文章编号：2095-2163(2021)12-0022-06 中图分类号：TP183 文献标志码：B

基于多特征融合的羊养殖问句相似度评价方法

李茂胜，王天一

（贵州大学 大数据与信息工程学院，贵阳 550025）

摘 要：问句相似度算法是常见问题集(Frequently Asked Questions,FAQ)问答系统的核心,本文旨在对问句相似度算法进行改良,提高羊养殖 FAQ 问答系统的准确率。针对于此,本文提出了一种基于多特征融合的相似度计算方法。该方法提取几种常见的自然语言处理特征,以及 3 个改良深度学习模型提取的特征,采用集成学习模型堆叠(stacking)处理这些特征,训练一个分类器,对问句对进行相似判断。通过相关书籍及爬虫,构建了一个 72 660 对的养羊问句对数据集进行实验。实验证明,该相似度算法能够有效提高羊养殖 FAQ 问答系统的准确率,并且达到了 98.8%。

关键词：问答系统；羊养殖；多特征融合；深度学习；模型堆叠

A similarity evaluation method for sheep breeding questions based on multi-feature fusion

LI Maosheng, WANG Tianyi

(College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

【Abstract】The similarity algorithm is the core of the common problem set question and answer system. In this paper, we aim at improving the question similarity algorithm and the accuracy of the sheep breeding FAQ Q & A system. This paper proposes a similarity calculation method based on multi-feature fusion. This method extracts several common natural language processing features and the features extracted by three improved deep learning models. Finally, these features are trained with integrated. Through the related book and crawler, 72660 pairs of sheep questions are collected to form a dataset used for experiment. Experiments shows that the similarity algorithm can effectively improve the accuracy of the sheep breeding FAQ Q & A system and reach 98.8%.

【Key words】question answering system; sheep farming; multi-feature fusion; deep learning; model of stack

0 引 言

进入互联网时代以来,各行各业的数据呈指数级增长,面对这些海量的数据,以往的搜索引擎[1]已逐渐不能满足用户准确、快速获取信息的需求。传统搜索引擎主要存在以下几点不足：

(1)通过关键字匹配进行搜索,导致返回的信息多且杂乱,有太多无关的信息被搜索引擎获取。

(2)没有对用户输入的句子进行句法、语法、语义方面的分析,只是简单地模糊搜索(同义词搜索),考虑的因素太少,影响搜索的准确率。

(3)由于用户界面的目录一般是按照类别进行分类,用户需要逐级进行搜索才能找到最终的类别和结果,导致速度较慢,耗费了大量时间。

面对以上传统搜索引擎的不足,问答系统[2]应运而生。问答系统是一种高效、智能的信息检索系统。用户可以使用自然语言来进行输入,系统通过相关的自然语言处理技术,对用户的输入进行分析,通过对数据库中的问句进行匹配,最终返还给用户一个简洁、准确的答案。问句相似度计算方法是 FAQ 问答系统的核心,很大程度上决定问答系统的好坏。问句相似度计算方法主要分为以下 3 种[3],分别是基于字符串的方法、基于知识库的方法和基于统计的方法。

基于字符串的方法主要是对输入的句子或组成句子词语的重复度、长度、词序等直接进行比较[4-5]。主要方法有编辑距离、最长公共子序列算法、N-Gram 模型、Jaccard 系数等。基于知识库的方法主要分为两类:一是基于结构化词典来衡量词语间的相似度,从而计算句子的相似度。常见的词典有《知网》(HowNet)、《同义词词林》、《WordNet》等,另一类则是通过网络知识的方法。该方法主要是利

基金项目：贵州省科学技术基金(黔科合基础-ZK[2021]一般 304);贵州省科学技术基金(黔科合平台人才(2018)5616)。

作者简介：李茂胜(1997-),男,硕士研究生,主要研究方向：自然语言处理；王天一(1989-),男,博士,副教授,硕士生导师,主要研究方向：量子通信、大数据与人工智能。

