合肥工业大学

项目名称: 基于Spark平台的人工智能知识的知识图谱构建

负 责 人: 文华

所属院校: 合肥工业大学宣城校区

所属院系: 计算机与信息系

指导教师: 周波

联系电话: 18856302551

2019年9月

“基于 Spark 平台的人工智能知识的知识图谱构建”

项目说明书

目 录

1. 数据获取 3
2. 数据预处理 3
3. 模型构建 4
4. 数据处理 5
5. 可视化 UI 设计 6
6. 映射函数设计 7
7. 图谱呈现 7
8. 项目创新 9
9. 商业价值 10
1） 应用场景 10
① 图谱问答（语音助手、智能电视） 10
② 学习工具（知识分析、计算、推理） 10
③ 商用知识图谱（公安、金融、旅游等行业） 10
2） 经济前景 10
① 图谱问答 10
② 学习工具 11
③ 商用知识图谱 11
参考文献 11

1. 数据获取

本项目拟构建人工智能知识的知识图谱（Knowledge Map）[1]，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（Distributed Crawl）获取内容是唯一有效的方法。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。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。因此，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，我们拟借助分布式爬虫实现 PathFinder 算法[2]，根据相关度阈值获取内容。相关度高的爬取内容有助于达到知识图谱构建的预期目的：知识面向给定领域、高度组织、系统化。

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。

数据获取结束后开始展开下一步工作。

2. 数据预处理

爬取的数据杂源异质，一般存在大量的重复与无效内容，此时需要对数据进行清洗，以便后续操作。当数据量较大时，不论是人工亦或传统的单机串行处理，都会面临极大的挑战，复杂度与耗时极其可观。为此，我们借助于 Spark 框架优秀的并行化处理能力，并应用其完备的 API 对文本进行去重并过滤无意义的数据。Spark 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[3]。

在完成数据的预处理后，接下来的工作是考虑模型的构建。

3. 模型构建

模型构建是构建知识图谱的关键一环，之后的工作都将围绕它展开，因此，我们在此有必要对整个模型构建的脉络进行必要的描述。

知识图谱在 2005 年由陈悦等率先在中国引入并命名 ，是人工智能领域一项重要的技术分支，它通过结合自然语言处理、机器学习、数据统计分析和大数据处理等技术来构建一个高效可靠的知识仓库，并以图的形式展示各个知识模块节点的关联性，从而帮助使用者更好的获取知识并从知识的联系中得到有价值的信息[1]。值得一提的是，目前知识图谱的构建并没有成熟的理论基础与技术路线，且开源的知识图谱工具大都缺乏实用化的条件。综述，基于我们对知识图谱的认知，对模型构建进行大胆尝试是不可避免的。

我们拟定的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[4]。因此，利用 Word 分词组件[5]完成分词，将所得数据各个术语提取出来，并用 word2vec[6]对术语构造向量;

b.关系抽取：将关系抽取看作是一个分类问题，采用深度学习的方法对 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;

