“互联网+”大学生创新创业大赛

项目计划书

项 目 名 称 AI 云学习 —— 一款基于 Spark 构建知识图谱的人工智能学习工具

项 目 类 型 “互联网+”信息技术服务业

项目负责 人 文华

申 报 日 期 2020 年 7 月 19 日

目 录

1 项目概述

1.1 研发背景

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。构建知识图谱的一般技术流程如图 1.1.1 所示。

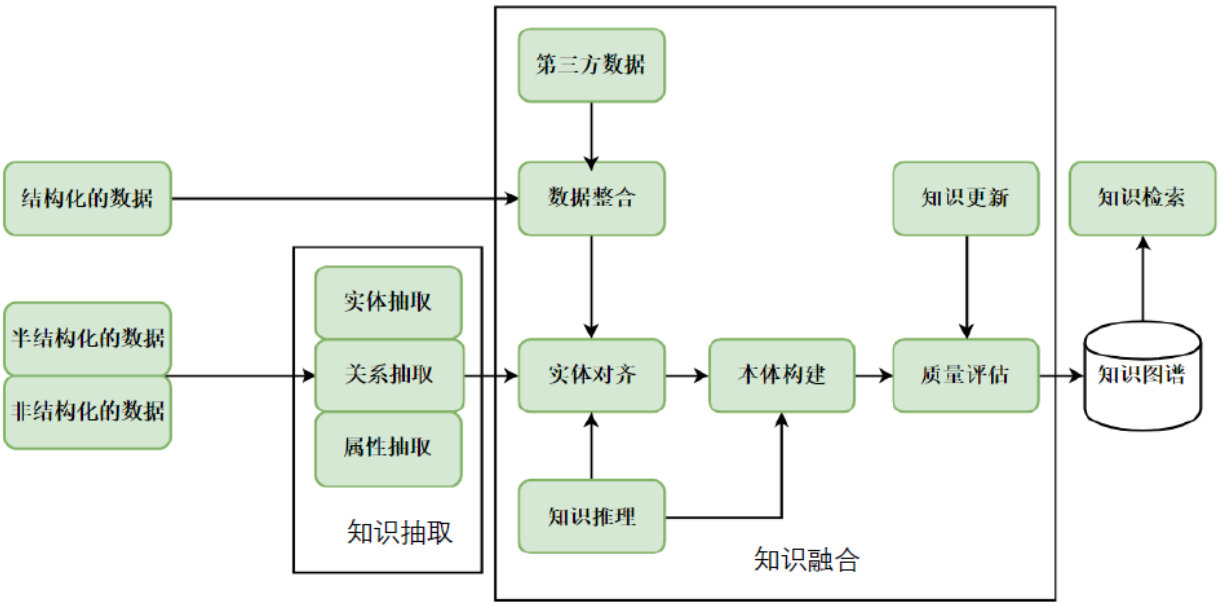

图 2.1.1.1 知识图谱构建流程

1.2 产品概况

本产品为“AI 云学习 —— 一款基于 Spark 构建知识图谱的人工智能学习工具”，其基于 Spark 大数据平台并充分利用了数据爬虫获取、实体识别、关系抽取、可视化分析等技术，构建了一个人工智能领域的垂直知识图谱，以期为知识服务系统提供知识的高效检索、组织和管理，为知识间关联关系的发现奠定基础。该图谱可提供力导向布局图作为可视化界面展示百科知识的直观方式，并且具有响应快、规模可扩展、跨平台等优点。本产品包括优化的 Python 爬虫元数据获取系统、知识图谱构建系统、手机 APP（Android 与 iOS 端）、轻量级应用服务器。用户可以通过本产品解决在特定应用场景下的知识检索问题，高效、完整、准确地学习相关知识，如：①准确、快速地检索“人工智能”相关术语并提供解释，且给出术语的联想结果，利于用户进一步学习；②突出学科在行业中的发展形势与学科热门应用领域，給学生就业、择业提供参考；③形象化地展示“人工智能”知识的脉络、历史沿革与发展趋势，为学生复习、深入学习提供参考。

目前本产品已经完成了所有的开发、调试与部署，正在通过多渠道宣传本产品，并向各方面争取投资，下一步的工作将在收集充分的用户反馈与筹集足够经费的基础上，对知识抽取算法进行优化，同时对产品服务器进行升级。本产品获得过 2019 年 iCAN 国际创新创业大赛安徽赛区的省级二等奖，以及经学校“大学生创新创业大赛”专家组审核通

过，可见本产品拥有扎实的技术积淀。

我们学思结合，敢为人先，勇于挑战，更有充足信心将所学专业知识转化为实际成果，也因此我们坚信：随着大数据与人工智能技术的蓬勃发展，传统的学习方式将被会逐渐替代甚至颠覆，取而代之的是更为现代化、效率更高、可重用性更强、传播更快的模式，而本产品的推出正顺应了这一时代背景，在行业未来的发展中必将大放异彩。

1.3 市场优势

①本团队产品顺应技术发展潮流，在同领域的产品属于首创，具有绝对的市场独占率与技术优势；②在互联网+同大数据与人工智能日新月异的时代背景下，我们团队以产品和服务为载体，技术创新与社会需要的融合臻于化境，抢占市场，获得利润；③本团体具有明晰的研发、宣传、营销目标，集中团队优势，抢占市场空缺；④本团队的产品有望成为首款面向人工智能学习者的知识图谱辅助学习工具；⑤本团队产品已经获得两项省级奖项，得到了审评人员的高度认可，具有巨大的发展空间。

1.4 市场预期

2020 年是知识图谱行业发展过程中非常关键的一年，首先，从外部宏观环境来讲，转变经济增长方式，严格的节能减排对知识图谱行业的发展都产生了深刻的影响。知识图谱行业需求持续火热，资本利好知识图谱领域，行业发展长期向好。2019 年居民人均可支配收入 28228 元，同比实际增长 6.5%，居民消费水平的提高为知识图谱行业市场需求提供经济基础。传统知识图谱行业市场门槛低、缺乏统一行业标准，服务过程没有专业的监督等问题影响行业发展。互联网与知识图谱的结合，缩减中间环节，为用户提供高性价比的服务。90 后、00 后等各类人群，逐步成为知识图谱行业的消费主力。通过对市场环境的分析，结合产品本身特征和目标市场定位，我们估计本团队产品在同行业产品中相对垄断，市场地位和市场份额可达 50%以上。本团队将在提供整体解决方案的基础上，通过先进的技术和完善的服务提高用户对产品的认可度，培养客户粘性。

1.5 销售预期

第一年：团队产品运营初期，预计将会服务用户 1000000 人次。全年实现毛利润 50

万元，力争实现净利润 32 万元。

第二年：团队产品运营初期，预计将会服务用户 2000000 人次。全年实现毛利润 90 万元，力争实现净利润 80 万元。

第三年：团队产品更新换代，服务优化，预计将会服务用户 4000000 人次。全年实现毛利润 130 万元，全年力争实现净利润 115 万元。

1.6 融资方式

本团队运营资金来源方式主要为：创业贷款。

2　产品服务与创意

2.1　产品项目介绍

2.1.1　产品名称

AI 云学习 —— 一款基于 Spark 构建知识图谱的人工智能学习工具。

2.1.2　产品 logo

图 2.1.2.1 产品 logo

2.1.3　产品研发团队

牛头冲八仙下海创业团队。

2.1.4 产品系统组成

（1）基于 PathFinder 算法的主从分布式 Python 爬虫元数据获取子系统。

（2）基于 Spark 平台的元数据预处理子系统。

（3）基于 Jiagu 模型的知识关系抽取子系统。

（4）基于 PHP 与 MySQL 的关键词检索子系统。

（5）基于 amChart 4 的图谱渲染与展示子系统。

（6）云端服务器。

（7）Web 应用。

（8）手机 APP。

2.1.5 产品功能说明

（1）对用户输入的不在数据库中的关键词进行预检索处理，即以当前关键词作为主从分布式 Python 爬虫元数据获取子系统的输入来获取相应的元数据。

（2）对分布式 Python 爬虫元数据获取子系统所得到的元数据进行文档去重、清洗、网页标签过滤、敏感词筛除与文本自组织标记。

（3）对 Spark 平台的元数据预处理子系统所得到的预处理数据进行自然语言模型训练并提取相应的知识关系。

（4）对 Jiagu 模型的知识关系抽取子系统所生成的三元组数据进行格式重调、MySQL 存储并给出用户使用与自定义知识图谱所需的“增删查改”功能。

（5）对 PHP 与 MySQL 的关键词检索子系统所返回的 json 格式数据进行力导向图渲染与展示。

（6）云服务器是部署知识图谱后端的主要平台，负责对用户数据的检索、元数据获取、文本预处理、知识关系抽取与知识图谱展示等一系列功能。

（7）Web 应用是供 PC 端与手机端用户实时检索所需知识图谱的前端平台，免去了安装专门应用的烦琐操作。

（8）手机 APP 分为 Android 与 iOS 版本，分别供 Android 用户和 iOS 用户安装使用，手机 APP 增强用户使用知识图谱的稳定性与安全性。

2.1.6 产品的技术领先性

（1）产品核心技术：

◆借助主从分布式 Python 爬虫实现 PathFinder 算法。

◆基于大数据处理平台 Spark 的文本预处理系统。

◆基于国产开源自然语言工具 Jiagu 实现高效、快捷的知识关系抽取。

◆基于数据仓库平台 hive 实现微秒级的数据库管理操作。

◆基于 amChart 4 完成艺术级的图谱渲染效果与知识节点展示。

◆云服务器实现了对用户输入数据的全自动元数据流式获取、文本预处理、知识抽取、数据库存储与图谱节点反馈。

◆Android、iOS 与 Web 应用提供了多种知识图谱访问操作。

（2）产品技术、应用与运营模式创新：

①技术创新

a. 借助主从分布式 Python 爬虫实现 PathFinder 算法

本项目拟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实现 PathFinder 算法，根据相关度阈值获取内容。

理论计算与实验数据证明，本项目采用的 Python 爬虫方法在显著地提高了数据获取效率的同时，还极大地保证了数据的相关度。

b. 基于大数据处理平台 Spark 的文本预处理系统

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。

c. 基于数据仓库平台 hive 实现微秒级的数据库管理操作

hive 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。

②应用创新

a. 基于国产开源自然语言工具 Jiagu 实现高效、快捷的知识关系抽取

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。

b. 基于 amChart 4 完成艺术级的图谱渲染效果与知识节点展示

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。

c. 云服务器实现了对用户输入数据的全自动元数据流式获取、文本预处理、知识抽取、数据库存储与图谱节点反馈

为了提高产品的可用性，本项目所设计的知识图谱除了提供对本地存储的知识节点查询外，还能以用户输入的关键词进行图谱拓展，概而言之就是：当输入关键词不匹配数据库内的任何结果时，将其作为 Python 爬虫的输入关键字爬取相关文本，并将所得文本按照既定的技术流程操作，得到与新关键词有关的知识图谱。这一方式使知识图谱的进一步拓展成为了可能。

③模式创新

a. Android、iOS 与 Web 应用提供了多种知识图谱访问操作

本产品提供了多种操作终端，最大化地覆盖了各个平台的用户，以期为产品盈利带来更为广阔的使用人群，这增大了产品的被动测试与 BUG 反馈案例，为后期产品优化提供了绝佳的参考。

b. 产品提供免费与付费双重个性化服务

本产品面向广大用户提供日均一定数量的免费知识图谱检索服务的同时，设置了付费检索服务：付费用户凭支付一定量的费用享受次数更多、自定义操作更完善的知识节点检索服务。付费服务是本产品盈利的重要来源。

2.2 产品系统总体技术方案

如 1.1 节所述，知识图谱构建主要分为三个步骤：知识抽取、知识融合与知识检索，下面就每个方面进行详细介绍。

（一）数据类型

构建知识图谱的元数据有三种常见的类型：结构化数据、半结构化数据与非结构化数据。

结构化的数据是指可以使用关系型数据库表示和存储，表现为二维形式的数据。一般特点是：数据以行为单位，一行数据表示一个实体的信息，每一行数据的属性是相同的。常见的结构化数据为数据库。

半结构化数据是结构化数据的一种形式，它并不符合关系型数据库或其他数据表的形式关联起来的数据模型结构，但包含相关标记，用来分隔语义元素以及对记录和字段进行分层。因此，它也被称为自描述的结构。对于半结构化数据，属于同一类实体可以有不同的属性，即使他们被组合在一起，这些属性的顺序并不重要。常见的半结构数据有 XML 和 JSON 格式。

非结构化数据是没有固定结构的数据。各种文档、图片、视频/音频等都属于非结构化数据。对于这类数据，我们一般直接整体进行存储，而且一般存储为二进制的数据格式。

（二）知识抽取

知识抽取是指把蕴含于信息源中的知识经过识别、理解、筛选、归纳等过程抽取出来，存储形成知识元库。知识抽取是构建知识图谱的首个关键步骤与基础，直接影响了后续工作的成效与最终构建所得图谱的质量。知识图谱构建中知识抽取分为：实体抽取、关系抽取与属性抽取。

实体抽取又称命名实体识别，包括实体的检测（find）：识别命名实体的文本范围，实体的分类（classify）：分类为预定义的类别，学术上所涉及一般包含三大类，实体类、时间类、数字类和 7 个小类，如人、地名、时间、组织、日期、货币、百分比等。

