Computer Science and Application 计算机科学与应用, 2021, 11(12), 3126-3135
Published Online December 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.1112316

Hans汉斯

面向水产养殖领域的对虾知识图谱云平台设计与实现

曹　亮[1,2,3,4]，庄振胤[1,2,3,4]，刘双印[1,2,3,4,5*]，李湘丽[1,2,3,4]，徐龙琴[1,2,3,4]，尹　航[1,2,3,4]，罗智杰[1,2,3,4]，刘同来[1,2,3,4]，郭建军[1,2,3,4]，刘建华[1,2]

[1]仲恺农业工程学院信息科学与技术学院，广东 广州
[2]广东省高校智慧农业工程技术研究中心，广东 广州
[3]广州市农产品质量安全溯源信息技术重点实验室，广东 广州
[4]仲恺农业工程学院智慧农业创新研究院，广东 广州
[5]石河子大学机械电气工程学院，新疆 石河子

收稿日期：2021年11月27日；录用日期：2021年12月23日；发布日期：2021年12月30日

摘　要

针对互联网中存在的松散型、碎片化和难整合的水产养殖领域知识现状，将知识图谱应用于对虾养殖领域，面向多元异构数据源进行知识抽取，采用BI-LSTM-CRF模型进行命名实体识别、TextCNN模型进行关系识别、Neo4j数据库存储获取的知识数据，建立基于SpringBoot框架的对虾知识图谱云服务平台，将分散的对虾知识有效整合为一个规范化、标准化和系统化的知识库，并采用SparkMlib的朴素贝叶斯分类算法完成问题模板的匹配，实现基于知识图谱的对虾智能检索、智能推荐、智能问答和疾病辅助诊断等功能，为养殖户、企业和科研人员提供便捷、有效和系统化的对虾领域知识。

关键词

知识图谱，Neo4j图数据库，命名实体识别，智能问答

Design and Implementation of Shrimp Knowledge Graph Cloud Platform for Aquaculture Field

Liang Cao[1,2,3,4], Zhenyin Zhuang[1,2,3,4], Shuangyin Liu[1,2,3,4,5*], Xiangli Li[1,2,3,4], Longqin Xu[1,2,3,4], Hang Yin[1,2,3,4], Zhijie Luo[1,2,3,4], Tonglai Liu[1,2,3,4], Jianjun Guo[1,2,3,4], Jianhua Liu[1,2]

[1]College of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou Guangdong

[*]通讯作者。

文章引用: 曹亮, 庄振胤, 刘双印, 李湘丽, 徐龙琴, 尹航, 罗智杰, 刘同来, 郭建军, 刘建华. 面向水产养殖领域的对虾知识图谱云平台设计与实现[J]. 计算机科学与应用, 2021, 11(12): 3126-3135. DOI: 10.12677/csa.2021.1112316

[2]Intelligent Agriculture Engineering Research Center of Guangdong Higher Education Institutes, Guangzhou Guangdong
[3]Guangzhou Key Laboratory of Agricultural Products Quality & Safety Traceability Information Technology, Guangzhou Guangdong
[4]Academy of Smart Agricultural Engineering Innovations, Zhongkai University of Agriculture and Engineering, Guangzhou Guangdong
[5]College of Mechanical and Electric Engineering, Shihezi University, Shihezi Xinjiang

Received: Nov. 27th, 2021; accepted: Dec. 23rd, 2021; published: Dec. 30th, 2021

Abstract

The independence, fragmentation, and looseness of aquaculture knowledge contents on the internet lead aquaculture hard to search and obtain integrated and accurate knowledge. For this purpose, a knowledge graph cloud platform of shrimp in the aquaculture field based on knowledge graph and SpringBoot framework is established. In which, the scattered knowledge of shrimp is effectively integrated into a standardized, and systematic knowledge. The platform includes four parts: the Bidirectional Long and Short Term Memory Network Conditional Random Fields (BI-LSTM-CRF) model is used for named entity recognition for the sake of extracting the knowledge from multiple data sources. The Text Convolutional Neural Network (TextCNN) model is used for relationship recognition; Neo4j database is used to store the acquired knowledge. The intelligent question answering and the intelligent search and information recommendation of aquaculture knowledge based on knowledge graph are realized by naive Bayesian classification algorithm to complete the matching of question templates. Our study provides users with a good environment for knowledge learning and using.