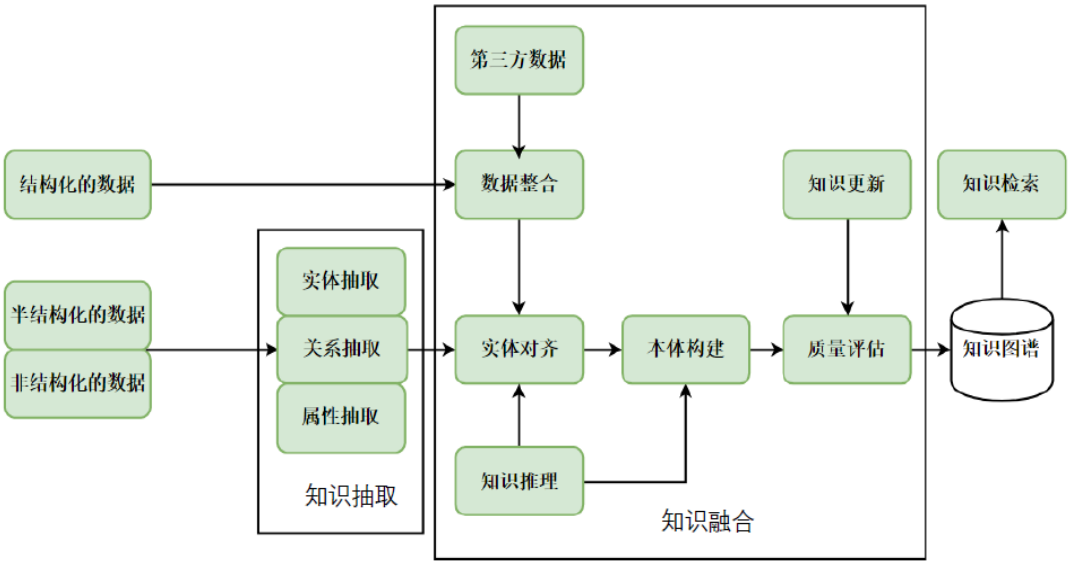
c.本体构建：本体是用于描述一个领域的术语集合，本项目的本体构建暨术语的提取与分类在知识抽取阶段已完成;

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。

图 1 知识图谱构建流程



在概述模型构建的步骤之后，下面介绍数据处理。

4. 数据处理

本阶段所面对的主要任务是“三、模型构建”中所提到的知识抽取工作，实体抽取已在“二、数据预处理”中基本完成，本节重点讨论的是关系抽取与属性抽取。

关系抽取一般是指从一个语句中判断两个实体是否有关系，是个二分类的问题，指定某种关系。同时，关系分类一般是指判断一个语句中两个实体是何种关系，属于多分类的问题[7]。我们将属性的判别视作分类问题，因此不对属性抽取另作描述。

深度学习方法在文本分类方面达到的效果与取得的成就是其他方法所无法比拟的，尤其是近年来神经网络的流行与发展为领域提供了大量高效的创新理论

和方法[7]。结合项目实际需求，我们采用卷积神经网络（Convolutional Neural Network，CNN）来做关系抽取，并将其与 Spark 平台有机结合，充分利用 Spark 并行处理的优势提高架构流畅度。借助开源深度学习工具 TensorFlow，我们可以方便地实现 CNN 来完成所需任务，目前这方面已经有了许多成熟的可借鉴的工作，文本分类指标的准确率、召回率与 F1 值较传统机器学习方法都得到了极大提高[8]。

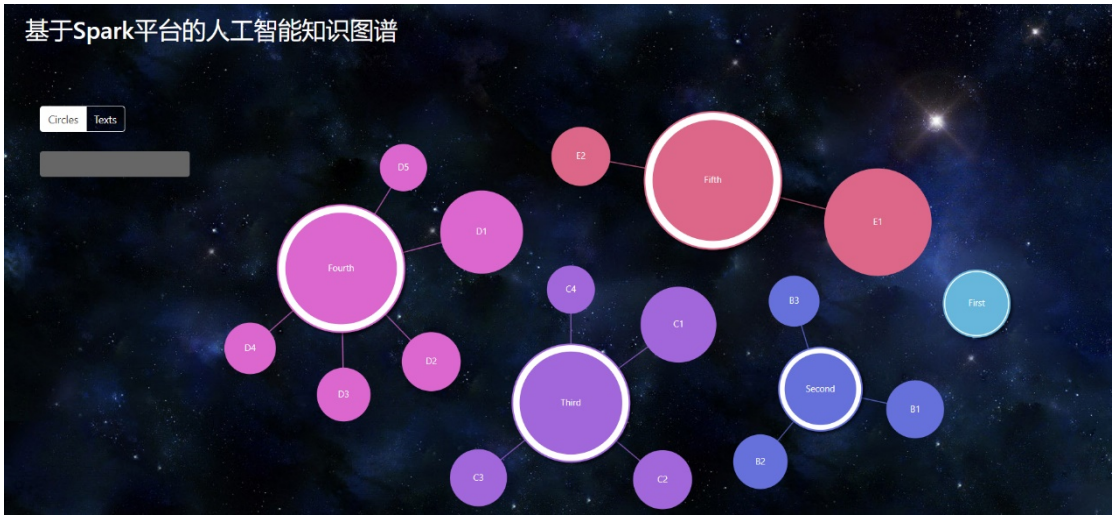
构建知识图谱所需的数据准备工作至此基本完成，此后的工作将着眼于知识图谱的展示与优化。

