“基于Spark的人工智能知识的知识图谱构建”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，以期給广大学习者尤其是本科生提供有益的学习参考。

理，

1 项目背景

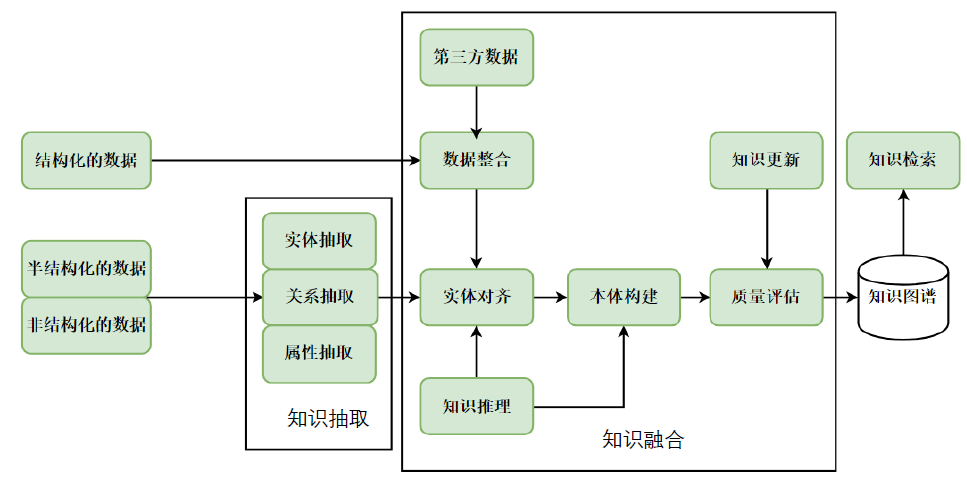
随着Web技术飞跃式发展，互联网先后经历了三个时代，它们分别具有不同的特征：文档互联的“Web 1.0”时代，数据互联为特征的“Web 2.0”时代以及当下正在发展的知识互联的崭新“Web 3.0”时代。知识互联为人们的学习与交流提供了极大便利，人类的知识交互达到了历史的新高峰。然而，互联网上的知识来源复杂、良莠不一，零散混乱、体系松散，尤其是在大数据的时代背景下，这給内容的筛选、组织与评价带来了极大挑战。知识图谱（Knowledge Graph）是人工智能（Artificial Intelligence，简称AI）领域一项重要的技术分支，具有强大的语义处理能力与开放互联能力。值得注意的是，目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。



3 项目方案

1. 既定方案

如下所述为本项目的原定技术方案。

知识图谱构建流程大致如下：

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；

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χ²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。