Keywords

Knowledge Graph, Neo4j Graph Database, Named Entity Recognition, Intelligent Question Answering

1. 引言

水产养殖行业中存在的“水、种、饵、混、轮、防、管、密”[1]八字真言，对应着养虾领域的关键技术要领。可见，养虾需要具备较全面和系统化的专业知识。当前，互联网虽然提供了对虾丰富的育苗技术、养殖模式、病害防治技术和加工技术等信息资源，但是这些知识来源广泛且分散，没有形成规范统一的知识，导致很多有效的养殖经验和先进养殖技术无法直接被有效共享和使用。

知识图谱作为一种知识管理技术，能有效重组海量信息数据并可根据用户需求提供多元化和个性化服务，可有效解知识的杂乱无章、零星无序和难以获取等问题。将知识图谱应用于对虾养殖领域，可对其相关知识数据进行知识抽取，将有效且完整度较高的知识进行知识融合，存储在图数据库 Neo4j 中，构建对虾知识图谱云服务平台，提供对虾知识的智能问答、智能检索、智能推荐和疾病辅助诊断与决策

等功能服务，并以可视化方式展示相关知识，以便满足用户在该平台学习及查询对虾养殖知识，给用户一个良好的知识体验。

2. 平台设计

2.1. 平台需求分析

1) 系统数据管理，系统数据包括系统用户信息管理，用户角色管理和用户权限管理以及字典信息数据管理。

2) 知识图谱的构建，要具备对知识库的基本管理，如知识数据的管理，知识模式层数据的管理。将知识图谱的基本应用迁移到对虾养殖领域，如对虾养殖知识可视化展示，以及对虾养殖知识的智能问答，对虾知识的智能检索与信息推荐等，形成一个系统化知识服务体系。

2.2. 平台功能模块

平台功能模块如图 1 所示。

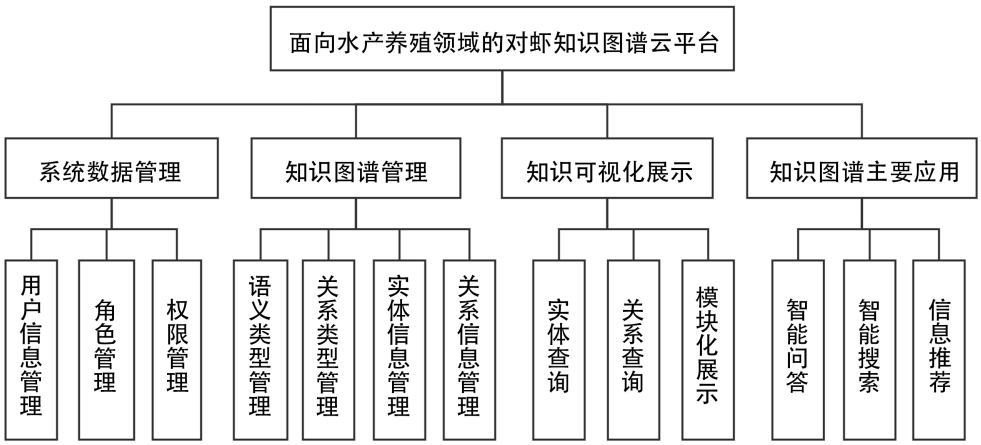

Figure 1. Platform functional modules
图 1. 平台功能模块

1) 用户管理，系统数据模块，管理使用用户的信息。

2) 用户角色以及权限管理，包括技术人员和领域专家等，不同角色有着不同功能模块管理权限。

3) 字典管理，系统可选项的基本数据信息管理。

4) 知识图谱模式层管理，模式层是对知识图谱知识规范化、统一化的知识表示定义，是较为核心的部分。系统需要提供对语义类型、类型的属性和关系类型基本的 CRUD 功能。