5. 可视化 UI 设计

可视化有许多优秀工具可供选择，如：OpenGL、VTK、plotly、R ggplot2 等，为了降低实现的复杂度并切合项目实际需求，我们选用 JavaScript 的 D3 即所谓的 D3.JS 作为本项目可视化工具。D3 的全称是 Data-Driven Documents，是一个被数据驱动的文档，它是 JavaScript 的函数库，主要用于数据的可视化。D3 功能强大，操作简单，工作量小。

综合考虑数据的嵌入与图谱良好的可视化展示，我们采用单中心向外延伸的模式，效果如图 2 所示。

图 2 单中心向外延伸模式图



6. 映射函数设计

用户在使用知识图谱工具时，必将频繁地查询关键字，有鉴于此，我们选取 Trie 树作为可视化映射的数据结构。将数据处理得到的源数据存储于数据库中。先建立存储术语的表，每一行存储 20 个相关度高的术语，同时每一行中将与其他关联词相关度最高的术语放在第一列。然后再建立多张表，每张表存储与 20 个术语所对应的源数据，表名以相关度最高的术语命名。

当得到用户输入的关键词后，利用 Trie 树将关键词与术语表中的第一列字符进行比较查询。得到相关度最高的一行术语，再根据术语名找到对应的表，对源数据进行展示。

除此之外，还需将用户输入的关键词，加入查询得到的那一行术语中。定期将术语表根据每行术语个数进行排序，以方便将热门搜索内容尽快展示给用户。

Trie 树是哈希树的变种。典型应用是用于统计，排序和保存大量的字符串(但不仅限于字符串)，所以经常被搜索引擎系统用于文本词频统计。它的优点是：利用字符串的公共前缀来减少查询时间，最大限度地减少无谓的字符串比较，查询效率比哈希树高。

7. 图谱呈现

目前本项目进展为成功构建了知识图谱模型并开始进行知识抽取，但囿于客观因素整个项目的工作仍在开展之中。为了提前展示项目预期成果，我们采用人工干预的方式对部分数据进行了图谱构建，如图 3 至图 6 所示。我们有理由相信，项目的最终成果将不亚于图示效果。