收稿日期：2021-09-14

用维基百科、百度、搜狗等网络知识库资源，通过网页链接、内容进行相似度计算。基于统计的方法在近年来取得了重要进展，该方法假设一个文本的语义只与其组成的词语有关，与上下文、词序、语法及句子结构无关。通过把词语映射为向量，把句子用向量来表示，计算两个句子向量间的距离来表示句子的相似度。基于统计的方法主要分为向量空间模型（如 VSM[6]）和主题模型（如 LSA[7]、PLSA[8]、LDA[9]）。近年来，随着深度学习的高速发展，通过神经网络训练词向量成为一种主流方法。一些词向量生成工具，如 word2vec[10]、Glove[11]、FastText[12]等，逐渐成为词向量预训练手段。基于神经网络的词向量相比于传统的哑编（one-hot）有以下优点：

（1）词向量维度明显降低。把词语用固定长度的向量来表示，大大减少了内存和计算量。

（2）基于神经网络的词向量能够很好地表示词语之间的语义。

由于单一的相似度计算方法主要是针对某一方向进行优化，均存在一定缺点。许多学者融合多个相似度计算方法，对算法进行改良，取得了不错的效果[13-14]。本文通过对多种自然语言处理（Natural Language Processing，NLP）常见特征以及 3 种改良的深度学习模块提取的特征进行融合，设计了一种问句相似度模块，并运用在羊养殖 FAQ 问答系统中。该问答系统将为养羊户提供简洁、具体的答案，相信通过该问答系统，能够解决养羊户在养殖过程中的常见问题。

1　构建数据集及训练词向量

1.1　问句对数据集构建

本文实验所使用的问句对来源于《山区肉羊高效养殖问答》、《农区科学养羊技术问答》、《现代羊病防制实战技术问答》等羊养殖相关问答书籍。经过整理，得到 2 500 个养羊基础问句对。在这 2 500 个问句对的基础上，用 Python 语言编写对应“百度问答”网站的爬虫代码，对这 2 500 个养羊基础问句分词后的句子进行爬虫。每个养羊基础问句爬虫了 10 个相似的问句，经过人工筛选并剔除语义无关的语句后，形成包含 2 500 组、11 000 个问句的数据集，每组相似句子 2~8 个不等。数据集实现步骤如下：

Step 1　根据羊养殖相关问答书籍整理得到 2 500个养羊基础问句。

Step 2　对 2 500 个养羊基础问句分词后进行爬虫后，人工筛选并剔除与语义无关的语句，得到 2 500组共 11 000 个问句，每组相似的句子 2~8 个不等。

Step 3　在 2 500 组问句中，每次选取 1 组问句 $Q_i = \{q_{i1}, q_{i2}, \cdots, q_{ik}\}$。

Step 4　由步骤 2 可知，Q_i 中 $q_{i1}, q_{i2}, \cdots, q_{ik}$ 都相似，因此任取 2 个问句 q_{i1}、q_{i2} 就可构成一个相似问句对。只需取尽 Q_i 中 2 个句子的组合就得到了 Q_i 全部的相似问答对。对于不相似问句对，由于各组句子都不相似，所以只须从 Q_i 中任取一个句子和其他组句子任意组合就可构成。最后，把得到的全部问答对存入数据集中。

Step 5　重复上述步骤，对 2 500 组进行处理，则可得到全部的数据集。

通过以上处理可以得到 72 260 个问句对。其中相似问句对和不相似问句对的比例为 1∶1，数目均为 36 130，避免在深度学习训练中造成数据平衡性对相似和不相似的权重的影响。生成的数据示例样本见表 1：其中 index 为句子对的编号，s1 和 s2 分别代表句子对的句子 1 和句子 2，label 则表示句子对的相似度，1 表示相似，0 表示不相似。

表 1　数据集样本

Tab. 1　Dataset sample

问句 1	问句 2	相似度标签
羔羊大肠杆菌病有什么特征？	羔羊大肠杆菌病症状有哪些？	1
怎样保存鲜奶才能不变质？	保存鲜奶的方法有哪些？	1
山羊口炎发病是什么原因？	绒山羊饲养应掌握哪些技术？	0
国外毛用山羊品种有几种？	山羊口炎发病是什么原因？	0

1.2　问句对数据集的词向量生成

在 NLP 任务中，需要将人类语言建模为向量的形式，这一过程首先需要对句子进行分词处理，再把词语嵌入到向量空间中，即词嵌入。本文使用 jieba 作为分词工具，并把养羊相关领域词语加入自定义词典，这样可以正确切分“农膜暖棚式”、“青绿饲料”等羊养殖专业词汇。经过 jieba 分词前后的样本见表 2。