5) 知识数据层管理，数据层是模式层的实例信息，是知识图谱应用的数据支持。系统需提供对实体信息和关系实体关系的基本 CRUD 功能，并提供可视化的操作方式。

6) 知识的可视化查询，前台能提供知识的可视化展示，对较为聚集的知识图谱做出基本的整合以展示给用户，并提供基本的实体查询以及关系查询。

7) 智能问答功能模块，智能问答是一种更符合人们知识表达的思维和方式，能根据用户的提问直接到知识库中查询相关信息并返回结果给用户。

8) 智能检索与推荐模块，该模块是对虾知识图谱的一个主要应用方向，能有效解决对知识的有效利用问题，为用户提供更好的搜索服务。

2.3. 平台架构设计

系统架构采用 B/S (Browser/Server)模式，基于 SpringBoot 框架进行整个系统的开发，层次架构主要包括数据层、业务层和表现层(UI 层)，平台架构如图 2 所示。

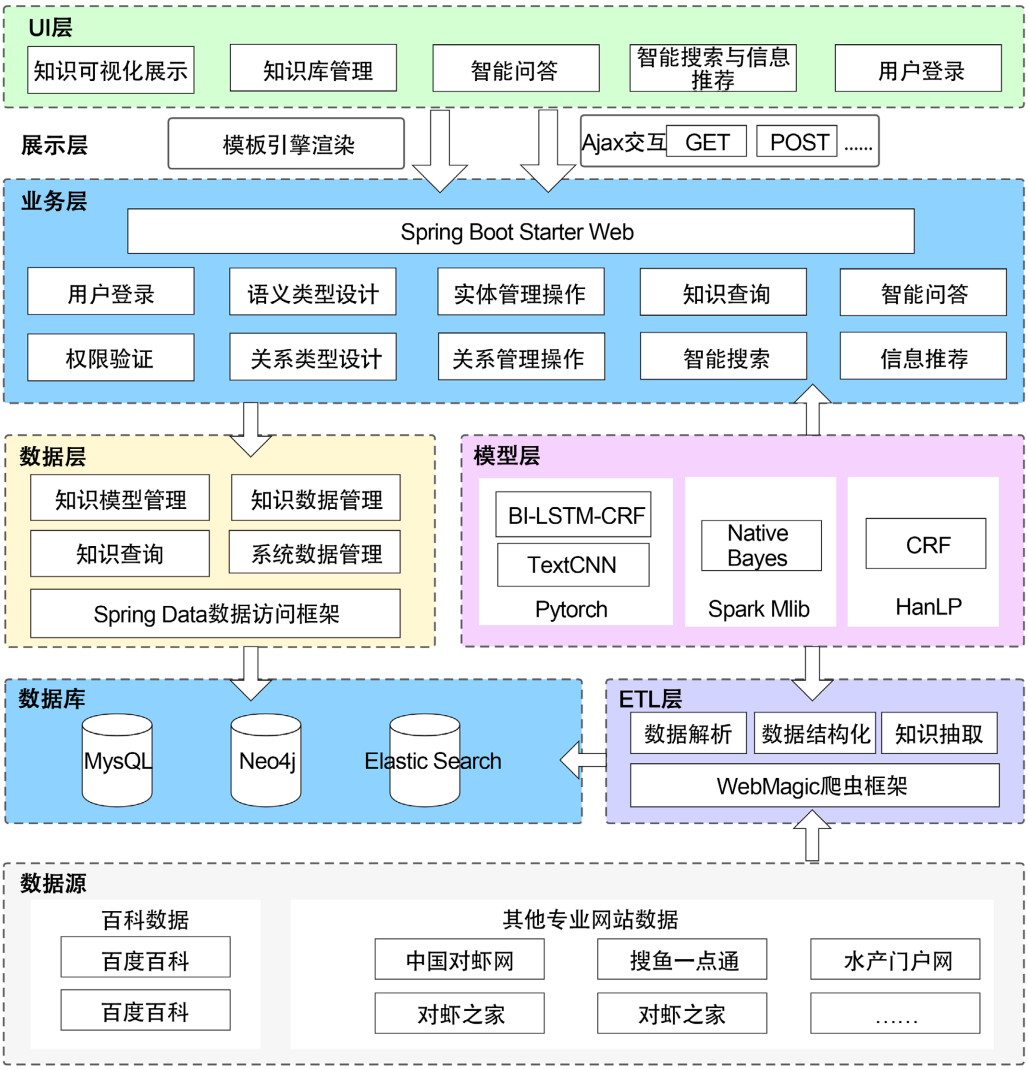

Figure 2. Platform architecture diagram
图 2. 平台架构图

1) 数据层：是整个系统的存储核心，其中 Neo4j 数据库是一种高性能的 NoSQL 图形数据库，他将结构化的数据以网络的方式存储，而不是以表的方式存储，因此，非常适合存储和表达知识图谱的实体、实体属性和实体关系。MySQL 数据库是一种安全、跨平台、高效的关系型数据库，以表格的方式存储数据，数据存储容量大，作为用户数据存储和知识模型的存储。Elastic Search 是一个开源的分布式、RESTFUL 风格的搜索和数据分析引擎，它的底层是开源库 Apache Lucene，用于处理纯文本的数据，检索非常高效，作为对虾养殖知识信息的存储。Neo4j 数据库数据主要运用网络爬虫技术，从百科数据以及对虾专业网站获取原始数据，并经过数据清洗、知识抽取和结构化处理等工作后导入数据库，此任务由 ETL (Extract-Transform-Load)层完成。知识抽取需要引用模型层的服务。