关系抽取主要负责从文本中识别出实体，抽取实体间的语义关系，在知识图谱构建中一般以三元组的形式来表征。

属性抽取的任务为识别实体的属性名与识别实体的属性值，而属性值结构一般是不确定的。

（三）知识融合

知识融合，即合并两个知识图谱(实体及其对应关系)，其基本问题是研究怎样将来自多个来源的关于同一个实体或概念的描述信息融合起来。由于知识图谱中的知识来源广泛，存在知识质量良莠不齐、来自不同数据源的知识重复、知识间的关联不够明确等问题，所以需要进行知识的融合。知识融合是高层次的知识组织，使来自不同的知识源的知识在同一框架规范下进行异构数据整合、消歧、加工、推理验证、更新等步骤，达到数据、信息、方法、经验以及人的思想的融合，形成高质量的知识库。

经过上述步骤后，方能得到可供进行知识检索的有效知识图谱。接下来详细介绍本产品构建知识图谱的技术流程。

2.2.1 数据来源

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。

a. 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）[1]，并在其上实现 PathFinder 算法，据所列关键词的

相关度按阈值排序获取特定内容。

本项目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 2.2.1.1 所示，主从分布式爬虫的逻辑结构如图 2.2.1.1 所示，爬虫的类图结构如图 2.2.1.2 所示。

表 2.2.1.1 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

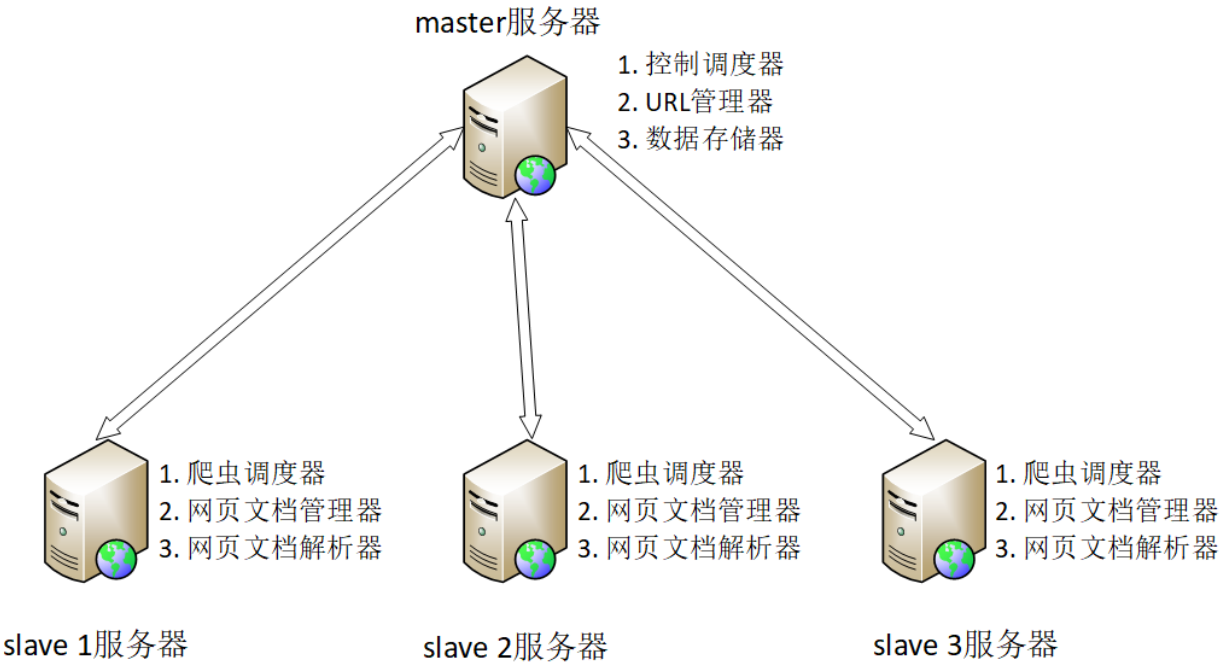

图 2.2.1.1 主从式分布爬虫逻辑结构

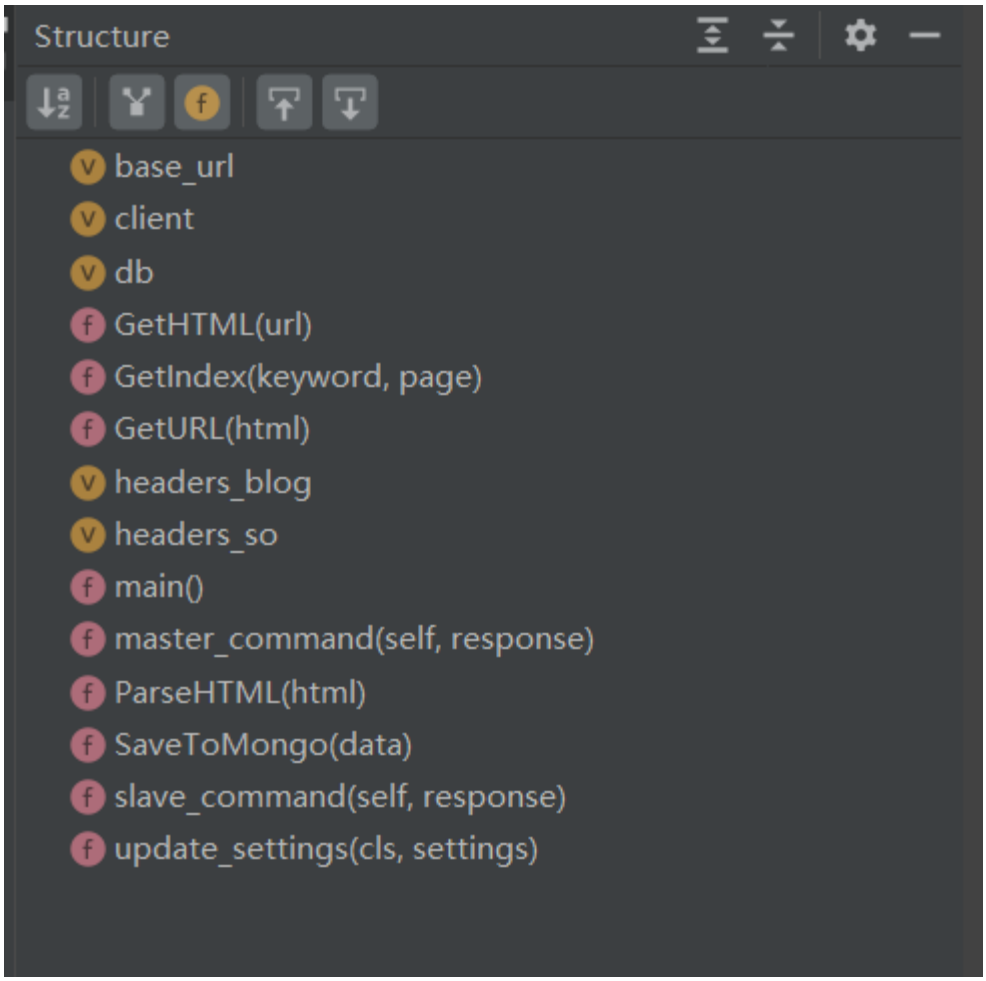

图 2.2.1.2 爬虫程序的类图结构

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。

2.2.2 Spark 与 Jiagu 模型

①Spark 与 hive 平台

Spark[2]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[3]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，本项目采用 MySQL。

本项目将使用 Spark 平台的相关工具进行数据预处理。

②数据预处理

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[4]，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 2.2.2.1，数据预处理的程序类图如图 2.2.2.2 所示，预处理后的部分数据如图 2.2.2.3 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 2.2.2.1 spark-shell 处理命令

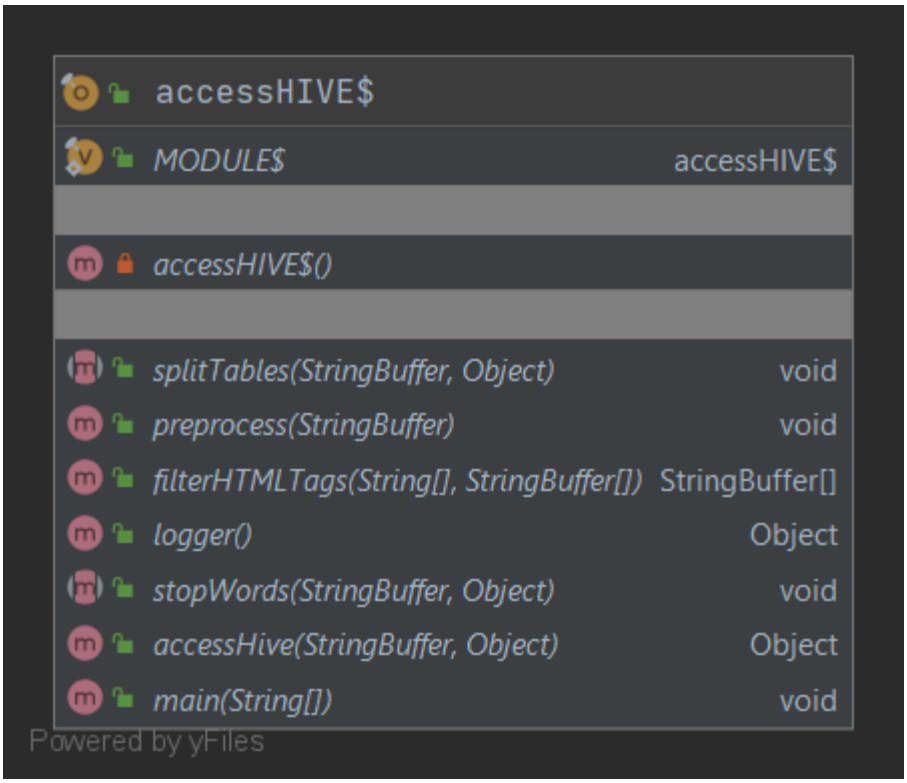

图 2.2.2.2 数据预处理程序的类图

```
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 2.2.2.3 预处理后的部分数据

③Jiagu 模型

Jiagu 模型[5]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。

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[6]。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[7]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 2.2.2.1 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 2.2.2.2 所示。实验结果用 held-out 方法[8]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 2.2.2.4 所示，所得部分三元组如图 2.2.2.5 所示。

表 2.2.2.1 数据集的统计信息

数据集	关系数量	语料行数
训练集	10	2435796
验证集		634547
测试集		1849620

表 2.2.2.2 所用训练参数

Learning rate	Iterations	Damping coefficient
0.001	50000	0.85

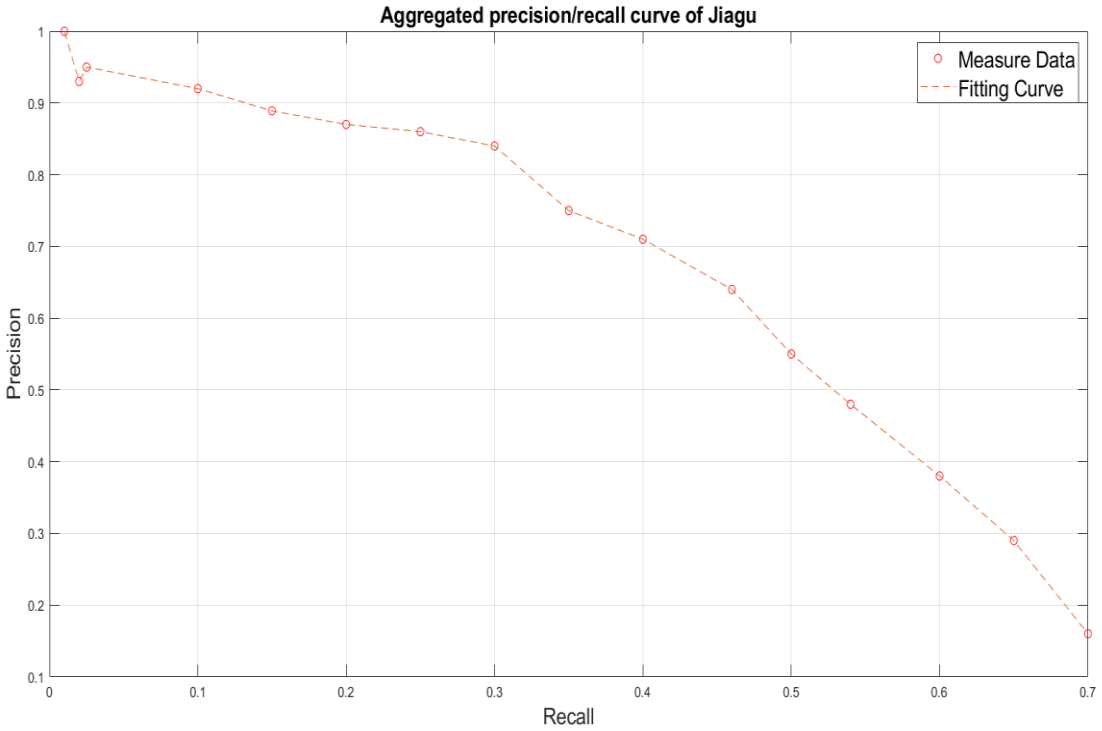

图 2.2.2.4 准确率/召回率

```
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 2.2.2.5 三元组数据

2.2.3 知识图谱的可视化

①三元组的转化

本项目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[9]。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 2.2.3 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 2.2.3.1 所示。