表 2　问句分词结果

Tab. 2　Question partition results

处理前的问句	处理后的问句
羊场选址的基本条件是什么？	羊场选址 的 基本条件 是 什么
怎样防止怀孕母羊流产？	怎样 防止 怀孕 母羊 流产
怎样防止羔羊白肌病？	怎样 防止 羔羊 白肌病

分词后的句子向量化，使用谷歌开源的训练词

向量工具 word2vec 完成,将 jieba 切分后的句子向量化,作为深度学习特征提取模块的输入。本文采用连续词袋模型,通过对上下文单词,来预测当前单词出现的概率。词向量维度设置为 300,窗口大小设置为 5,通过养羊相关的领域知识库进行词向量训练。

2 特征提取及最终特征融合模型

2.1 多特征融合模型结构

本文提出了一种基于多特征融合的相似度计算方法。该方法的实现过程是输入问句对,转化为词向量矩阵,然后进行特征提取,最后通过 stacking 用来训练一个分类器,对问句对进行分类。多特征融合模型结构总体框架如图 1 所示。其中,特征提取是其中的核心部分,主要分为两个方向:一是直接对数据集提取一些常见的 NLP 特征,如编辑距离、n-gram 相似性等 。二是通过神经网络模型计算问句对的相似度作为特征。主要通过 3 个深度学习模块(曼哈顿相似度、注意力机制相似度、比较-聚合相似度)分别提取特征。

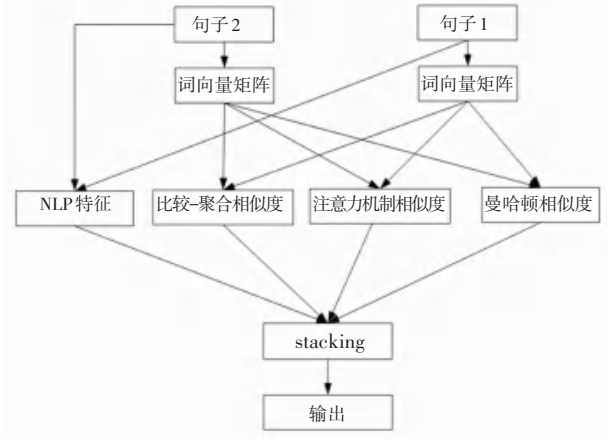

图 1 多特征融合模型结构

Fig. 1 General structure of multi feature fusion model

2.2 NLP 数据特征

NLP 数据特征主要是通过编写相对应的代码,提取问句对常见特征。本文提取了 10 个 NLP 特征,并把提取的特征用列的形式保存,构成 10 维向量,作为多特征融合模型输入数据的一部分。本文提取的常见特征如下:

(1)长度上的不同:计算两个句子长度的差。

(2)编辑距离:指两个字符串 A、B, A 编辑为 B 所需的最少编辑次数。如果其编辑距离越小,则其越相似。反之,亦然。

(3)N-Gram 相似性:将文本里面的内容按照字节进行大小为 N 的滑动窗口操作,形成长度是 N 的字节片段序列。每一个字节片段称为 gram,对所有的 gram 的出现频度进行统计,并且按照事先设定好的阈值进行过滤,形成关键 gram 列表,也就是这个文本的向量特征空间。

(4)句子中词语个数的特征:将提取 6 种与词语个数相关的特征。如,两个句子中相同词的个数分别除句子的最大及最小以及平均个数作为特征;两个句子不同词的个数分别除以对应句子词语总个数;计算两个句子的交集除以并集得到杰卡德(Jaccard)相似度等等。

(5)两个语句词向量组合的相似度:主要利用经 word2vec 训练的词向量对两个句子进行向量表示,再计算两个语句词向量组合的余弦相似度。

2.3 神经网络特征

在介绍深度学习相似度模块之前,先对后面用到的孪生神经网络(Siamese network)和长短时神经网络(Long Short-Term Memory, LSTM)进行简单介绍。

孪生神经网络也叫做“连体的神经网络”。神经网络的“连体”是通过共享权值来实现孪生神经网络。首先将两个输入分别输入两个神经网络中,再通过这两个神经网络分别将输入映射到新的空间,形成输入在新的空间中的表示,最后计算 Loss,评价两个输入的相似度。