2) 业务层：是介于表示层与数据层之间的一个层次，它是整个系统的功能逻辑核心。其中部分功能需要调用到模型层的服务，包括负责命名实体识别模型 BI-LSTM-CRF、负责关系识别的 TextCNN (基于 Pytorch)和朴素贝叶斯分类器(基于 Spark Mlib)。业务层与数据层的交互使用 Spring Data 数据访问框架实现，包括 Spring Data Neo4j、Spring Data Jpa、Spring Data ElasticSearch 等。

3) 表现层：表现层与业务层的交互基于 RESTful API 的请求格式，以 JSON 数据格式实现前后台的数据传输，提升交互数据处理的效率。前台在采用 H5 技术的基础上，使用 UI 框架 Layui 进行界面的开发。其中，对于对虾养殖知识的可视化展示是系统主要功能之一，使用 Echarts 技术实现。

3. 功能模块与实现

3.1. 知识图谱概述

知识图谱是显示知识发展进程与结构关系，揭示实体之间关系的网络图形，以“知识抽取、知识表示、知识融合和知识推理”为主要特征。其中，知识抽取主要是对数据进行数据实体、数据关系和数据属性抽取；知识表示是通过各种模型(如复杂关系模型、距离模型、矩阵分解模型等)将知识因子与知识关联起来；知识融合是通过整合、去歧及更新等操作将各种不同的知识源融合成一个特需型的知识库；知识推理是在已有知识库中挖掘和分析出类似于人类思维的知识对象之间的层次关系[2]。

3.2. 实体抽取及设计

3.2.1. 实体抽取方法概述

实体抽取设计采用机器学习方法，主要是将命名实体识别问题转化为字符串标注问题，利用算法模型抽取实体方式，主要方法包括隐马尔科夫模型(HMM)、最大熵(ME)、支持向量机(SVM)、条件随机场(CRF) [3]和基于双向长短期记忆模型和条件随机场(BI-LSTM-CRF)的组合模型[4]。

3.2.2. BI-LSTM-CRF 模型

对虾养殖知识的命名实体的识别，采用 BI-LSTM-CRF 组合模型，其中，BI-LSTM 模型是双向的 LSTM 模型，其结构如下图 3 所示[5]。该模型主要分为输入层、Embedding 层、BI-LSTM 层、CRF 层以及输出层。