图 3 “视觉测量”知识图谱效果图

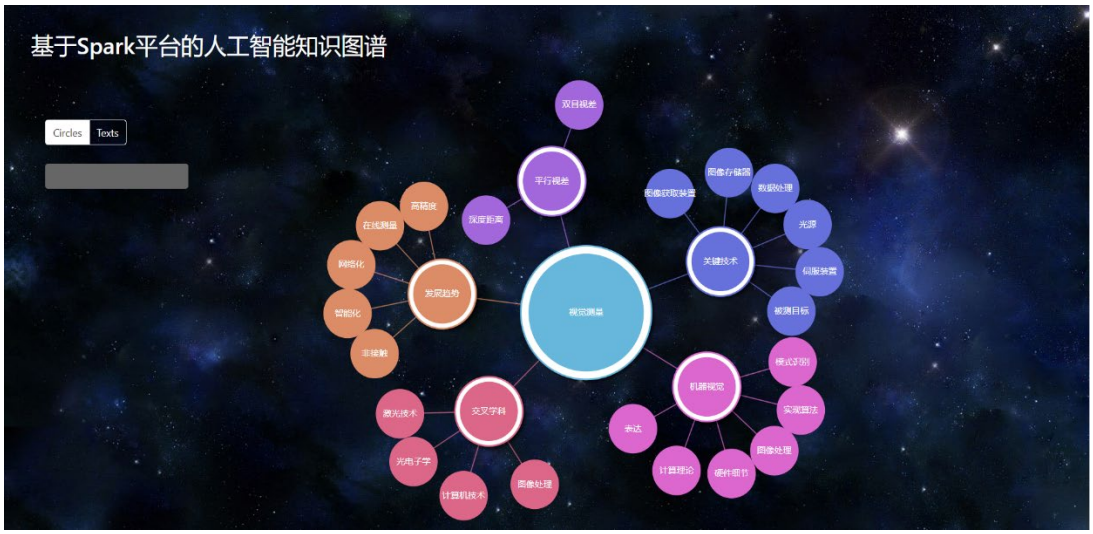


图 4 “感知器”知识图谱效果图

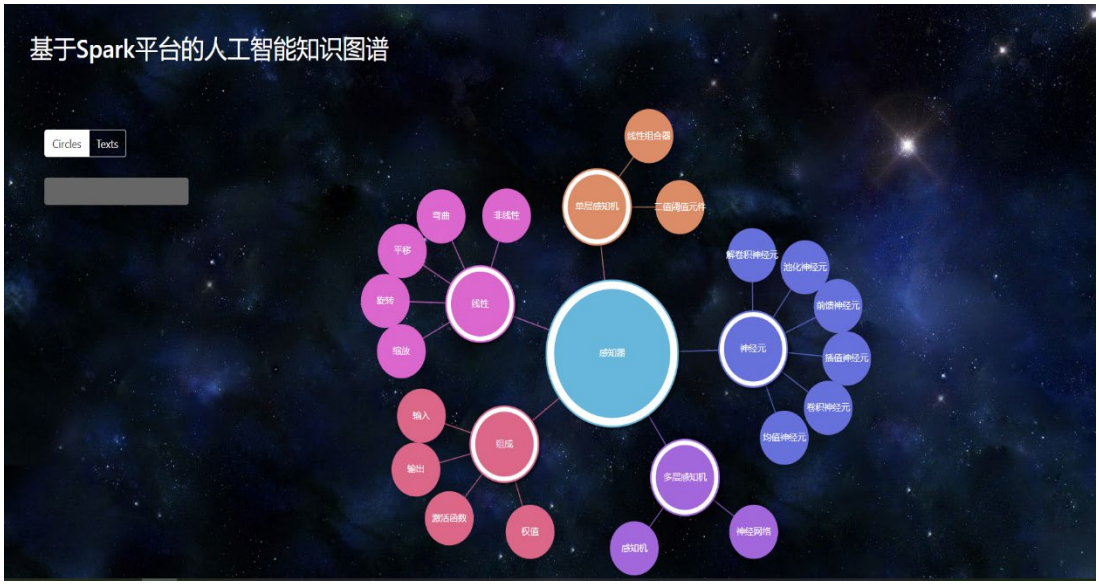
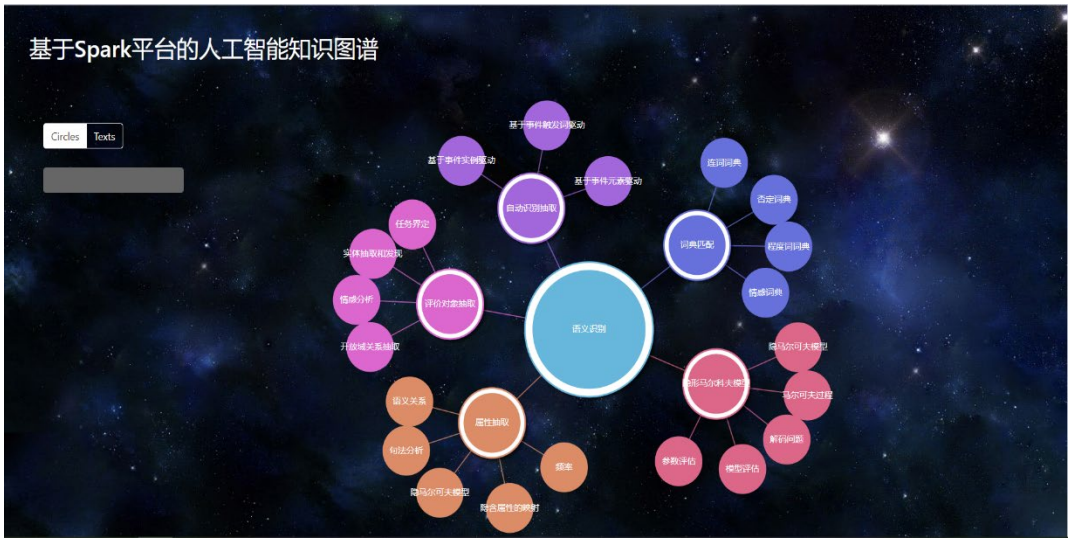