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。

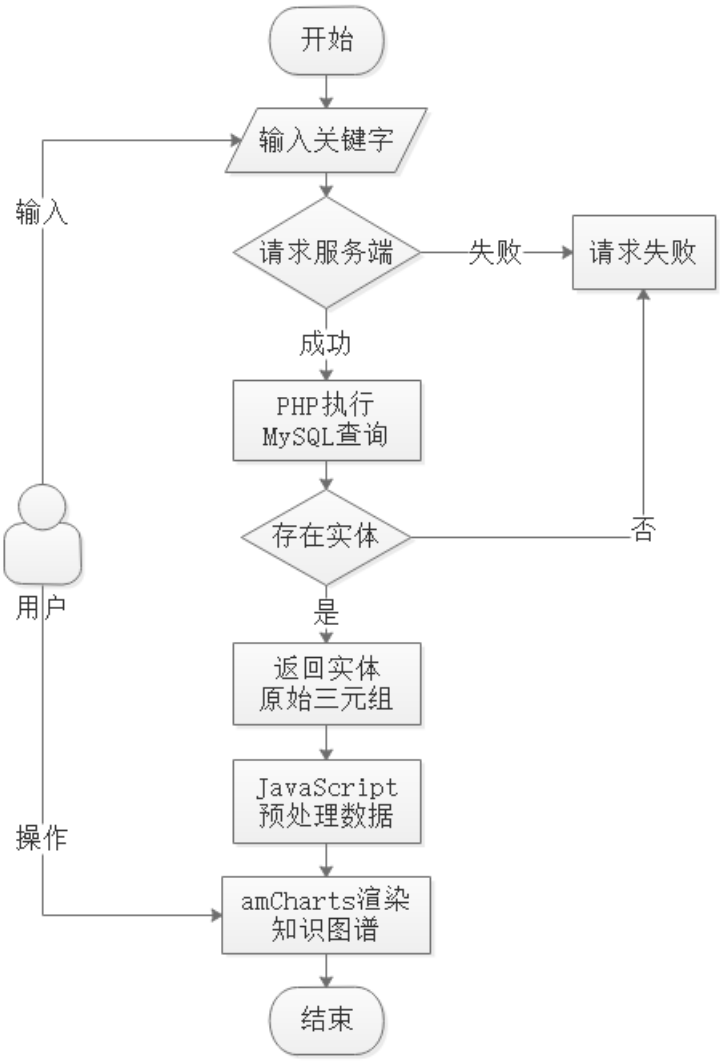

图 2.2.3.1 知识图谱可视化流程图

2.3 产品硬件配置、软件截图与说明

2.3.1 产品硬件配置

云端服务器配置信息，即产品硬件配置信息如图 2.3.1.1 所示。

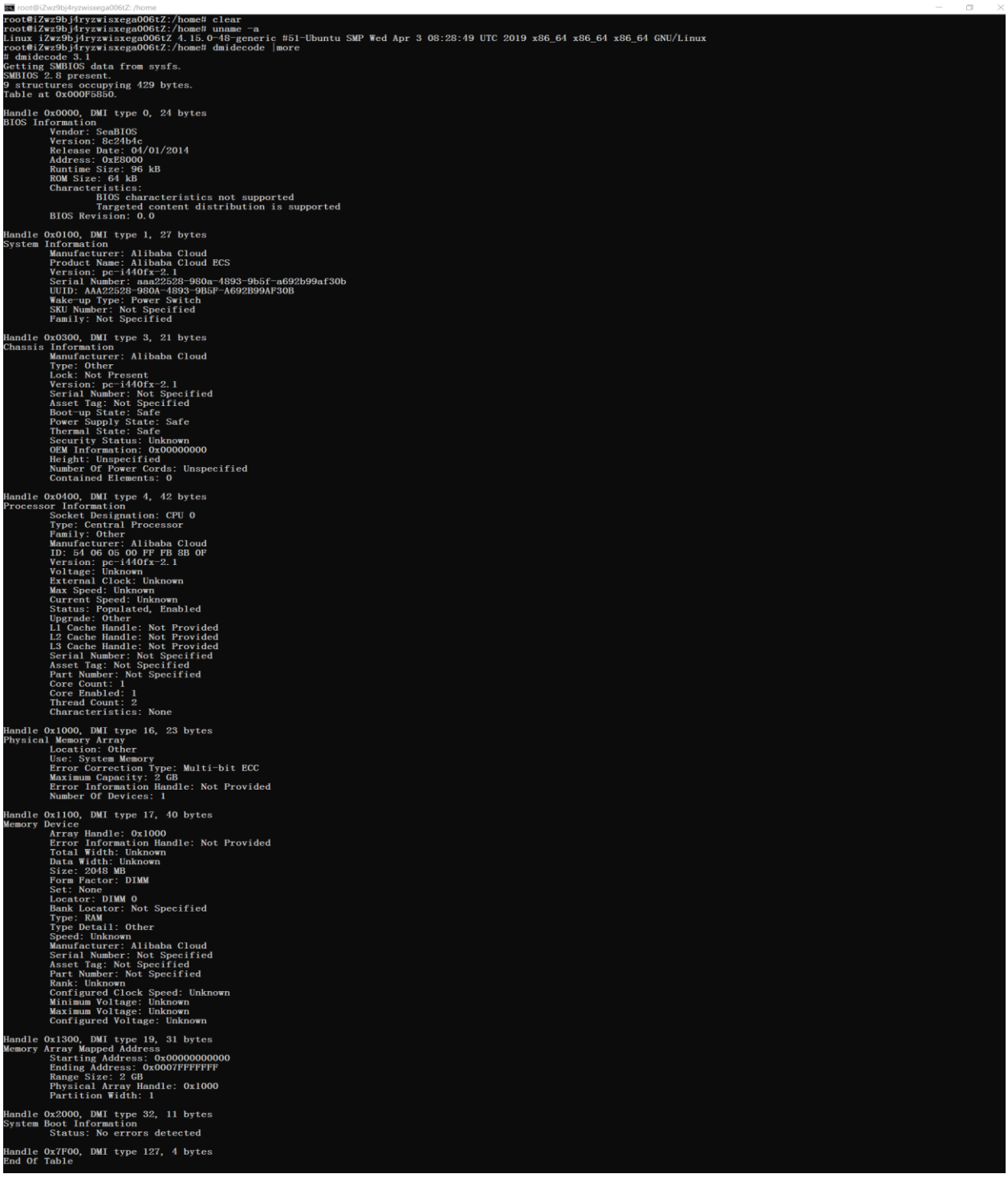

图 2.3.1.1 云端服务器硬件配置信息

2.3.2 软件截图

Web 应用运行状况，如图 2.3.2.1 至图 2.3.2.6 所示。

图 2.3.2.1 运行截图-1（检索关键词：机器人）

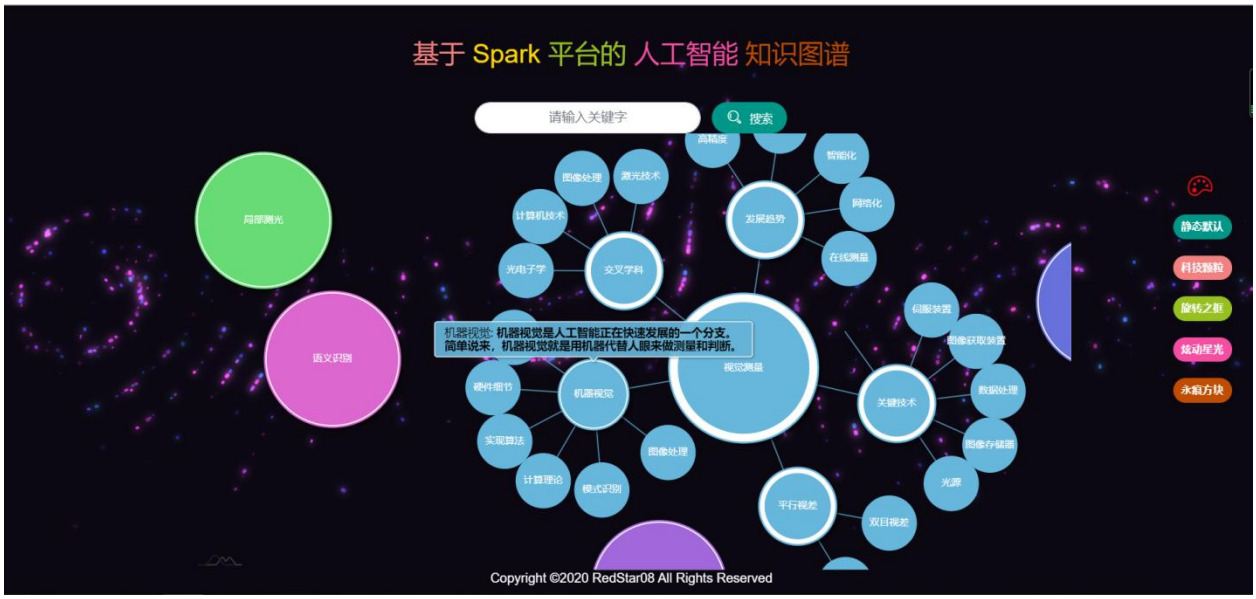

图 2.3.2.2 运行截图-2（检索关键词：视觉测量）

图 2.3.2.3 运行截图-3（检索关键词：人工智能）

图 2.3.2.4 运行截图-4（检索关键词：AI 开发）

图 2.3.2.5 运行截图-5（检索关键词：k 近邻算法）

图 2.3.2.6 运行截图-6（检索关键词：NLP 技术）

本项目所采用的图谱可视化工具支持多种主题背景的选择，如图 2.3.2.7 至图 2.3.2.10 所示。

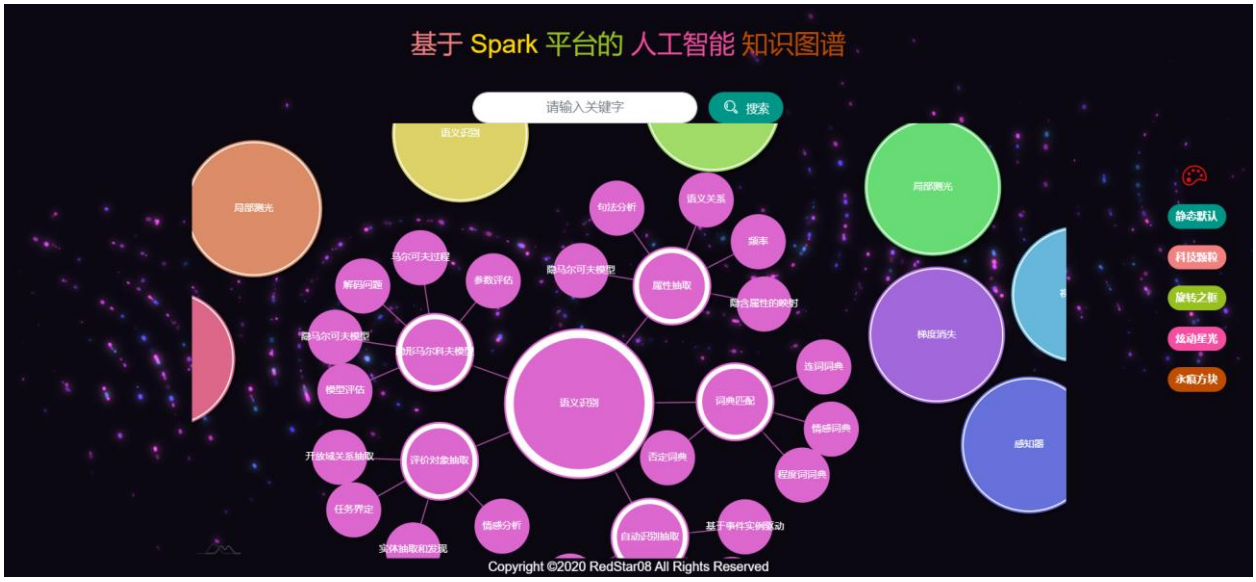

图 2.3.2.7 主题 2“科技颗粒”

图 2.3.2.8 主题 3“旋转之框”

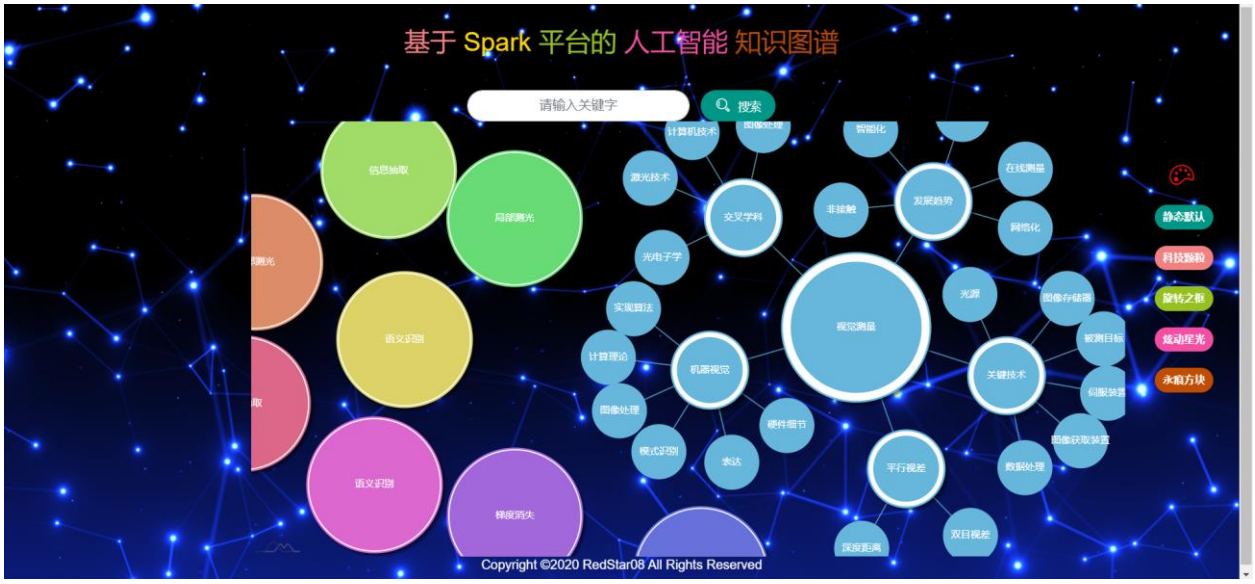

图 2.3.2.9 主题 4“炫动星光”