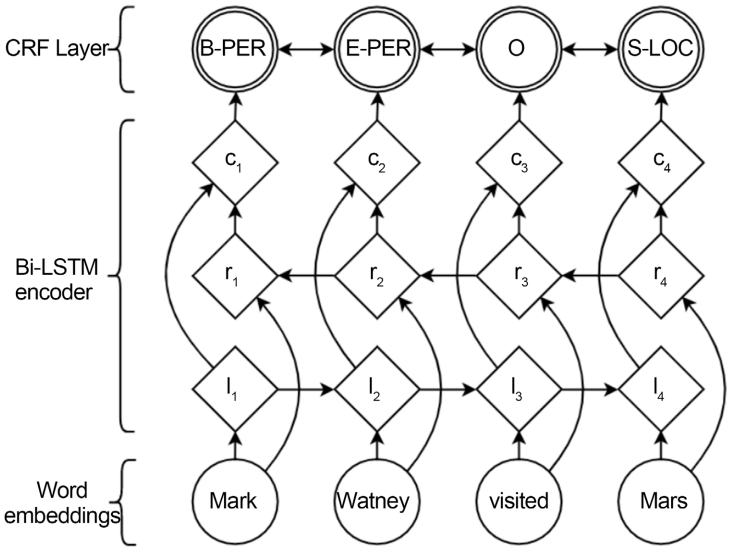

Figure 3. BI-LSTM-CRF model diagram
图 3. BI-LSTM-CRF 模型图

其中，Embedding 层主要是将每个输入层输入的字符转换成特征向量。BI-LSTM 层主要是使用从上一层中获取特征向量矩阵，然后根据输入的特征向量的正向序列和反向序列进行计算，最终得到两组不同参数的上文特征向量和下文特征向量；CRF 层主要是有效使用过去和将来的标签来预测当前标签，能在训练过程中记录句子输入的特征顺序，可以很好地避免上层预测结果的无效问题。

3.2.3. 设计实现与分析

首先进行语料制作，主要是标注语料。语料的标注便是根据模式层语义类型，对语料文本数据进行标注，标注方式可以分为基于词的和基于字的标注。本设计采用基于字的 BIOES 标注方法，在语料标注完成后，进行模型训练，其设计参数如表 1 所示。