循环神经网络(RNN),具有记忆的功能,能够很好地处理和预测时间序列,适用于文本处理。但伴随着序列输入越长,越可能导致梯度消失和梯度爆炸问题。而 RNN 网络的变种 LSTM ,通过特别构建的门结构,对细胞状态进行删除或添加信息,就能较好地解决这两个问题。

2.3.1 曼哈顿距离相似度模型(siamese-lstm-mandist, SLM)

曼哈顿距离相似度模型是以孪生神经网络和 LSTM 为基础构建的相似度模型[13]。由输入层、嵌入层、LSTM 层、自定义的曼哈顿层和输出层 5 部分组成,模型结构如图 2 所示。输入层把每个词语在词典中得到一个词语编号序列传递给嵌入层;嵌入层则将各个词语编号映射为 word2vec 词向量后,作为 LSTM 层的输入。两个 LSTM 开始对句子 1、2 的词向量进行学习,两个 LSTM 层彼此共享权重;将 LSTM 学习后的向量输入曼哈顿层,计算曼哈顿距离;再通过 dropout 来防止过拟合,以及归一化来提高计算速度;最后在输出层通过 *softmax* 函数输出结果。

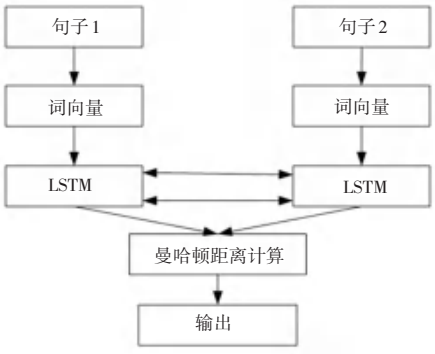

图 2　曼哈顿距离相似度模型结构图

Fig. 2　Manhattan distance similarity model structure

2.3.2　注意力机制相似度模型（Siamese－lstm－attention，SLA）

注意力机制相似度模型是以 LSTM 和注意力机制（Attention）为基础构建的相似度模型[17]。由输入层、嵌入层、LSTM 层、注意力层和输出层 5 部分组成，模型结构如图 3 所示。输入层、嵌入层、LSTM 层的功能同上个模型。其不同之处在于，LSTM 学习到的向量加上注意力层，对 LSTM 输入的向量分配不同的权重，进行选择性输入，更好地表示句子的语义。再通过全连接层降低维度，dropout 来防止过拟合，归一化来加速收敛速度；最后在输出层输出问句对的相似结果。