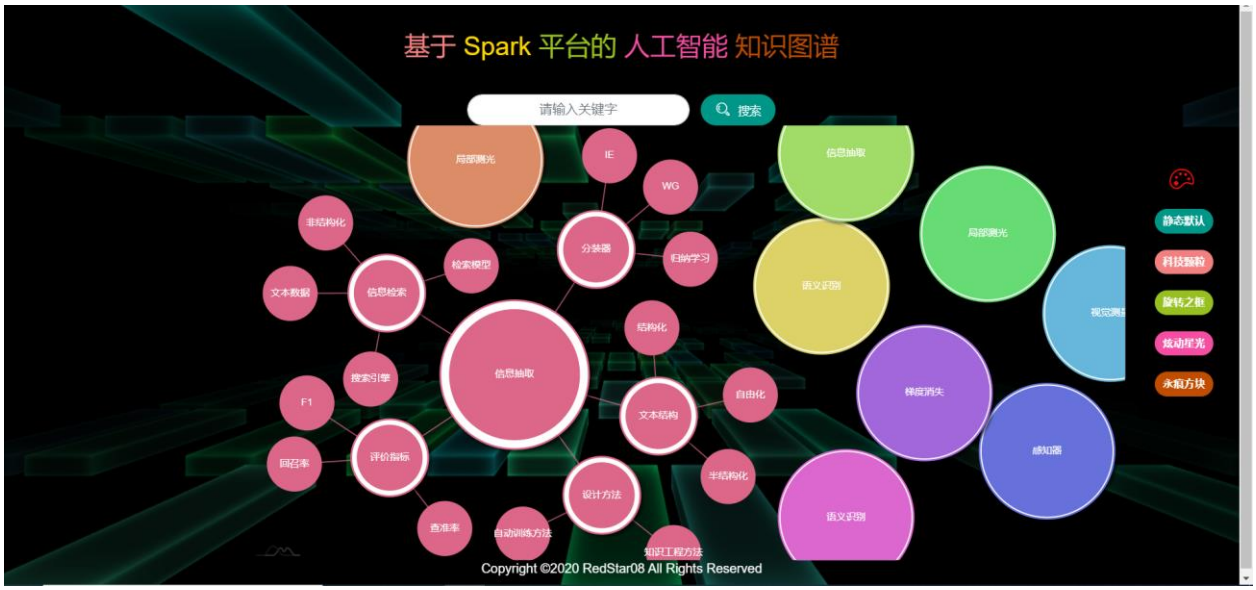

图 2.3.2.10 主题 5“永痕方块”

本产品的 Android 与 iOS 端应用正在紧急开发中，不日即可上线服务。

2.4 产品升级计划

1、知识图谱的升级

（1）目前知识图谱本地存储的知识节点数量较小，在未来的上线服务与用户反馈后，将逐步增大知识节点的数量，扩展图谱规模。

（2）目前采用 Jiagu 自然语言处理工具所提供的知识关系抽取功能需要提供大量的人工标记数据进行模型训练，人工标记数据耗费大量的人力与时间，在下一步的研究中将会尝试使用远程监督模型对原始数据进行标记，减少人力成本的同时，提高了工作效率。

2、云端服务器的升级

（1）当前云端服务器是租用阿里云的轻量服务器，性能一般，将来随着用户的增多与产品盈利，将会改换为性能更优越的服务器，按需增加服务器数量。

（2）用户日均访问量增长的同时会带来巨大的流量消耗，届时将采用分布式云服务器处理框架，并对每台服务器负载均衡技术，减轻单台服务器的计算压力。

3、客户端产品的升级

当前的知识图谱工具只能从 Web 端访问，随着 Android 与 iOS 端应用开发完成，本产品将如期向所有平台的用户提供全方位的知识图谱检索服务。

3 市场分析

3.1 市场环境分析

知识图谱是一种重要的知识表示形式,能够打破不同应用场景下的数据隔离，通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论与计量学引文分析、共现分析等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构从而达到多学科融合目的的现代理论。知识图谱能为学科研究提供切实的、有价值的参考，因此其发展备受重视，行业前景较为广阔。

3.1.1 知识图谱行业 PEST 分析

1. 市场因素

2020 年是知识图谱行业发展过程中非常关键的一年，首先，从外部宏观环境来讲，经济增长方式的转变，严格的节能减排对知识图谱行业的发展都产生了深刻的影响，另外还有来自通货膨胀、人民币升值、人力资源成本上升等等因素的影响；从企业内部来讲，产业链各环节竞争、技术工艺升级、出口市场逐步萎缩、产品销售市场日益复杂等问题，都是企业决策者所必须面对和亟待解决的。

2. 经济因素

知识图谱行业需求持续火热，资本利好知识图谱领域，行业发展长期向好。

今后五年经济社会发展的主要目标是：经济保持中高速增长，到 2020 年国内生产总值和城乡居民人均收入比 2019 年翻一番，主要经济指标平衡协调，发展质量和效益明显提高；创新驱动发展成效显著；发展协调性明显增强；人民生活水平和质量普遍提高；国民素质和社会文明程度显著提高；生态环境质量总体改善；各方面制度更加成熟更加定型。

我国知识图谱行业如何透视现状、锚定未来、战略前瞻、科学规划，寻求技术突破、产业创新、经济发展，为引领下一轮发展打下坚实的基础。

下游行业交易规模增长，为知识图谱行业提供新的发展动力。

2019 年居民人均可支配收入 28228 元，同比实际增长 6.5%，居民消费水平的提高为知识图谱行业市场需求提供经济基础。

3. 社会因素

进入互联网时代，应用的特点发生了变化，大部分都是大规模开放性应用。同时大数据时代也给新时期知识库技术的发展带来了机遇。在大数据时代，我们拥有了前所未有的算力和数据，有着花样繁多的模型，大规模的众包平台，以及高质量的用户内容，这使得自动化知识获取、知识图谱构建自动化成为可能。

4. 技术因素

科技赋能 VR、大数据、云计算、知识图谱、5G 等逐步从一线城市过渡到 2、3、4 线城市，实现知识图谱行业科技体验的普及化。

知识图谱行业引入 ERP、OA、EAP 等系统，优化信息化管理施工环节，提高了行业效率。

3.1.2 知识图谱行业发展现状分析

知识图谱市场热度高涨，其应用市场得到跨越式发展的根本原因在于技术、安全、品种的革新。用户需求的爆发式增长极大丰富了知识图谱的应用市场。

一方面，知识图谱的产业链中原料和供应商的进一步推动，有利于产业源端的重组升级，优化产业流程；另一方面知识图谱技术、品质、品种的更新迭代，有利于产品的不断升级和质量改进，进一步满足用户的新需求，这些都有利于产业进一步发展。多方的推动使得知识图谱应用将在未来 5 年得到爆发式发展。良好的社会环境也为本团队发展提供了非常肥沃的土壤条件

3.1.3 行业规模分析

据协会统计，2019 年我国知识图谱产销较快增长，产销总量再创历史新高，比上年同期分别增长 14.5%和 13.7%，高于上年同期 11.2 和 9.0 个百分点。12 月产销比上月分别增长 1.7%和 4.0%，比上年同期分别增长 15.0%和 9.5%。[10]

1. 产品销售同比增长 14.9%

2019 年，产销比上年同期分别增长 15.5%和 14.9%，增速高于总体 1.0 和 1.2 个百分点，其快速增长对于产销增长贡献度分别达到 92.3%和 94.1%。其中，同比增长 3.4%；12 月产销量比上月分别增长 0.2%和 3.2%；与上年同期相比，产销量分别增长 13.6%和 9.1%，产销同比均呈较快增长。

2. 销售同比增长 5.8%

2019 年，同比分别增长了 8.0%和 5.8%，增幅进一步提高；同比增长 11.2%和 8.8%，12 月环比增长 12.4%，同比增长 25.1%；环比增长 10.5%，同比增长 12.1%。

3. 产品销售同比增长 53.0%

2019 年比上年同期分别增长 51.7%和 53.0%。比上年同期分别增长 63.9%和 65.1%；比上年同期分别增长 15.7%和 17.1%。

另据艾瑞咨询统计推算，2019 年涵盖大数据分析预测、领域知识图谱及 NLP 应用的大数据智能市场规模约为 106.6 亿元，预计 2023 年将突破 300 亿元，年复合增长率为 30.8%，其中 2019 年市场中以金融领域和公安领域应用份额占比最大[11]。其市场规模发展趋势如图 3.1.3.1 所示。

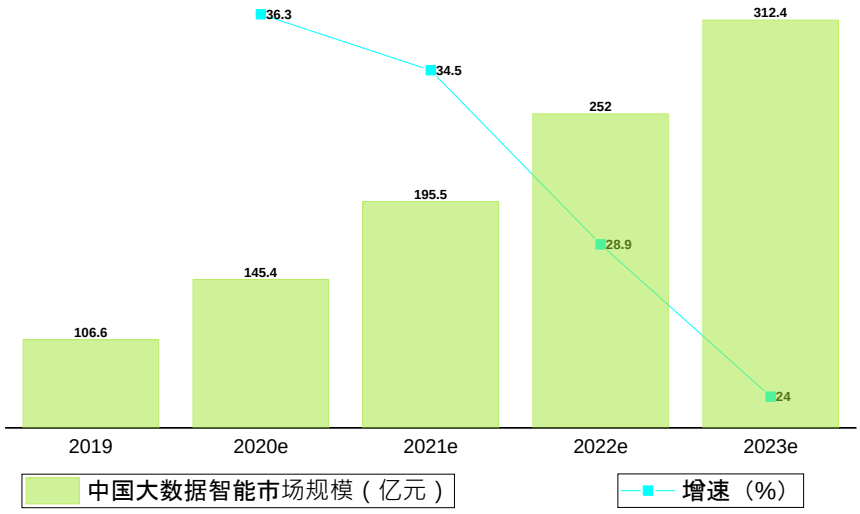

图 3.1.3.1 中国大数据只能市场规模（来源艾瑞咨询研究院）

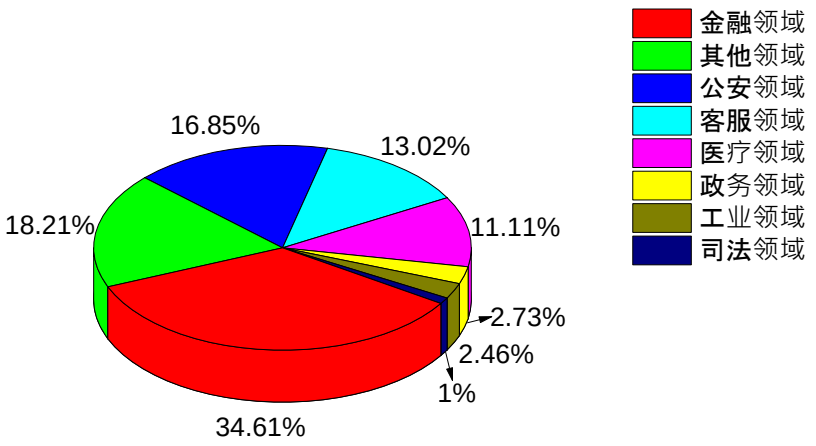

图 3.1.3.2 市场应用结构（来源艾瑞咨询研究院）

3.1.4 中国对知识图谱行业政策分析

在 2019 年工信部曾发文明确指出，2020 年将围绕工业大数据融合应用、民生大数据创新应用、大数据关键技术先导应用、大数据管理能力提升 4 大类 7 个细分方向着重发展，而知识图谱作为集大数据和人工智能于一身的综合技术，也将成为重点关注领域。

由中国电子技术标准化研究院联合数家企业与高校联合编写的《知识图谱标准化白皮书》（2019 版）也已发布。白皮书从哲学层面、政策层面、产业层面、行业层面、技术层面、工具层面、支撑技术等多个层面对知识图谱的实际需求、关键技术、面临的问题与挑战、标准化需求、展望与建议等进行了梳理，涉及智慧金融、智慧医疗、智能制造、智慧教育、智慧政务、智慧司法、智慧交通等十五个领域，并初步提出了知识图谱技术架构和标准体系框架等，以期对未来知识图谱在更多行业的推广应用及标准研制提供支撑。[12]

并且由中国电子技术标准化研究院提报的国家标准《信息技术人工智能知识图谱技术框架》（计划号：20192137-T-469）、IEEE 标准《知识图谱架构》（项目编号：P2807）和《知识图谱技术要求与评估规范》（项目编号：P2807.1）均已获批立项。

种种政策表明我国正在努力推进知识图谱建设，为知识图谱行业的飞速发展保驾护航

知识图谱行业国内外对比分析如表 3.1.4.1 所示。

表 3.1.4.1 知识图谱行业对比

	国外知识图谱	国内知识图谱
价值定位	聚焦于发掘早期初创型企业并助力其成长，快速提升其商业价值	
	促进创客文化形成，获得高技术商业回报	响应宏观及产业政策号召，吸引资源导入
价值创造	通过提供服务、资本增值、社会回报获得有形及无形价值	
	技术交易、股权价值回报为主	增值服务、资金补贴为主
价值实现	寻找合理的商业逻辑与实现渠道，获得企业成长与收益获得的双赢	
	股份转让、IPO 等获取收益	政府补贴、税收分成、培训…….
价值传递	形成品牌效应，吸引更多优质企业和初创团队，扩散传播价值	
	理念宣传、技术交流	政府站台、双创活动、人脉推广…….