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。

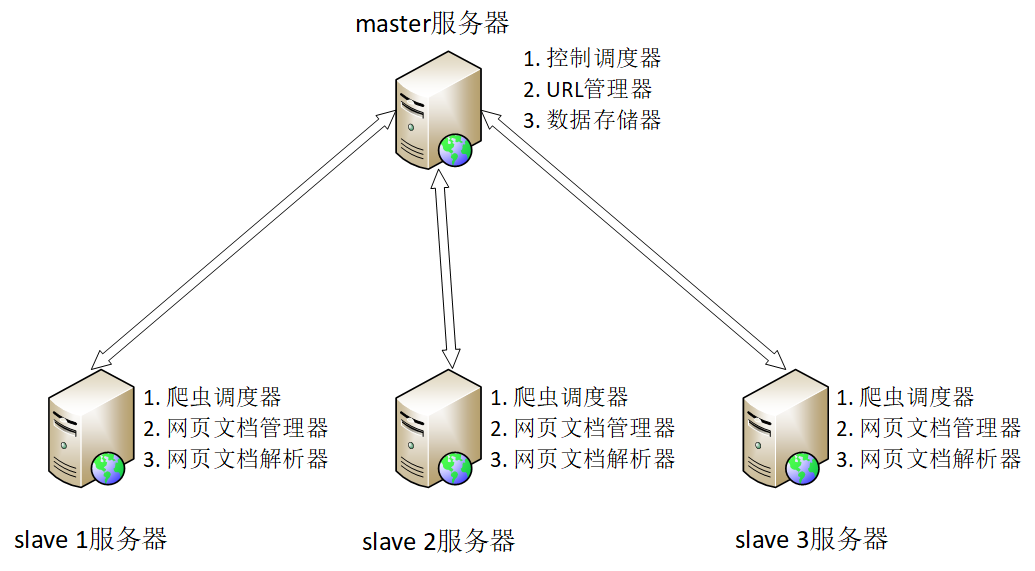
创建完成：

至此，知识图谱的创建工作基本结束。

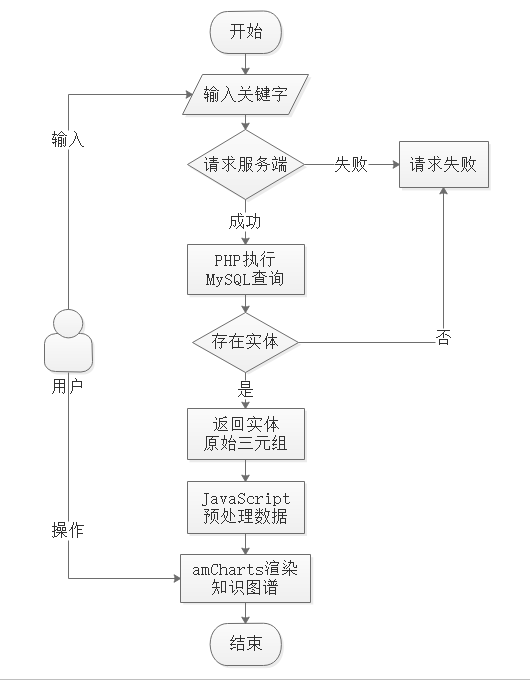
1. 实际方案

原定方案在实践中遇到诸多困难，结合实践我们有所修改。

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。



1. 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。
2. 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。



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。

|  |  |
| --- | --- |
| 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

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。

1. 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。

3）经费使用

本项目原申请了经费8000元人民币，实际使用了XXXX元，未超过申请的项目经费额度。其中XXXX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；XXXX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

本项目的难点主要有如下几个。

（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。

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。

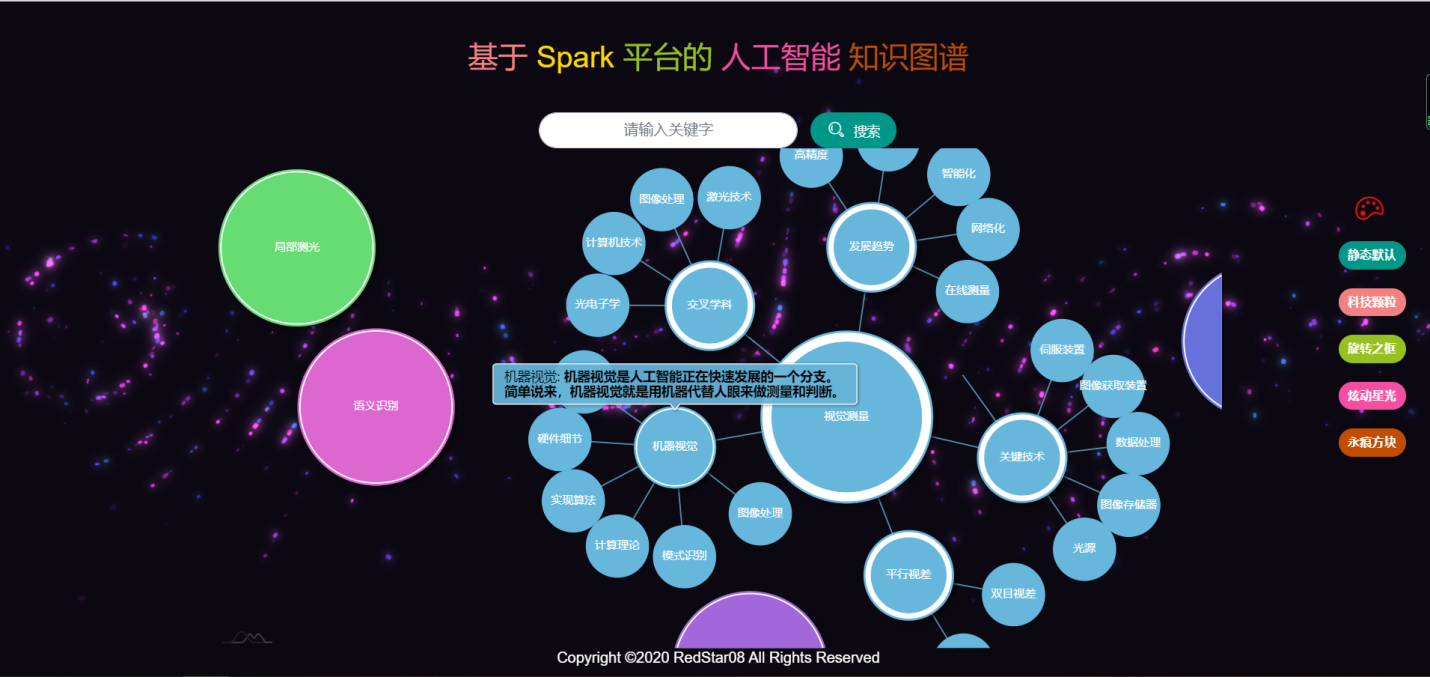
必须清醒地认识到，知识关系的抽取是一个十分复杂的工作，而且由于汉语作为一门分析语所具有的固有特征，对汉语文本进行知识抽取较之英语更加复杂，个人短时间内独立开发一套工具进行知识抽取很难，因此借助开源工具是很有必要的。目前开源的许多工具都是以英语作为语料和语义环境开发出来的，对汉语的兼容性较差，经过权衡我们最终选择了国产的开源自然语言处理工具Jiagu。目前Jiagu的作者还在维护这个项目，可以预期今后的版本的功能将更加强大。

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。





1. 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；
2. 本项目在2019年iCAN比赛中荣获安徽省二等奖。

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