Table 1. Design parameters of BI-LSTM-CRF model
表 1. BI-LSTM-CRF 模型设计参数

参数名称	参数值
学习率	0.02
衰减率	0.05
dropout	0.5
batc_size	16
epoch	30

在模型训练过程中，基于 tensorBoad 的训练可视化效果如图 4 所示。

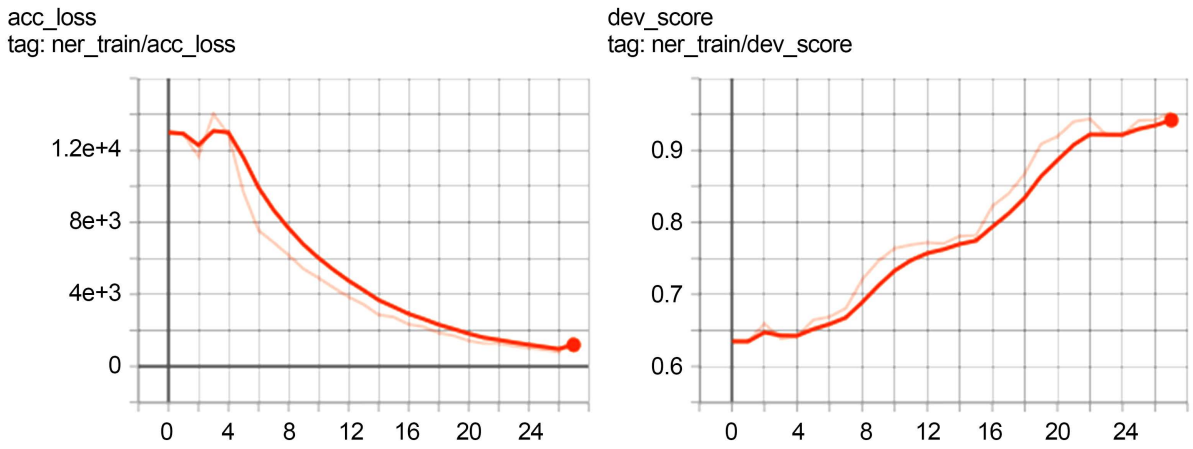

Figure 4. Visualization of BI-LSTM-CRF training
图 4. BI-LSTM-CRF 训练可视化图

训练结束后对训练模型进行评估，最终结果的准确率为 94.3%、召回率为 90.1%、F1 值为 92.8%。

3.3. 关系抽取及设计

3.3.1. 关系抽取方法概述

关系抽取采用监督学习方法，主要是将关系抽取转为文本分类。其抽取模型包括朴素贝叶斯(Native Bayes)、卷积神经网络(CR-CNN)、基于文本的卷积神经网络(TextCNN)、带注意力机制的卷积神经网络(Att-CNN)和带注意力机制的双向长短期记忆神经网络(Att-BiLSTM)等[6]。

3.3.2. TextCNN 模型

为获取虾类养殖知识实体之间关系，使用 TextCNN 模型进行关系抽取，其结构如图 5 所示[7]。

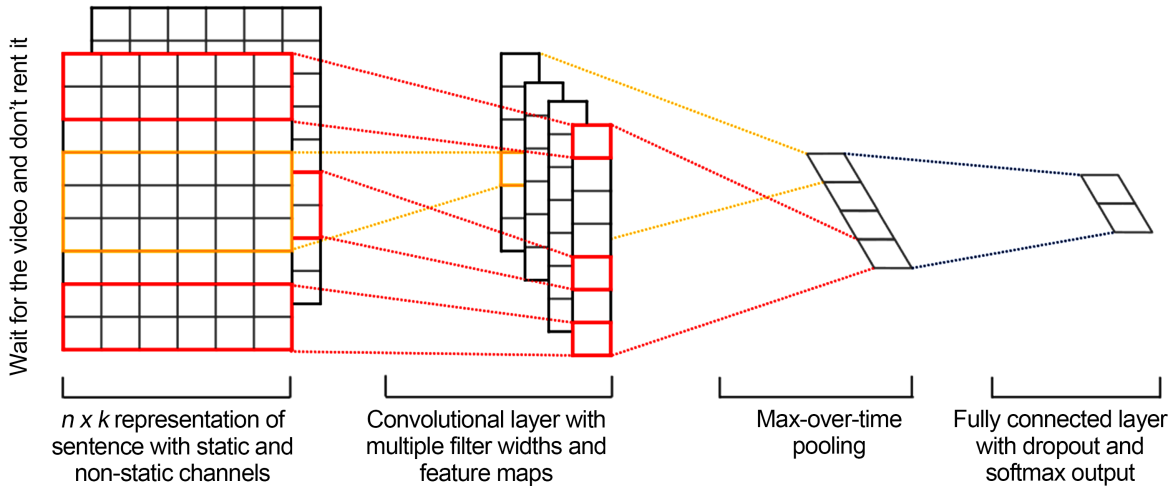

Figure 5. Structure diagram of TextCNN model
图 5. TextCNN 模型结构图

该模型主要分层为输入层、卷积层、最大池化层、全连接层。其中，输入层将输入的字符数据转化为向量，并使用一些已训练好且质量高的词向量(如 Word2Vec)，提高工作效率。卷积层根据输入层初始化的词向量进行卷积操作。池化层工作与上层大致相同，但需获取池化核中的平均值或者直接保留其中的最大值作为最终结果。全连接层即输出层，是池化后级联，为防止过拟合，dropout 设置范围为 0.3~0.5，使用激活函数分类输出。

3.3.3. 设计实现与分析

首先进行标注训练文本，以供后续模型训练。监督学习进行关系抽取的主要方式是将其转换成文本的分类问题，在训练文本完成后进行模型训练，模型设计参数如表 2 所示。

Table 2. Design parameters of TextCNN model
表 2. TextCNN 模型设计参数

参数名称	参数值
学习率	0.002
衰减率	0.05
dropout	0.5
卷积核数	100
卷积核大小	3~5

在训练结束后对训练模型进行评估，最终结果的准确率为 84.5%、召回率为 80.5%。

3.4. 智能问答设计

3.4.1. 基于 Spark 的朴素贝叶斯模型训练

朴素贝叶斯分类算法在多个领域都得到广泛的应用，并取得非常好的结果。但随着样本集数量增加，

其训练和运行效率会逐渐下降。因此，本设计采用基于 Spark 并行化的朴素贝叶斯分类器实现对虾知识图谱的智能问答，可以很大程度上提高分类算法的训练效率以及运行效率，其训练流程如图 6 所示。