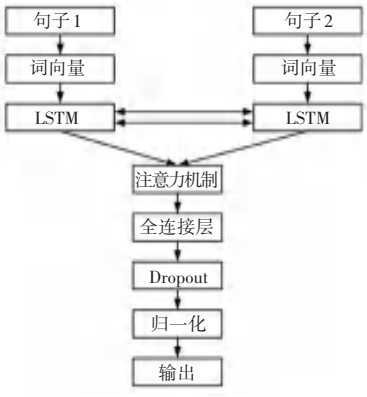

图 3　注意力机制相似度模型结构图

Fig. 3　Structure of attention mechanism similarity model

2.3.3　改进的比较-聚合相似度模型（Approve-Compare-Aggregate-Model，ACAM）

该模型基于 Compare－Aggregate－Model[15] 进行了改进，模型结构如图 4 所示。该模型主要做了 3 部分改进。

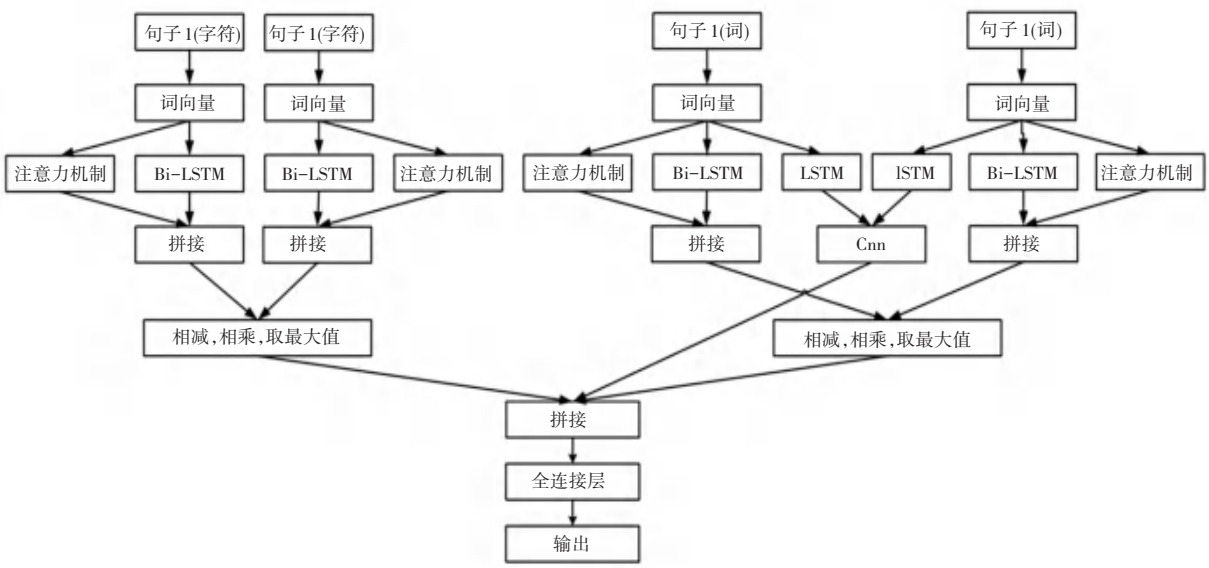

图 4　改进的比较-聚合相似度模型结构图

Fig. 4　Structure of improved comparison-aggregation similarity model

第一部分：在左边词向量输入的情况下，右边添加了字符向量作为输入。主要是为了提取词语内字间的信息，以及对超出词典的词进行表示。

第二部分：对 embedding 层分别用注意力机制和 bi-lstm 对向量分别表示，然后把这两个向量进行拼接。加入注意力机制后使得语句向量在词上有了重心，分配的权重不同。

第三部分：对前面处理过的句子向量进行更多的交互处理，对输入的两个句子向量进行相乘、相减、取最大值等操作。通过对句子向量进行这几种交互处理能够更好地比较句子之间的语义，提高模型准确率。

最后,模型把交互处理后的向量和前面 LSTM 和 CNN 网络提取的向量进行拼接,并通过全连接层进行降维,最后通过输出层进行输出。

2.4 stacking 建立分类模型

2.4.1 构建多个特征组成的数据集

通过代码提取了 NLP 数据特征,以及通过 3 个不同深度的学习网络提取了 SLM 特征、SLA 特征、ACAM 相似度特征。把提取的 12 个特征组合在一起构建一个 12 维的向量数据集,用来训练一个机器学习的分类模型。

2.4.2 构建分类模型

构建的分类模型主要是基于 Sklearn 库的常见分类方法,通过 stacking 算法进行融合得到。主要把分类器分为了两级:第一级别的分类器(初级学习器)有随机森林、朴素贝叶斯、支持向量机、逻辑回归,第二级分类器(次级学习器)为决策树。前面多个模型提取的特征所组成的数据集输入初级学习器分别进行训练,并将训练后所得到的结果作为次级学习器的输入,最后通过次级学习器进行分类以后,输出最终结果。总体 Stacking 计算过程如图 5 所示。