图 5 “梯度消失”知识图谱效果图

图 6 “语义识别”知识图谱效果图



8. 项目创新

1) Spark 结合 CNN 构建人工智能知识图谱；
2) 实现 PathFinder 算法限制阈值爬取数据；
3) 实现了非传统思维导图的可视化效果。

9. 商业价值

1）应用场景

① 图谱问答（语音助手、智能电视）

针对用户提出的问题，对关键词进行知识图谱构建，并对数据进行可视化展示。显然，其直观形象、易于理解。

② 学习工具（知识分析、计算、推理）

作为一款学习工具，对用户所需的知识点进行知识图谱构建，帮助用户分析、计算、推理一些复杂的数据，从而帮助用户理解对应知识，相对于传统的课本优势在于简单，易懂。

③ 商用知识图谱（公安、金融、旅游等行业）

借助Spark处理大数据的优势，可以迁移本项目的技术路线，譬如：对一些数据量较大的或者复杂的数据构建知识图谱，帮助各个行业分析、预测以及总结所需要的数据，节省数据分析时间，提高各个行业工作效率，帮助其发现并及时解决问题，调整策略。

2）经济前景

① 图谱问答

现在几乎人手一部智能手机，家家户户有智能电视。如果将此套构建图谱的技术，应用于智能手机、智能电视等领域，不但市场广大，而且能将相关图谱直观的展示给用户，让其体验到知识图谱不一样的乐趣。

② 学习工具

对于学习工具，目前市场上充斥着大量形色各异的产品。将此套技术应用于学习行业，可以针对孩子启蒙教育的学习、中小学生知识的学习、成人工作培训的学习、老人生活中知识盲点的学习等等，设计适用于不同年龄层次的人群。应用于学校、家庭、教育机构、培训中心等等，市场前景广阔。

③ 商用知识图谱

此套构建知识图谱的技术可以在金融、公安、旅游等行业进行投资。如金融行业的经济关系图、经济效益图；公安系统中人物人际关系图谱；旅游行业人流量、消费量、热门地区等重要指标的图谱。都可以帮助各个行业提高工作效率，预测并及时提出下一步的方案。

参考文献

[1]陈悦,刘则渊.悄然兴起的科学知识图谱[J].科学学研究,2005,Vol.23(2).

[2]汤天波.Pathfinder 算法优化研究[J].计算机应用与软件,2015,Vol.32(11).

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

[4]马力,李沙沙.基于词向量的文本分类研究[J].计算机与数字工程,2019,Vol.47(2).

[5]ysc.Word 分词文档.https://github.com/ysc/word.2018.9.28.

[6]张敬谊,张亚红,李静.基于词向量特征的文本分类模型研究[J].行业应用,2017(5).

[7]C´ıcero Nogueira dos Santos,Bing Xiang,Bowen Zhou.Classifying Relations by Ranking with Convolutional Neural Networks[J].Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 626–634, 2015.

[8]LinlinWang,Zhu Cao,Gerard de Melo, Zhiyuan Liu. Relation Classification via Multi-Level Attention CNNs[J].Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1298–1307,2016.