3.2 目标市场定位

我们团队以“更好地为学生尤其是大学生学习提供便利”为短暂目标，希望能够使用知识图谱这种方式来帮助同学们更好地学习，这样能让原本繁杂的知识能够清晰明了地展现出来，并以此为申引，在一些其他方向发展，将产品融入更多其他行业，从而满足目标消费者的需求。

当前知识图谱更多的是应用在电商、图情（图情知识图谱是指聚焦某一特定细分行业，以整合行业内资源为目标的知识 图谱。提供知识搜索、知识标引、决策支持等形态的知识应用，服务于行业内的 从业人员，科研机构及行业决策者）、企业商业及创投（创业投资）等方面，而用于学习和通用行业的知识图谱大多不存在或者不够成熟。目前市场上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；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。

因此，本团队将主要目标分成两级，目前主要发展对象是学生群体，再逐渐转向普通企业用户或个体用户，致力于将知识图谱应用融入生活，在这个大数据互联网时代给用户带来更好的生活体验和更加便捷的用户体验。由此引出几个可发展方向，具体如下：

1. 知识检索关联

知识检索依托创投知识图谱，可以在原有知识全文搜索的基础上实现语义搜索并引出相关信息及其他关键词应用形态。其中，语义搜索提供自然语言式的搜索方式，由机器完成用户搜索意图识别。例如，如果搜索“人工智能”这个词语，节点展开后能够显示其基本解释并引出其他相关知识，如“语言识别”,“图像识别”等。这一功能对学生学习有较大帮助，能够让学生在巨大的知识库中最快速找到所需知识，并完成对相关知识的学习，有助于更全面，更快捷地进行学习。

2. 金融：识别及预防欺诈

反欺诈在金融风控中举足轻重，但基于大数据的反欺诈存在两个难点：一是如何整合不同来源的结构化和非结构化数据，并有效地识别出身份造假、团体欺诈、代办包装等欺诈案件。二是不少欺诈案件涉及复杂的关系网络，如组团欺诈。知识图谱是基于关系的表达方式，可轻松解决以上两个问题，因此在反欺诈中获得广泛应用。首先，知识图谱可以提供非常便捷的方式来添加新的数据源。其次，知识图谱本身是直观的关系表达方式，可以帮助更有效地分析复杂关系中存在的特定的潜在风险。

3. 农业：多媒体知识指导

大量的农业资料以不同格式分散存储，传统的关系数据库模式不适用于复杂多变的领域，无法实现定义所有可能的知识点并构建关键数据库模式，而知识图谱这种更加灵活的知识表示模型可以实现管理。利用抽取挖掘技术从各种多源异构数据中获取相应的知识，并用统一图谱进行表示，形成完整的知识库，刻画作物知识、土壤知识、肥料知识、疾病知识和天气知识等。

4. 智能分析

由于缺乏诸如知识图谱此类背景知识，各类工具理解大数据的手段有限，限制了基于大数据的精准与精细分析，大大降低了大数据的潜在价值。因此尽管越来越多的行业或者企业积累了规模可观的数据，但这些数据非但未能创造价值，甚至可能因消耗大量的运

维成本而成为负资产。

知识图谱的发展提供了强大的背景知识支撑，可以赋能舆情分析、商业洞察、军事情报分析和商业情报分析此类基于大数据的精准分析。

知识图谱和基于此的认知智能为精细分析提供了可能。如汽车制造厂商等制造企业都希望实现个性化制造运用于精细分析案例。知识图谱构建关于汽车评价的背景知识，如汽车的车型、车饰、动力、能耗等，提取消费者对汽车的褒贬态度、消费者改进建议、竞争品牌等评价与反馈，并以此为据实现按需与个性化定制。

知识图谱应用方面，未来将会出现更多应用形态，随着知识表示技术和推理技术的发展，结合一些新型的可视化方法，我们还可以展望一些预测分析类的应用形态，如疾病预测、行情预测、政治意识形态检测、城市人流动线分析。除此之外，知识图谱在辅助多媒体数据处理方面也是一个有待深入研究的方向，如物体检测、图像理解等。本团队的发展方向也会随着时代发展而不断向前，将知识图谱在越来越多的领域找到能够真正落地的应用场景，在各行各业中解放生产力，助力业务转型。

3.3 市场容量估算与预测

随着我国城市化进程的加快，社会稳定和城市安全等问题逐渐显现，知识图谱技术是实现基础建设的关键技术。因此，随着社会经济及信息技术的进一步发展，知识图谱的应用将是未来的一个新趋势。

我国知识图谱行业市场规模前景预测：

知识图谱技术在人们日常生活、工作中的应用越来越广泛。随着我国社会经济脚步的不断加快，对于知识图谱的应用需求也将越来越大。

随着中国新兴市场的据起，消费量急剧上升，中国知识图谱市场已经成为各大国际巨头势在必夺的重要市场。同时，随着发达国家生产成本的居高不下，国际大型制造商为了保持竞争力，降低生产成本，纷纷将生产制造基地转移至中国、印度等具有较强需求潜力的发展中国家。

知识图谱采购的本土化，将为中国知识图谱企业带来发展机遇。项目的发展具有一定程度的地域性和传承关系。随着中国知识图谱市场的发展，合资品牌的逐渐增多，多样化的技术路线也随之引入中国市场。

相关行业专家表示，在很长一段时间中，中国的技术路线不会统一，而是会呈现百

家争鸣的发展态势。无论是哪一种类型的变速器，发展的核心都是基于对能源方面的考虑，追求低碳、高效、低成本，这三大特点是技术发展的动力源泉。

随着我国消费升级，消费者的偏好也在发生转变，年轻化，智能化等消费趋势让越来越多的消费者开始青睐。根据 2018 年的消费者趋势调查显示，72%的消费者倾向于在未来选购。

根据测算，需求方面，未来五年，细分市场年均增速可达 25-30%，远超行业平均 56%的水平。产能供应方面，各大主流供应商纷纷扩张产能，产能增幅较快。即便如此，未来五年，旺盛的需求依然会持续领先行业的供给水平。

综合以上分析知识图谱行业的市场需求、现状、规模、前景预测等行业调研。根据知识图谱行业以往投资回报率，结合行业的近几年的复合增长率分析，未来几年的知识图谱产业行业投资预期客观，预期将会达到 120%以上。

4 现状与规划

4.1 人工智能发展现状

人工智能最早能够追溯到 1936 年，英国数学家 AM. Turing 在论文《理想计算机》中提出了图灵机模型，然后 1956 年在《计算机能思维吗》一文中提出机器能够思维的论述（图灵实验）.之后计算机的发明和信息论的出现为人工智能发展奠定了良好的基础.1956 年在达特茅斯会议上，Marvin Minsky、John Mccarthy 等科学家围绕“机器模仿人类的学习以及其他方面变得智能"展开讨论，并明确提出了“人工智能”一词。

人工智能的发展经历了 2 次发展热潮。第 1 次是 1956-—1966 年，1956 年，Newe 和 Simon 在定理证明工作中首先取得突破，开启了以计算机程序来模拟人类思维的道路；1960 年，McCarthy 建立了人工智能程序设计语言 LSP.上述成功使人工智能科学家们认为可以研究和总结人类思维的普遍规律并用计算机模拟它的实现，并乐观地预计可以创造一个万能的逻辑推理体系。第 2 次是 20 世纪 70 年代中期至 80 年代末，在 1977 年第五届国际人工智能联合会会议上，Feigenbaum 教授在特约文章《人工智能的艺术：知识工程课题及实例研究》中系统地阐述了专家系统的思想并提出“知识工程”的概念。至此，人工智能的研究又有新的转折点，即从获取智能的基于能力的策略变成了基于知识的方法研究。此后，人工智能的发展进入平稳发展期。

近些年，大数据时代的到来和深度学习的发展象征着人工智能的发展迎来了第 3 次发展热潮。1997 年，IBM 的深蓝（Deep blue）机器人在国际象棋比赛中战胜世界冠军卡斯帕罗夫，引发了人类对于人工智能的思考。2016 年英国初创公司 DeepMind 研发的围棋机器人 AlphaGo 通过无监督学习战胜了围棋世界冠军柯洁，让人类对人工智能的期待提升到了前所未有的高度，在它的带动下，人工智能迎来了最好的发展时代。2019 年，上海举办了世界人工智能大会，会议集聚了全球人工智能领域最具影响力的科学家和企业家以及相关政府的领导人，围绕人工智能领域的技术前沿、产业趋垫和热点问题发表演讲和进行高端对话，开启人类对于人工智能发展的新一轮探索[13]。

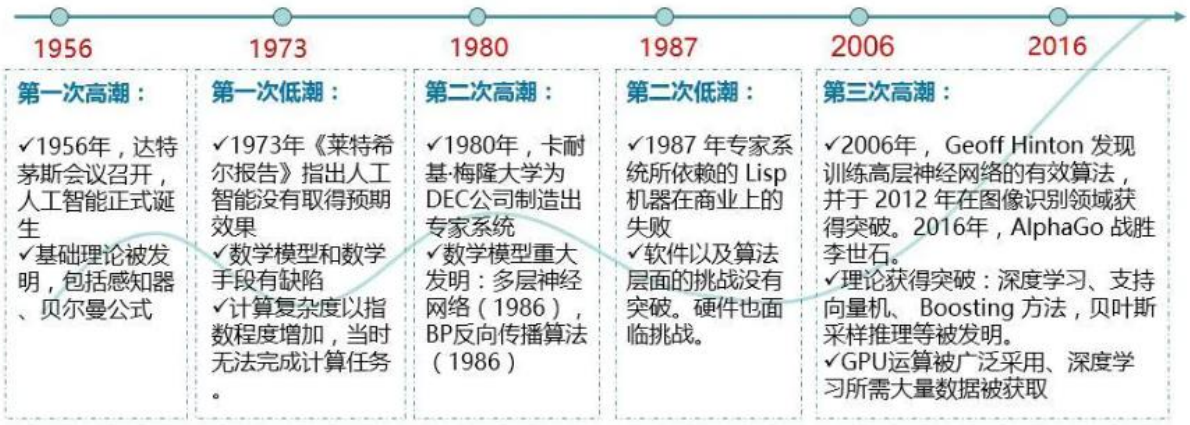

图 3.1.4.1 人工智能发展浪潮

4.2 知识图谱发展现状

知识图谱自 2012 年推出以来,进展迅速,已经成为大数据时代的重要知识表示之一,极大地推动了智能化的发展进程。目前知识图谱技术已经在大规模简单应用场景中取得了显著效果。但近年来,知识图谱的需求从数据丰富的大规模简单应用场景转向专家知识密集但数据相对稀缺的小规模复杂应用。这一转向过程给知识图谱带来了新的挑战。

4.2.1 知识图谱实现功能

首先从知识表示层来看，知识图谱的研究和落地，现在只是完成了大规模简单应用所需要的表示。知识图谱本质上是大规模语义网络。知识图谱首先是一种大规模知识表示，所以它通常包含海量的实体，往往是数以亿计。大规模也体现为多样的关系，成千上万的关系。正是因为它规模大，往往需要做出质量妥协，所以很多时候知识图谱也允许出错。现在没有人敢说自己数千万、数亿规模的知识图谱百分百正确，永远是 99.999%，允许错误。也允许 schema 不完善，从而包容更多实例，精良的模式在很多图谱里面是缺失的。语义网社区投入巨大精力推动通用 schema 的建设，但是遇到很多挑战。

它支撑的应用，大部分是简单应用：以实体（词汇）为中心的知识表示，表达的往往是实体的属性和关系；它的推理极为简单，往往都是基于路径或者上下位词的简单推理，以及基于分布式表示的推理。所以知识图谱这几年的发展，解决了大规模简单应用的场景。

其次实现了简单推理。符号知识存在的根本价值在于能做推理。当前知识图谱的大

部分推理是简单推理例如，用户搜索周杰伦，很多平台给用户推荐他的歌。这是因为知识图谱知道刘德华是歌手，因此一定会有相应歌曲。这是基于上下位关系推理。搜索唐太宗，推荐李世民，这是同义关系推理；搜索战狼 1，那么平台可能会推荐战狼 2。因为它们都是同类型的电影，并且是同一个导演、同一个主演，这是基于路径的推理。

现实中大部分应用利用这些简单推理就能解决，并且即便只用这种简单推理也能解决很多以前搜不到、问不清的痛点问题，并且效果显著。大家现在看到的很多应用场景、应用知识图谱所解决的根本问题，都是搜索、推荐和问答。

4.2.2 知识图谱瓶颈

而最近两年最大的变化就是我们面临着应用场景的变换。我们正在从大规模、简单的应用场景向小规模、复杂应用场景切换。知识图谱的前期应用场景都是以 BAT、TMD 为代表，它们属于大规模简单应用场景，模式单一，其应用的知识是众人皆知的。但是现在越来越多的是石油、能源、工业、医疗、司法、金融这种小规模复杂应用场景，它有着密集的专家知识、有限的数据资源和深度的知识应用等鲜明特性，这都是新场景给我们提出的全新挑战。这与知识图谱在互联网应用中用到的衣食住行这类通用知识显著不同。这一新的形势对于获取隐性的专家知识提出了新挑战。一方面专家知识往往是隐性的，难以直接从文本中抽取。另一方面，专家知识有着一定的门槛，只有少部分行业从业人员才能完成专家知识的众包工作。除此之外，在盘点数据的时候，会发现大部分的场景数据是稀缺的。首先领域数据本身就稀缺。其次还缺乏高质量的标注数据。我们很多机器学习模型需要标注数据，哪怕有资金可以投入人力标注，但是领域任务往往是不明确的，而专家资源又很昂贵，那么标注也会非常困难。如果不采用人工标注，而利用外界爬取的数据进行融合，也会十分困难，因为领域数据融合代价通常也非常大。所以总体上来讲，虽然很多时候我们觉得有大数据，但是相对于很多领域智能化应用而言，我们的数据还是十分“贫乏”[14]。

