域的脉络、历史沿革与发展趋势，为用户复习、深入学习提供参考。

下一步的工作将从几个方面进行研究：采用知识联想等方法增加知识图谱中的知识实体规模，进一步优化知识关系抽取，改善知识融合等。

6. 结束语

垂直知识图谱的应用前景广阔，囿于构建技术尚在发展、仍未成熟，相关的产品较少。本文大胆对目前热门的人工智能领域进行了知识图谱构建，初步探索出了相关图谱的构建步骤，得到了效果较为理想的实验结果。本文的构建方法可以应用于大多数针对特定学科或领域的垂直知识图谱的构建，以期在扩大训练语料的基础上得到较本文实验结果覆盖率更广的领域知识，即规模更为庞大的 RDF。值得一提的是，本文在构建图谱的过程中认识到：汉语作为一门分析语所具备的固有特点是构建汉语知识图谱的障碍之一，在后续工作中或可以考虑以英文语料为基础构建知识图谱，待完成后再行翻译。

本文还以人工智能领域的机器学习、自然语言处理与机器视觉三个分支为例，介绍了构建相关垂直知识图谱的技术流程。以期能够抛砖引玉，使其他有志之士有所参考。

参考文献：

[1]邹蕾, 张先锋. 人工智能及其发展应用[J]. 理论与研究, 2012 年第 02 期.

[2]国务院. 2017 年国务院政府工作报告[R]. 第十二届全国人民代表大会第五次会议, 2017 年 3 月 5 日.

[3]国务院. 2018 年国务院政府工作报告[R]. 第十三届全国人民代表大会第一次会议, 2018 年 3 月 5 日.

[4]国务院. 2019 年国务院政府工作报告[R]. 第十三届全国人民代表大会第二次会议, 2019 年 3 月 5 日.

[5]陈劲, 吕文晶. 人工智能与新工科人才培养：重大转向[J]. 高等工程教育研究, 2017 年 06 期.

[6]金婧, 万怀宇, 林友芳. 融合实体类别信息的知识图谱表示学习方法[J]. 计算机工程, https://doi.org/10.19678/j.issn.1000-3428.0057353

[7]XIE Ruobing, LIU Zhiyuan, SUN Maosong. Representation learning of knowledge graphs with hierarchical types[C]// International Joint Conference on Artificial Intelligence. New York, NY, USA: AAAI Press, 2016: 2965 - 2971

[8]杨玉基, 许斌, 胡家威, 仝美涵, 张鹏, 郑莉. 一种准确而高效的领域知识图谱构建方法[J]. 软件学报, 2018,29(10): 2931-2947. http://www.jos.org.cn/1000-9825/5552.htm

[9]孙昊天, 杨良斌. 基于带权三元闭包的知识图谱的构建方法研究[J]. 情报杂志, 2019, 38(6): 168 - 173.

[10]董永强, 王鑫, 刘永博, 杨望. 异构 YANG 模型驱动的网络领域知识图谱构建[J]. 计算机研究与发展, 2020 年 04 期: 699 – 708.

[11]Bjorklund M. RFC 7950: The YANG 1.1 Data Modeling Language[OL]. IETF, 2016[2019-12-01]. https://tools.ietf.org/html/rfc7950

[12]熊晶,焦清局,刘运通. 基于多源异构数据的甲骨学知识图谱构建方法研究[J]. 浙江大学学报(理学版), 第 47 卷第 2 期: 131 – 150.

[13]刘燕,傅智杰,李姣,侯丽. 医学百科知识图谱构建[J]. 中华医学图书情报杂志, 2018 年 6 月, 第 27 卷第 6 期: 28 – 34.

[14]白如江, 周彦廷, 王效岳, 王志民. 科学事件知识图谱构建研究[J]. 情报理论

与实践, http://kns.cnki.net/kcms/detail/11.1762.G3.20200317.1708.008.html.

[15]https://www.ltp-cloud.com/

[16]陈成, 陈跃国, 刘宸, 吕晓彤, 杜小勇. 意图知识图谱的构建与应用[J]. 大数据, 2020 年 02 期: 57 – 68.

[17]刘泽华, 赵文琦, 张楠. 基于 Scrapy 技术的分布式爬虫的设计与优化[J]. 信息技术与信息化, 2018 年 2 - 3 期: 121 – 126.

[18]赛金辰. 基于 Spark 的 SVM 算法优化及其应用[D]. 北京邮电大学, 2017 年 1 月.

[19]李爽. 基于 Spark 的数据处理分析系统的设计与实现[D]. 北京交通大学, 2015 年 6 月.

[20]https://github.com/fighting41love/funNLP

[21]https://github.com/ownthink/Jiagu

[22]徐增林, 盛泳潘, 贺丽荣, 王雅芳. 知识图谱技术综述[J]. 电子科技大学学报, 2016 年 7 月, 第 45 卷第 4 期: 589 – 606.

[23]刘哲宁, 朱聪慧, 郑德权, 赵铁军. 面向特定标注数据稀缺领域的命名实体识别[J]. 指挥信息系统与技术, 2019 年 10 月, 第 10 卷第 5 期: 14 – 18.

[24]MINTZ, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data[C]// Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Proceeding of the AFNLP. Stroudsburg: ACCL, 2009: 1003 – 1011.

[25] 孙启民, 胡莉丽, 黄威. 基于 SNMP&Amcharts 的性能监测技术在动环监控系统的应用[J]. 技术创新, 2016 年 02 期: 35 – 38.