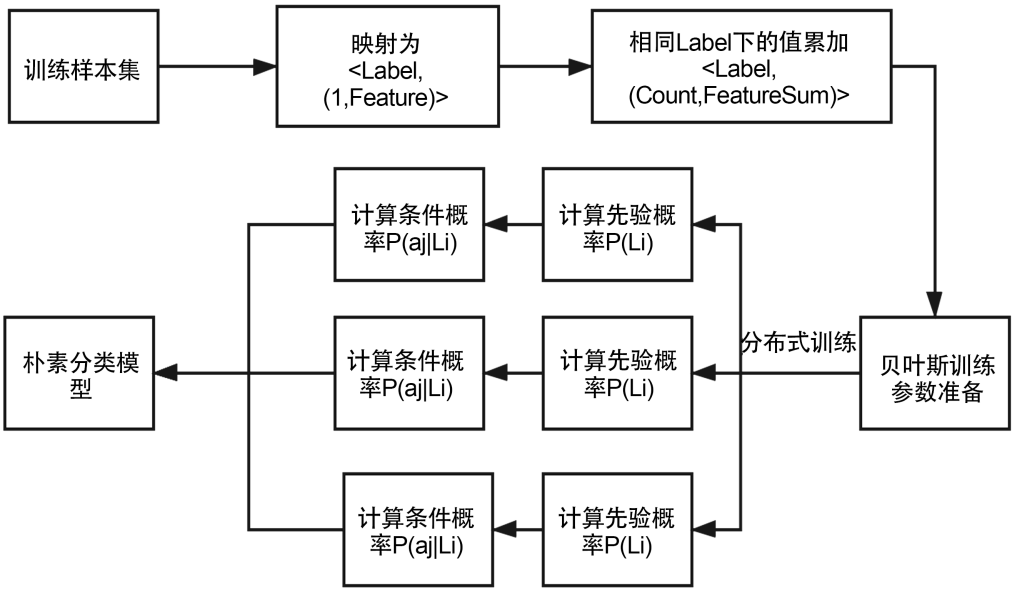

Figure 6. Training flow chart of Naive Bayes classification based on Spark
图 6. 基于 Spark 的朴素贝叶斯分类训练流程图

1) 在完成训练样本的构建之后，创建 RDD 并将本地的特征数据映射成<Label, (1, Feature)>格式。利用 one-hot 编码将文本信息映射为词向量，得到特征向量 Feature。

2) 对每一个 Map 数据进行聚合计算，统计全局样本数和特征向量总和，最终计算出<Label, (Count, FeatureSum)>结果。

3) 获取分类数量 Label 和分类数量 n，计算分类先验概率。

4) 对每个类型进行循环计算，获取每个类别条件概率。

5) 训练结束，获得朴素贝叶斯分类模型。

最后，将测试提交的自然语言问句向量化成矩阵输入到训练好的朴素贝叶斯模型中进行计算，针对每一个分类都计算出其可能存在的概率值，取其中的最大值作为最终的分类特征索引即可获取当前问句所属分类。

3.4.2. 设计实现与分析

根据用户输入的问句，采用拓展字典的方式，通过 HanLP 对查询的问句进行实体识别，将句子中的信息进行抽象化从而提取出特征信息；其次根据问句生成的词向量输入到朴素贝叶斯模型中进行计算，将每一个分类问题的概率全部计算出来，并取其中的最大值作为最终的结果；然后将得到的模板索引匹配到对应的模板问句，并将其特征词替换成当前的实体名称信息，还原问句生成计算机可识别的 Cypher 语句；最后查询 Neo4j 数据库，将查询结果返回到用户界面。输出结果展示如图 7 所示。