图 4.2.2.1 知识图谱应用场景转变

因此我们考虑到人工智能，是当前最热门研究专业领域之一，其相关方向的人才匮乏也正越来越成为（市场）关注的议题，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。而与此同时当前的知识图谱也存在着无法在专业领域得到有效应用的问题。所以团队选择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、自然语言处理与机器视觉等三个领域作为代表。

4.3 产品现状

我们目前已完成人工智能中机器学习、自然语言处理与机器视觉这三个比较热门的三个领域的知识图谱。用户可以使用我们的产品对这三个领域的相关知识进行检索，我们也会针对用户输入的关键词进行扩展，展现给用户与其输入的关键词相关联的知识。同时对用户界面进行优化，满足不同用户对知识表示方式的需求。

4.3.1 产品成本

我们的知识获取主要是基于国内的科技论坛网站，用 Python 语言编写爬虫程序进行自动化获取的。这些论坛网站的讨论基本上是与当前人工智能领域的最新发展内容息息相关的，从而可以保证用户能够得到最前沿的信息。为了能够获取大量的知识并进行相关的存储同时要保证产品的反应速度，我们需要对电脑进行不同程度的升级，但这些花销即可

获取大量的数据。相对而言，成本是非常低的。

4.3.2 产品功能

首先我们利用知识图谱使得大规模自动化知识获取基本可行。针对人工智能这一领域，我们基本实现了从数据获取->知识抽取->知识融合三个环节的自动完成。

其次我们利用知识图谱完成了许多元数据之间的关联。比如，搜索人工智能时，其往往可以表示为 AI，这样一种关联就可以告诉我们这两个字段是可以匹配的，而关联就能创造价值。所以，我们利用知识图谱作为数据融合的指引，当在搜索框内输入关键词并点击搜索后。主光圈即为输入的关键词，而周围的光圈即为其关联得性质与详细信息。因此学习人工智能的学生通过一个关键词就可以了解到多方面的知识。

同时我们利用知识图谱解决了语言表达鸿沟问题。很多时候用户所提供的搜索关键词与我们提前存在数据库里的词汇表达是有一定的差异的，特别是对于初学者。另外不同专业的人在对人工智能中同一件事情的描述所使用的语言极有可能是不一样的。而与此同时有些实体本身就有若干种说法。我们通过建设大量词汇知识图谱，包含领域的同义词、缩略词、上下位词等关系，有效解决语言表达鸿沟的问题。

相较于传统的以简单的知识应用与常识为基础的知识图谱，我们实现了能应用于专业领域，方便学习的知识图谱。现在越来越多的高校开设人工智能专业，同时国家也在这一领域投入大量资金。根据教育部在 2020 年 2 月份公布的 2019 年度普通高等学校本科专业备案和审批结果，据统计中国人民大学、北京化工大学、北京邮电大学、北京师范大学、中国传媒大学、复旦大学等 180 所高校新增人工智能本科专业。这是人工智能（AI）本科专业被纳入我国本科专业的第二年，去年仅有 35 所高校获批，今年这一数量涨势迅猛，超过去年的 5 倍。人工智能的热潮越来越高，而且人工智能方面的人才也非常的少，所以这是很多高校开设人工智能专业的原因。我们当前的产品可以供学子们进行人工智能相关内容的学习，也能够根据学子们的搜索关键字频率，将当前最热门的内容展现给他们。一定程度上也有助于人工智能的推广与发展。

4.3.3 产品价值

我们的知识图谱作为一种语义网络拥有极强的表达能力和建模灵活性：首先可以对现实世界中的实体、概念、属性以及它们之间的关系进行建模；其次，知识图谱是其衍生

技术的数据交换标准，其本身是一种数据建模的“协议”，相关技术涵盖知识抽取、知识集成、知识管理和知识应用等各个环节。

同时我们的产品作为一种特殊的图数据。其中每个结点都有若干个属性和属性值，实体与实体之间的边表示的是结点之间的关系，边的指向方向表示了关系的方向。非常的直观美化，对于用户没有高的要求，使得任何人，都可以通过我们的产品查阅人工智能领域的相关资料，都可以进行相关的学习。

其次，我们的产品采用了人类容易识别的字符串来标识各元素；图数据表示作为一种通用的数据结构，可以很容易地被计算机识别和处理。产品的可扩展性良好，技术路线已经完成，针对不同的应用场景，更改数据源即可完成新的应用。

对于知识图谱如何应用于专业领域的这一问题，我们根据自己的创新性技术路线给出了回答。考虑到近两年应用场景正在逐步们从大规模、简单向小规模、复杂进行转变，我们产品的应用前景非常广阔。同时对于其它正在研究知识图谱相关内容的人员来说，我们也提供了一种新的技术路线，在一定程度上也能促进共同的进步。

最后，由于我们的产品部署在服务器上，消耗的自然资源非常少。相较于需要购买大量的书籍来掌握相关内容，我们的产品经济环保了许多。

4.4 产品规划

4.4.1 扩大应用范围

我们计划在一年内实现文档级的知识获取。考虑到人工智能发展的火热，在未来肯定会有越来越多的人工智能产品走入大家的生活中。而在现实情况中，我们买任意设备，经常会附赠一个说明手册，例如买冰箱都会有一个手册，但是手册的利用率极低，很少有家庭成员会真正的翻阅。然而当碰到问题想去查找的时候，我们也很难从手册中找到答案。更何况是人工智能领域的高科技产品。所以基于能否将这些鸡肋一般的手册全部淘汰掉，同时还能提升用户满意度的考虑。我们希望团队的人工智能知识图谱不仅仅适用于想要学习这一专业领域的人才，也能够帮助到其它人即使不太了解专业知识也能够应对这种生活中的突发情况。我们计划将手册变成知识库并存储在数据库中利用知识图谱实现知识问答。那么不仅仅是人工智能这个领域，还可以将比如冰箱的手册变成知识库进行储存，需要变换的就是数据库里的数据，整个技术路线我们已经完成。所以我们将可以为整个社

会解决手册这一巨大成本问题。实现这个目标的前提是文档级的知识获取。基于文档的信息抽取需要结合文档自身的结构，书写风格，和组织形式进行一定的迁移。业务文档结构化迫切需要从句子级别抽取发展到篇章级别抽取。

4.4.2 开发新的业务

我们计划在探究如何将知识图谱应用于专业领域获得的启发应用于平时生活中的简单场景。在两年内利用我们的知识图谱技术路线补全简单场景中缺失的因果链条（背景知识）。万事万物都处在一个复杂的因果网络中，当前的大数据多是业务结果数据，缺乏产生这些数据的背景因果。比如，数据挖掘中的经典案例尿片与啤酒，买尿片的人经常买啤酒。可是为什么会出现这种情况呢，其实如果我们能够推测男性用户为什么会同买啤酒与尿片的原因，这实际上可以帮助我们创造更大的商业价值。可能是家里有婴儿，而孕妇出行不便，因此必须得由作为父亲的他来买尿布，同时这几天由于工作他非常紧张与疲惫，所以买一点啤酒顺便缓解一下压力。如果我们能里用知识图谱把这个因果链条给补全，当男性用户再次买尿片时，我们推断他压力大，因而给他推荐心理咨询服务。由此得到启发我们可以推荐很多新的业务。又再次的扩大了整个产品的经济价值与应用前景。

5 竞争力分析

近年来，AI 一直在高新技术领域保持着相当大的热度，在未来 AI、5G、物联网、云计算和大数据等技术的成熟与广泛运用肯定会让“万物互联”的世界焕然一新。透过互联网思维进行横向观察，AI 已经普遍进入大众的视野并在生活、学习、工作等方面有着广泛的应用。作为 AI 云学习的工具，本项目不管是在理论层面还是技术层面都有着强有力的核心竞争力。为此，我们分别建立波特五力分析模型和 SWOT 模型分析本项目的核心竞争力。

5.1 波特五力模型

波特认为行业中存在着决定竞争规模和程度的五种力量，这五种力量综合起来影响着产业的吸引力以及现有企业的竞争战略决策。五种力量模型确定了竞争的五种主要来源，即供应商和购买者的讨价还价能力，潜在进入者的威胁，替代品的威胁以及最后一点，来自在同一行业的公司间的竞争。如下图 5.1.1 所示。

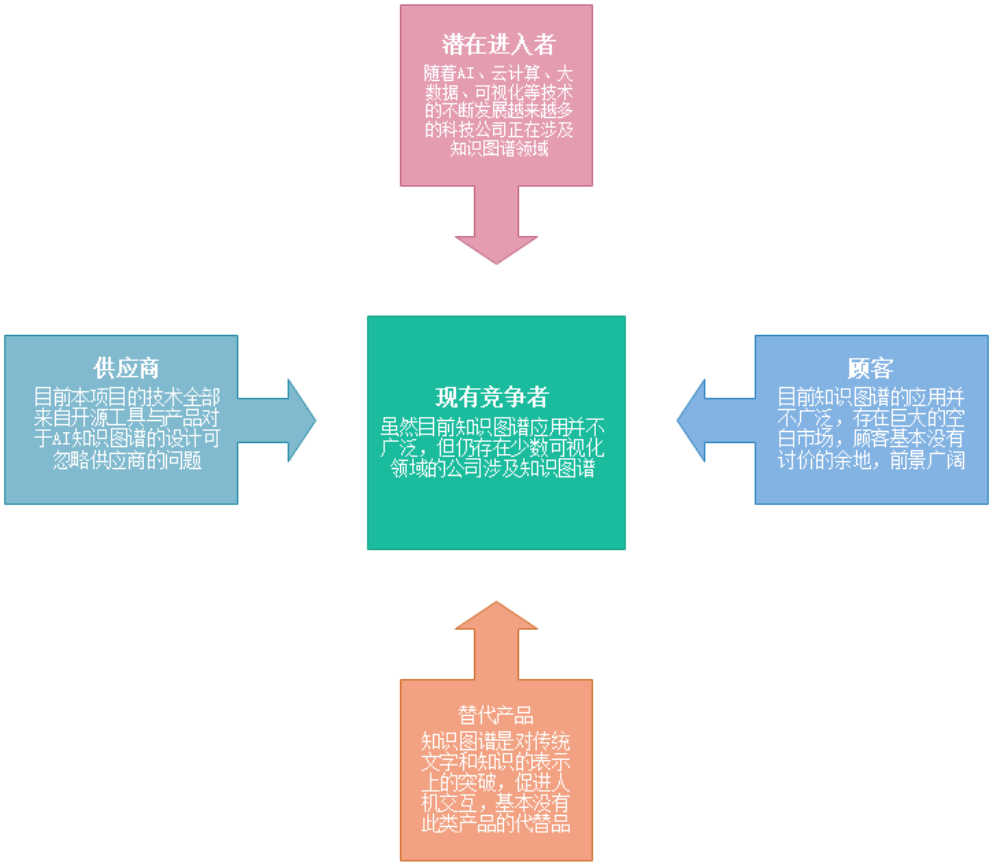

图 4.4.2.1 知识图谱项目波特五力模型分析

5.1.1 现有竞争者

作为 AI 领域的一项成熟的技术，目前在市场上知识图谱的应用反而不太广泛，大多数的商业公司甚至一些 AI 科技公司对知识图谱的应用与前景认识和把握得不够充分。诚然，目前确实有一批商业公司在知识图谱领域投入市场并进行商业运用，但目前没有大力投入知识图谱领域的科技公司，如百度搜索引擎中的一些关键词利用词云技术关联，图书馆中检索系统的书本关联和基本信息关联的知识图谱，目前还尚未成熟的人物人际关系分析系统等等。由于目前应用不太广泛，知识图谱在技术上基本没有创新，并且科技公司投入的研发力度不能尽人如意，传统的知识图谱构建技术已经出现了新的技术壁垒。不同于传统的知识图谱构建的技术路线，本项目在构建知识图谱的路线上运用自己独特的创新点，在数据爬取和数据处理方面，将 Spark 大数据计算平台强有力的并行处理能力，以及超快的数据处理速度，结合阿里云的云计算能力充分发挥了各大平台的优势，同时在可视

化方案的选择上，我们选择了功能强大的 amChart 4，整套技术路线是原创的，使用的工具全部是开源的，这些都属于本项目的核心竞争力。

5.1.2 潜在进入者

近年来 AI 领域的热度居高不下，各大科技公司纷纷进军 AI、云计算、大数据市场。但知识图谱是一个独特的存在，作为一项人工智能领域已经成熟的技术，传统的知识图谱大同小异，要么是知识冗余、关联度不高要么是效果呈现不好、应用不够广泛或者是产品基本没有更新或者更新迭代的速度不能适应使用的场景，加之传统的技术路线对于目前火热的 AI 领域市场，知识图谱的构建技术鲜有人去进行创新。就目前知识图谱的市场而言，现阶段的竞争者对于本项目构成的竞争影响不够大，毕竟本项目的核心竞争力就是全套原创的技术路线和开源的构建工具，这是本项目区别于现阶段其他竞争者的本质，本项目无论是在技术领域还是商业应用的领域无疑都可以在领先的技术水平下进行产品的升级与转型，在市场上保持自己的领先地位。

5.1.3 替代产品

目前，市场上的知识图谱应用仍然存在着大量的市场空白，传统的知识图谱构架技术针对复杂的应用场景，不能够灵活地进行产品迭代和转型，无法在短期的投入下看到成效。因此，目前市场上暂时还找不到知识图谱的替代产品，加之目前市场是知识图谱构建供不应求，许多需要应用知识图谱的领域往往由于技术原因而得不到充分的发挥，产品迭代速度跟不上产出的效能。而本套知识图谱构建的技术相对于传统的技术更为创新，其应用前景更为广阔，迭代速度更快，在生活、学习、商业等方面有着巨大的市场。

5.1.4 供应商讨价能力

本项目是一套软件构建的技术，针对于不同行业，不同人群，不同应用场景都可以进行自适应，且技术路线属于团队原创，正拟申请国家专利，构建工具也是遵守 Apache Licence 完全开源，不存在供应商讨价能力这一层面的影响。

5.1.5 顾客讨价能力

由于本项目是一套软件构建的技术，针对于不同行业，不同人群，不同应用场景都可以进行自适应。本项目技术流程需要针对不同的顾客、不同的使用人群进行智能匹配和迭代。对于目前急需知识图谱技术支撑的企业，由于可以构建高效知识图谱的科技公司极少，市场存在大片空白。顾客基本没有讨价的能力。此外，知识图谱一旦在学习、生活、商业尤其是商业应用带来的效益高于构建的投入时，新型的构建知识图谱技术对于顾客讨价的空间会经一步的缩小。

5.1.6 知识图谱领域环境总结

通过对本项目在知识图谱应用的场景和前景上进行行业五力竞争模型评估，在模型中行业竞争主要威胁是潜在的进入者，但本项目的核心竞争力就是全套原创的技术路线和开源的构建工具，这是却别于现阶段其他竞争者的，本项目无论是在技术领域还是商业应用的领域无疑都可以在领先的技术水平下进行产品的升级与转型，在市场上保持自己的领先地位。

5.2 SWOT 分析

基于内外部竞争环境和竞争条件下的态势分析，就是将与研究对象密切相关的各种主要内部优势、劣势和外部的机会和威胁等，通过调查列举出来，并依照矩阵形式排列，然后用系统分析的思想，把各种因素相互匹配起来加以分析，从中得出一系列相应的结论，而结论通常带有一定的决策性。运用这种方法，可以对研究对象所处的情景进行全面、系统、准确的研究，从而根据研究结果制定相应的发展战略、计划以及对策等。如图 5.2.1 所示。

图 5.1.6.1 AI 云学习 知识图谱 SWOT 模型分析

5.2.1 内部环境分析：优势、劣势及对策

优势：

◆在技术上，本项目结合人工智能、Spark 大数据平台、云计算等前沿技术，构建工具全部开源，技术路线完全自主创新，正积极申请专利。

◆在市场上，知识图谱市场空白、应用前景广泛、市场竞争小、项目灵活。可根据不同的使用场景进行自适应，可运用在学习、生活、商业、军事等环境，面向大众化人群。

◆技术路线的创新性和广阔的应用前景构成了本项目核心竞争力，针对不同人群、不同应用场景、整套知识图谱的构建流程大同小异，更改数据源即可进行不同场景和人群的自适应与匹配，更利于知识库的不断自我进化和更新，便于不同产品之间的更新与迭代。

劣势与对称：

◆初期项目计算机等硬件资源投入较大。由于数据获取、数据清洗等环节对于需要分析巨大数据量的知识图谱，前期需要投入一定的硬件成本支持大数据和 Spark 平台构建与运行，解决方案是前期的硬件资源可以分摊给多个 slave 机器，如本项目初期利用 1 台 master 云服务器和 3 台本地笔记本主机进行分布式爬虫获取数据，节约成本，后期可视化

属于软件部分等成本投入几乎为 0。

◆项目初期仅用于学习场景，没有稳定的客户进行场景自适应匹配。对学习以外的场景进行训练和自适应，如开放 API 和知识图谱的接口给公安系统中的人物关系知识图谱，利用 AI 算法帮助公安进行分析、计算与推理，亦或是开放旅游大数据的知识图谱 API 进行数据集的训练。

◆项目广泛运用工业流行的新型技术，针对项目成员的技术要求较高。由于本项目涉及的技术都是基于当下流行的开源技术，对于技术的创新仅存在与当前已存在并流行的工具和技术，解决方案是作为领跑者开放自身的技术路线并构建知识图谱生态系统，随着知识图谱的不断应用整个 AI 领域和知识图谱市场会诞生一系列优秀的产品，此时技术壁垒会被千千万万的科技公司一同打破。

5.2.2 外部环境分析：机遇与威胁

应用前景和经济前景广阔，在 AI、5G、物联网等技术前提下，万物互联带来一系列机遇与挑战。用知识图谱强大的语义化和可视化的双重冲击，便利使用者。

存在的机遇：

◆图谱问答（语音助手、智能电视）。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。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。

◆学习工具（知识分析、计算、推理）。作说到学习工具，目前市场上充斥着大量产品。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。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。

◆商用知识图谱（金融、公安、旅游等行业）。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。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。

潜在的威胁：

◆知识图谱目前没有一个完善的体系和商用化的标准，可能存在后期的技术壁垒。作为领跑者开放自身的技术路线并促进知识图谱生态系统的构建，随着知识图谱的不断应用整个 AI 领域和知识图谱市场会诞生一系列优秀的产品，此时技术壁垒会被千千万万的科技公司一同打破。

◆随着 AI、5G、物联网的应用，技术的不断升级，知识图谱的广泛应用会导致大量科技公司涌入知识图谱行业，压缩竞争本项目的生存空间。

6　组织与人员

6.1　团队目标

本项目团队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，通过 Spark 完成人工智能知识的重整，实现一个学习者尤其是本科生适用的知识图谱工具。

6.2　组织结构及各组职责分配

团队组织结构图如图 6.2.1 所示。

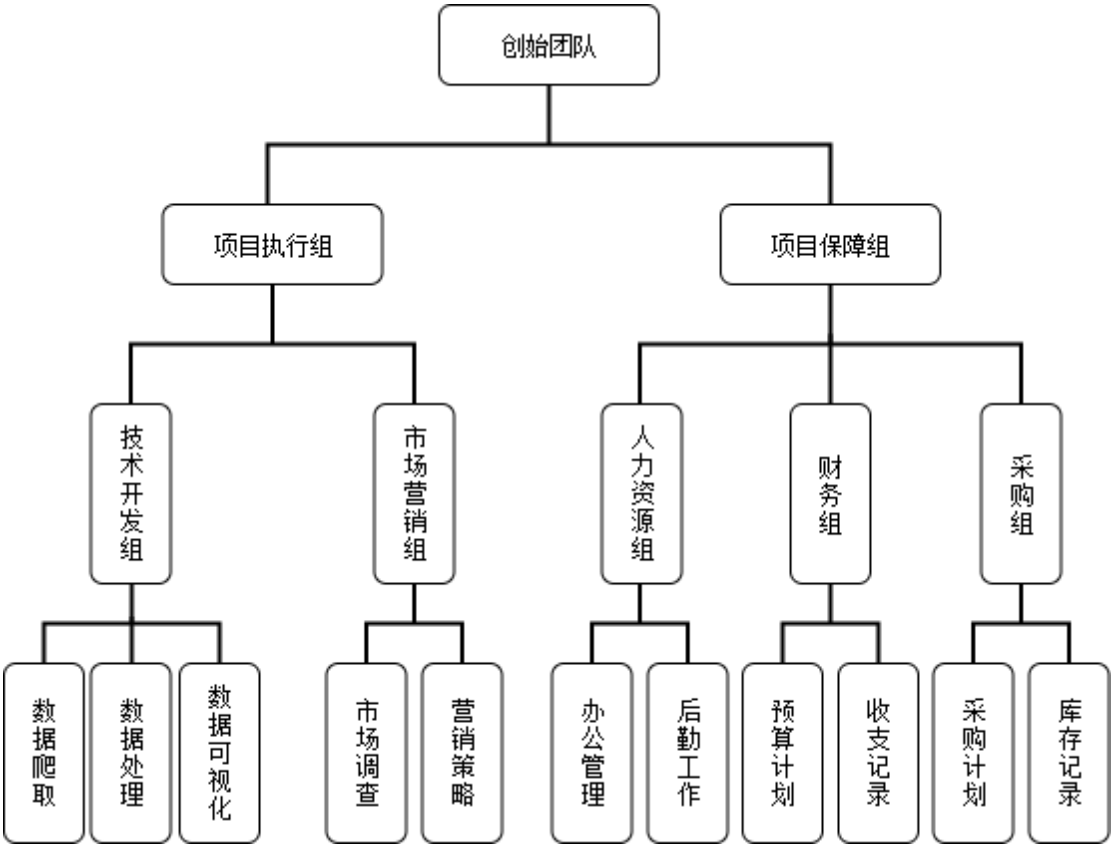

图 5.2.2.1 团队的组织架构图

1. 市场营销组

①对市场有着灵敏感知，了解趋向发展前景；

②对产品的功能的设计，适应人群进行综合把控；

③进行市场调查研究与分析，提出营销策划方案，提供团队决策依据；

④制订营销方案和计划，并付诸实施；

⑤制定并执行品牌定位及整体宣传策略、市场调查及竞争对手的动态收集；

⑥做好产品宣传工作；

⑦制定融资方案。

2. 技术开发组

①结合市场营销组提供的可靠信息进行产品设计；

②制定合理的创新技术研发计划和阶段性目标，并付诸实施；

③升级改善核心技术；

④从各个方面完善产品，使之更贴合消费者需求。

3. 财务组

①办理各种财务事务；

②制定季、年度财务预决算文件；

③合理分配、核算、监督团队在项目开发经营过程中的各种财务行为；

④制定各类财务制度；

⑤按天、月、季度、年分别做出清晰明了的财务报表。

4. 采购组

①制定项目所需采购计划；

②负责物资材料质量证明和相关资料的索取、下发和整理入档、负责购销合同的传递和入档管理；

③负责办理材料入库、记账、结算、报账等有关的业务手续。

④定期编报材料采购报表，分析采购价格及管理费用的开支，降低采购成本。

4. 人力资源组

①负责团队的人力资源管理；

②处理日常办公室工作；

③制定团队日常工作规章制度；

④整理保存各类文件档案；

⑤协助团队申报办理各类有关团队运作的手续；

⑥负责日常的后勤工作；

⑦执行办公室的规章制度，管理好办公室的日常办公秩序。

6.3 主要成员

6.3.1 前期主要成员

表 6.3.1.1 团队前期主要成员

团队成员	所在部门	工作分配
文华	技术开发部	负责产品数据处理及项目的统筹与规划等
刘宏鑫	技术开发部	负责项目数据可视化及前端页面设计等
周余	技术开发部	负责项目所需数据爬取及软件后端编程等

6.3.2 后期主要成员

表 6.3.2.1 团队后期主要成员

团队成员	所在部门	工作分配
文华	技术开发组	负责产品数据处理及项目的统筹与规划
刘宏鑫	技术开发组	负责项目数据可视化及前端页面设计
周余	技术开发组	负责项目所需数据爬取及软件后端编程
陈叶红	市场营销组	负责项目当前及未来相关市场调研数据分析
王文举	市场营销组	负责根据市场分析指定相应营销策略
刘城浩	采购组	负责材料采购并辅助完成文案设计
郭立程	财务组	负责财务管理并辅助完成文案编写
林聚	人力资源组	负责人力资源管理并辅助完成文档查验

6.3.3 指导老师

表 6.3.3.1 团队指导老师信息

指导老师	所在院系	研究方向	个人简介
周波	计算机与信息学院	人工智能，大数据，云计算	合肥工业大学副教授

6.3.4 团队概况

本项目团队主要由 8 名学生与 1 位指导老师组成，有负责前后端等技术的同学 3 名、负责调研和文案的同学 2 名以及负责人员协调和财务分析的同学 3 名，指导老师是从事数据分析相关领域多年的教授，可提供专业的知识理论支撑。

团队组建特征：

（1）团队发展目标清晰：把自身利益和社会利益相结合，作为一个创业团队，队内成员有多年负责活动及项目策划的经验，我们时刻清楚自身的目标，通过一个五年期的计划，把 AI 云学习打造成一流的 Web 应用及 APP 的实业，实现我们的创业梦，让万千学子在 AI 云学习中轻松实现高效学习。

（2）团结且价值观统一：团队崇尚开放、诚实、协作的办事原则，同时鼓励队员自主参与，并定期开展思想交流会，化解冲突的同时交换想法激发灵感，组织内部比较容易形成相互信任的环境。多数成员有党员背景及担任班委经历、奉献意愿强烈，渴望在团队内实现个人价值，因而调动了团队内工作积极性，使得项目完成进度速度大大提升；在项目进行中，本团队各成员合作紧密，各部门间工作沟通频繁，针对现状及时调整工作方向，提高工作效率。综上团队整体梯队结构合理，在所从事的技术方面均具有独挡一面的能力，同时积极进取、热心求学、拥有良好的团队精神。

（3）专业素养高且技能互补：作为计算机类本科生团队，本团队成员均有过软件开发背景，对于 B/S 架构软件的开发流程十分熟悉，在软件制作时分工明确，前端、后端及数据库均有相关专业人才负责开发和维护，良好的个人技能能够让我们能出色地完成任务，在该项目开发过程中各个技术模块间衔接顺畅，各项工作的完成速度及完整度非常高。