程序中首先对问题“白便病有哪些症状”进行了分词拆分和识别，实体“白便病”被成功识别并打上对应的标签；然后根据训练好的朴素贝叶斯模型，分别计算得出该问题在 7 个问题分类的概率值，最高的为第 3 个分类，概率约为 0.76；问题转换的最终结果为实体和实体关系，即白便病和症状。最终根据实体和实体关系，查询知识图谱数据得出答案。

```
========HanLP开始分词========
白便病/Disease
有/vyou
哪些/ry
症状/n
========HanLP分词结束========
句子抽象化结果：Disease 有 哪些 症状
the model index is 3.0
问题模板类型【0】概率：0.04673953681528046
问题模板类型【1】概率：0.0605224204421408
问题模板类型【2】概率：0.03976512155612536
问题模板类型【3】概率：0.7634903338776061
问题模板类型【4】概率：0.02641835065320439
问题模板类型【5】概率：0.029823841167094002
问题模板类型【6】概率：0.03324039548854883
句子套用模板结果：Disease 症状
原始句子替换成系统可识别的结果：白便病 症状
2020-04-25 21:46:26.192  INFO 28764 --- [nio-8080-exec-6] o.n.o
 (a:Disease)-[:Expression]-(b) where a.name = {0} return b.name
查询结果：游塘、偷死、肠道逐渐变白、拉白便、虾粪呈细长白色线状、肠壁细
 弱、空肠空胃、肝胰腺体萎缩、摄食缓慢、
```

Figure 7. Intelligent question and answer test results
图 7. 智能问答测试结果

4. 结束语

知识图谱是将复杂的知识领域通过数据挖掘和信息处理等方式，揭示知识领域的动态发展规律，为相关研究提供切实的、有价值的参考。构建面向产养殖领域的对虾知识图谱云平台，将互联网上的水产领域知识有效整合成规范化和通用性的知识库，便于各类用户便捷的获取相关知识，是知识图谱在水产养殖领域的一种尝试与探索。随着人工智能的快速发展，将水产领域知识图谱赋能认知智能，将会是一个有着广阔发展空间的方向[8]。

基金项目

国家自然科学基金(61871475、61471133)，广东省科技计划项目(2015A040405014、2017B010126001)，广东省教育科学规划项目(2018GXJK072、2020GXJK102)，广东省学位与研究生教育改革研究项目重点项目(2019JGXM64)，广东省教育厅科研项目(2017GCZX001、2016KZDXM001、ZHNY1903)，广州市科技计划项目(201903010043、201905010006、202103000033)，教育部产学合作协同育人项目(202002049009、201802235009、201802153181、201802153182、201802153191)，仲恺农业工程学院学位与研究生教育改革研究重点项目(新农科背景下高校农科硕士教育改革模式研究)，国家级大学生创新创业训练计划(202111347004、202111347001)，广东省大学生创新创业训练计划(S202111347057)。

参考文献

[1] 余开, 宋迁红. 听听专家怎么讲关于南美白对虾的那些事[J]. 科学养鱼, 2019(1): 20-23.

[2] 郑祉盈, 曹亮, 李湘丽, 等. 基于 Neo4j 图数据库的对虾养殖领域知识图谱研究[J]. 通讯世界, 2020, 27(11): 146-147, 150.

[3] 阚琪. 基于条件随机场的命名实体识别及实体关系识别的研究与应用[D]: [硕士学位论文]. 北京: 北京交通大学, 2015.

[4] Huang, Z., Xu, W. and Yu, K. (2015) Bidirectional LSTM-CRF Models for Sequence Tagging. *Computation and Language*, **3**, 1508-1991.

[5] Lample, G., Ballesteros, M., Subramanian, S., *et al*. (2016) Neural Architectures for Named Entity Recognition. *The* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics*, San Diego, California, 12-17 June 2016, 260-270. https://doi.org/10.18653/v1/N16-1030

[6] 曹春萍, 何亚喆. 融合 BSRU 和 ATT-CNN 的化学物质与疾病的关系抽取方法[J]. 小型微型计算机系统, 2020, 41(4): 794-799.

[7] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, Qatar, 25-29 October 2014, 1746-1752. https://doi.org/10.3115/v1/D14-1181

[8] 刘柳. 知识图谱的行业应用与未来发展[J]. 互联网经济, 2018(4): 16-21.