6.3.5 团队管理

我们在创业过程中，对团队实施了有效管理：①目标管理：团队树立了共同的知识服务目标。在统一的目标管理作用下，团队成员共享目标，在明确的方向下携手共进，共同努力完成团队目标。②团队合作：在项目进行中，加强团队合作，各部门经常性有效沟通，提高各部门之间的工作效率。③时间管理：对项目的进度进行了合理的安排，并且坚持按照既定的计划来实施每一个步骤。④流程化管理，团队以流程为主线的方法管理，强

调以流程为目标，以流程为导向来设计组织框架，同时进行业务流程的不断再造和创新，以保持企业的活力。⑤人员管理：对于项目各部分负责人，充分考虑其背景、经历等因素，做到将合适的人放到合适的岗位上。⑥学习管理：在项目进行的过程中，大家都怀着极大的热情来学习新知识，弥补自己的不足。

6.4 团队战略

6.4.1 团队定位

团队专注于智慧知识图谱构建与优化，利用数据采集技术，与移动互联网、人工智能高度融合，通过软件工程标准手段，开发建立搭载专家系统的知识图谱平台，为后续大规模的本科生宽泛的相关性概念理解或深入地系统学习提供良好的解决途径，技术产品和技术应用不仅符合中国大体量本科教育现状和人文特点，而且与国内外同类产品相比，具有较高性价比。未来本团队将不断完善技术创新、专利申请等工作，将团队的技术优势转化到知识服务领域的规模化应用中,以互联网、大数据、人工智能为依托，以知识检索解决方案的研发、销售为突破口，打造国内一流的互联网+创业团队，推动我国学习教育现代化进程，为我国信息化发展做出卓越的贡献。

6.4.2 团队愿景与使命

当前我们已经身处全民学习和人工智能的巨大热潮中，在知识信息爆炸的今天，每个人都需要一款可以有效梳理网上有效信息的工具，在我们的知识图谱平台下，用户可以体验到多个热门领域的知识图谱、相关知识进行相关性检索，并得到关键词的相关扩展结果，满足不同用户对知识的不同需求。打造一套符合人工智能新时代学习方法体系，并提供迎合实际需求的优质产品与服务，策划适用于广大知识用户的、性价比高、经济性的解决方案，铸就国内一流，对世界有影响力的相关信息检索和知识服务品牌。

6.4.3 团队理念

团队本着“为用户节省时间，以知识服务为核心，升级知识网络体系结构，创造新时代知识图谱”的核心理念，为客户服务、为当前互联网知识产业结构的升级发展出力。

本团队打造一流、先进产品的根本目的是服务知识体系于客户，通过优秀的产品和

优质的服务满足客户需求，更好的推动用户的知识积累，取得广大用户的信任，促进项目技术、管理等方面的不断创新，逐步推动团队发展，实现社会价值。

7 财务分析

7.1 创业资金来源

本团队成员都是在校大学生，创业初期资金非常匮乏，而本项目的资金花费少维护成本低，对于初期小规模创业很是契合，到后期进行业务拓展网站更新搜索引擎升级时，需要的资金可由前期获利弥补，可基本满足项目需求，但是本项目具有良好的发展前景特别是在大数据趋势愈发明显的现代生活办公中，而如何快速发展项目来及早参与到大数据的趋势中来就成了将本项目获利最大化关键问题，所以如何尽快获得一定数量是目前的主要问题，可使用的融资方式为：

创业贷款：利用政府大力扶持大学生创业的契机争取获得政策性贷款，政策性贷款一般是政府贴息的，贷款成本很低，我们可以充分利用这些优惠条件，为创业获得更多的启动资金。

本团队第一年需投入 75 万元人民币。

预计本团队前 5 年的年平均盈利达到 60 万元，结合对我们的技术优势、市场份额定位、产品销量、经济效益分析，整个项目估值约 500 万元。

7.2 资金使用分析

7.2.1 运营费用预期（第一年）

- 10 万用于购买办公设备
- 10 万用于团队日常经营费用
- 30 万用于团队人员工资
- 10 万用于市场推广费用
- 5 万不可预见费用
- 小计：65 万

7.2.2 生产流动资金预期

按第一年销售 500 套系统，每套系统采购成本以 1000 元计算，需要 50 万

采购资金；以每套售价 3000 元计算，大约有 100 万元营业收入。考滤到第一年的流动资产周转率比较低，按 1.5 次来计，我们总共需要的流动资金大约为：67 万元。

7.3 三年内销售盈利预测

表 7.2.2.1 团队三年内销售盈利预测

(万元)	第一年	第二年	第三年
收入预测	**100**	**200**	**300**
办公室租金	10	10	20
工资	30	50	70
材料成本	10	20	30
管理费用	4	8	15
销售费用	5	10	15
财务费用	4	8	15
支出合计	**63**	**106**	**165**
盈利情况预测			
毛利润	50	90	130
营业利润	37	94	135
所得税率	15%	15%	15%
所得税	5	14	20
净利润	**32**	**80**	**115**
净利润率	**32.00%**	**40.00%**	**38.33%**

8 风险与对策

随着互联网技术的快速发展、硬件设备的快速更新迭代与目前即将快速普及的 5G 技术的应用，在当今的市场，特别是在移动互联网行业，优秀产品层出不穷，市场方向多种多样，竞争也非常激烈。尤其是在 2020 年新冠疫情的影响下，线下实体行业首当其冲受影响最大，互联网行业虽然所受冲击较小且有借此迅猛发展的态势，但是仍不可掉以轻心。不确定的经营风险是企业投资经营前必须考虑的一个重要因素。在市场竞争，经营管理，技术，财务等方面都存在一定的风险。针对现有可能存在的风险，我们做了一定的分析，并对此做了相应的应对策略。

8.1 风险分析

8.1.1 市场竞争风险

现在市场上已经有许多大型知识图谱被构建出来，而且已经有商业公司在知识图谱领域投入市场并进行商业运用，与这些公司相比，我们起步晚，团队实力也不够雄厚，在市场竞争中处于很不利的地位。且 AI 云学习作为一款刚上线的 Spark 构建知识图谱的人工智能学习工具 app，在市场、技术、影响力上都无法与这些成熟公司相比。这会对 AI 云学习市场占有产生极大的市场竞争风险。

8.1.2 经营管理风险

目前 AI 云学习的创业成员均为在校大学生，经营管理方面有着很大的不足，运营团队经验欠缺，在决策执行方面也有所欠缺，与那些已经成熟的互联网公司相比处于劣势，这在创业过程中显然是一个不可避免的问题。且后续可能由于部分成员需要考研而离开团队，或者因就业问题部分成员去更大的公司发展，这就使团队面临人员流失的困境，过大的人员流动势必会影响整个团队的运转，进而对团队产生不利影响。

8.1.3 技术风险

互联网的快速发展给社会带来了便利与经济效益，但互联网与手机网络中存在病毒

黑客等问题也是不可忽视的，这些问题可能会导致 AI 云学习出现使用故障，或者服务器被攻击造成用户数据泄露等严重问题，对这些问题的预防和处理必然要极度重视。同时，作为一个在校大学生创业团队，我们要认识到自身技术上的不足，与已经成熟的公司相比我们在技术积累、人才积累、研发团队、产品的更新迭代研发上还有很大的差距。这些问题与技术差距会很大程度上制约团队的发展速度跟发展潜力。

8.1.4 财务风险

资金的供应流动对一个团队来说至关重要，投资方的投资意向对我们来说是机遇也是风险，一旦在产品初期投资方资金撤出势必会对产品产生影响，运营资金不足，团队发展也会陷入困难。除了投资方方面可能带来的问题之外，团队内部的资金使用也要引起重视，对于产品研发，测试，优化，运营，推广等费用都要进行严格透明的控制监督方法，防止内部原因导致资金链断裂，团队发展困难。

8.2 风险规避对策

8.2.1 市场竞争风险对策

目前市场上已经出现了一些商用的知识图谱工具，但我们发现这些知识图谱工具普遍存在着垂直知识图谱工具种类少、知识冗余混乱、组织零散、系统性差等缺点。它们的这些缺点正是 AI 云学习与其竞争的优势，解决这些问题正是 AI 云学习的目标之一，对这些问题的专门处理与解决就是我们在市场中的优势所在。

和其他的商用软件相比，AI 云学习的用户定位更集中于在校学习的大学生尤其是本科生，因此在前期宣传时可以把宣传资源集中在在校大学生方面。

基于以合肥工业大学为起点，积极利用同校同学之间的友好关系进行推广，同时可以沟通相关专业的老师询问是否可以推荐学生使用 AI 云学习帮助学习，同时充分发挥不同学校之间同学相互认识的优点，以合肥工业大学为起点积极向各个高校进行推广 AI 云学习，迅速占有市场。

技术更新与产品设计不断的进行优化调整，由于主要用户人群为在校大学生，除了要提供更便捷高效准确的学习帮助之外，也要提供更加多元化个性化人性化的使用体验。

在与当前成熟企业有资格对抗较量之前，对 AI 云学习产品的部分功能加以保密，逐

步解锁，确保始终保持在想法创新上领先对手。

8.2.2 经营管理风险对策

组织选拔有相关知识经验的人员成立一个专门的市场运营团队，主要用来负责团队日常运营时出现的问题，尽力避免因运营失误而产生的不良影响；在一些重大的决策实行之前，需要召开核心成员大会，大家一起对问题进行分析讨论，最后投票选出可行的方案。同时要提前建立一套行之有效的面对运营失误的高效处理方案，尽量将这类问题扼杀在萌芽之中，对没有避免掉的问题尽量将影响与损失降低到最小。

对于团队的人员流动问题要发挥创业团队中人员关系密切的优点，以同学关系为枢纽，在学校中各个成员关系密切，团队成员之间可以不定期的聚集在一起交流各自的想法，同时各个成员对自己未来的想法规划进行交流，避免出现有成员突然离开，对团队造成影响。

同时要运用战略性的人员管理思想，对关键岗位要提早实施人才储备制度，注意培养有能力有潜力的员工作为关键岗位的接班人，在不影响团队事务的前提下，多带一些有潜力的员工到相关的场合观摩学习，培养其以后面对此类场合与问题的能力。

建立完善的考核升职制度，同时聘用不同专业背景的员工，采取多元化人性化的管理措施，提升员工的认同感与自我价值的实现感。积极接纳有突出贡献与优秀能力的人才进入团队核心，激励员工的奋斗热情。

8.2.3 技术风险对策

在 AI 云学习软件的研发之初就要考虑靠安全性问题，在设计与编写软件代码时要注意系统的安全性，尽力避免可能出现的 BUG。在后期进行软件测试时要着重对软件的安全性进行测试，确保能够保证用户的个人信息安全。同时挑选优秀技术人员成立技术小组，负责软件的日程管理与优化，给用户良好的使用体验；同时负责在出现软件安全事故时，能够对问题及时的进行排查与处理，及时排除安全风险。

技术人才是好的产品与产品安全的重要保障，对此团队要积极招收技术人才，对新招收的技术人员有优秀员工进行培训，让他们尽快的能发挥自己所学。也可以在合肥工业大学相关的专业中寻找合适的同学加入，以提高技术人才储备。同时应当适当提高技术人员的薪资福利，留住技术人才，尽快的完成人才积累。

积极的对产品进行优化升级，在积累一定的技术人才之后可以适当的加快产品的更新迭代周期，同时研发新的技术，为软件添加新的功能，积极扩大优势牢牢抓住用户市场。

8.2.4 财务风险对策

在吸引投资方投资时也要进行适当的筛选，最佳选择是选择短期投资方，前期尽快完成自身的资金积累，在投资方拿到自身的回报之后我们也能及时摆脱对其资金上的依赖，避免团队发展受到投资方的影响与控制。

成立财务部门，对于产品研发，测试，优化，运营，推广等费用都要进行严格透明的控制与监督，同时定期向管理团队提交财务报告，确保内部资金流动的安全。设立完善的资金申请审批监督制度，防止内部人员出现报假账贪污挪用公款等现象，确保资金都用在了应该使用的地方。

设立一部分预算，用于为团队购买各种必要的商业保险，用来避免一些意外或者不可抗力因素对团队造成的损失。

参考文献

[1]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.

[2]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.

[3]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.

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

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

[6]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.

[7]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.

[8]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.

[9]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.

[10]中国产业调研网. 2020【知识图谱】行业市场调研及前景预测分析报告[R]. 2020.

[11]艾瑞咨询. 去往认知海洋的一艘船 中国知识图谱行业研究报告[R]. 2019.

[12]中国电子技术标准化研究院. 《知识图谱标准化白皮书》（2019 版）[R]. 2019.

[13]李晓理,张博,王康,余攀.人工智能的发展及应用[J].北京工业大学学报,2020,46(06):583-590.

[14]肖仰华.知识图谱的下半场：机遇与挑战[R]

附录

产品获 2019 年 iCAN 国际创新创业大赛安徽赛区省级二等奖证书

产品获合肥工业大学创新创业教育中心审核通过证书

证书

省级大学生创新创业训练计划项目《基于 Spark 平台的人工智能知识的知识图谱构建》（项目编号：S201910359345）结题成绩为通过。特发此证。

主持人：文华　指导教师：罗月童

成　员：刘宏鑫、周余

合肥工业大学创新创业教育中心

二〇二〇年五月

注：本证书可通过“合肥工业大学大学生创新创业训练项目智能管理系统”查询真伪，查询网址为：https://cxcy.webvpn.hfut.edu.cn