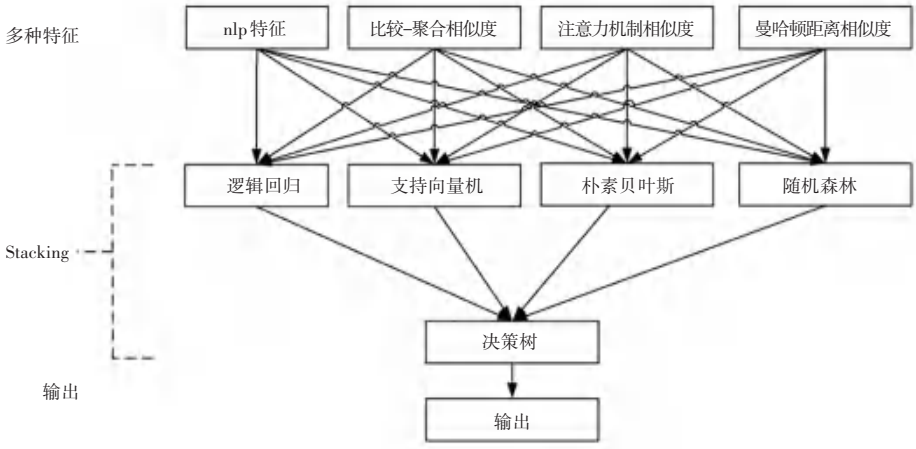

图 5 Stacking 计算结构图

Fig. 5 Stacking calculation structure

3 实例验证

实验环境设置为:Intel(R) Core(TM) i5-10200H CPU 8 核处理器;GPU 为 RTX 2060,运行内存 16 G。

实验通过对 72 106 个问句对划分为训练集、验证集,比例为 8∶2。

3.1 评测标准

本次实验通过精确度(*Accuracy*)、查准率(*Precise*)、查全率(*Recall*)、*F*1 分数这 4 个评测标准来衡量模型的性能,评测标准的具体公式见表 3。

其中,精确度(*Accuracy*)表示预测符合标准的样本与总样本的比例;查准率(*Precise*)表示正确预测正样本占实际预测为正样本的比例;查全率(*Recall*)表示正确预测正样本占正样本的比例。*F*1 分数是分类问题的一个衡量指标,其是 *Accuracy* 和 *Recall* 的调和平均数,能更好地衡量分类的好坏。所以在机器学习竞赛中,*F*1 常常作为最终测评的方法。

表 3 评测标准公式

Tab. 3 Evaluation metrics

	精确度(*Accuracy*)	查准率(*Precise*)	查全率(*Recall*)	*F*1 分数
表达式	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$	$Precise = \frac{TP}{TP + FP}$	$Recall = \frac{TP}{TP + FN}$	$F1 = 2 \cdot \frac{Precise \cdot Recall}{Precise + Recall}$

表 3 中 *TP*、*FN*、*FP*、*TN* 的含义见表 4。*TP* 表示正确地把正样本预测为正;*FN* 表示错误地把正样本预测为负;*FP* 表示错误地把负样本预测为正;*TN* 表示正确地把负样本预测为负。在机器学习中尤其是统计分类中,通过混淆矩阵能够很容易地看到机器学习是否混淆了样本的类别。矩阵的每一列表达了分类器对于样本的类别预测,矩阵的每一行表达了样本所属真实类别。

表 4　混淆矩阵

Tab. 4　Confusion matrix

	积极(Positive)	消极(Negative)
正确(True)	*TP*	*FP*
错误(False)	*FN*	*TN*

3.2　实验结果分析

4 种方法在养羊验证集上训练的结果见表 5。可以看出，本文方法是把前面提取的 NLP 特征及几种深度学习提取的特征作为集成学习 stacking 的输入，通过训练一个分类器，得到结果。相较前面 3 个方法，本文方法的各项评价指标都接近 99%，在 4 个评价指标上具有最优的综合性能，表明该方法能够很好地计算问句相似度，改善羊养殖问答系统的性能。但是，由于 stacking 方法需要计算前面多种模型的特征作为输入的数据，在时间效率上的对比，本文方法效率较低，还有值得改善的地方。在后续设计羊养殖问答系统时，可以结合问句分类模型，减少问答系统在相似度计算模块的匹配数量，以弥补本方法效率较低的问题。

表 5　多种相似度计算方法结果及时间效率

Tab. 5　Results of multiple similarity calculation methods

方法	精确度	查准率	查全率	F1 分数	测试组数	耗时/s
SLM	0.971	0.997 1	0.971	0.971	14 452	11.744
SLA	0.971	0.960	0.988	0.974	14 452	11.608
ACAM	0.983	0.979	0.990	0.984	14 452	71.283
本文方法	0.988	0.988	0.987	0.987	14 452	82.356

4　结束语

本研究通过对常用的 NLP 特征和 3 种深度学习方法进行特征提取分类，能够很好地考虑多种特征的情况，相比于提取单一特征的方法，该方法通过对多种方法提取的特征进一步分类，明显提升了 4 种评价指标，且在验证集上各评价指标都接近 99%。因此，本文提出的方法明显优于前 2 种方法，对比第 3 种方法也有了一定提升。

基于该相似度计算方法的羊养殖问答系统，能够准确匹配用户的问题，提供对应答案。相信这个羊养殖问答系统能够帮助养殖户解决许多养殖问题，能够促进养羊业更好的发展。

但本文的工作还有些不足，主要是在创建养羊数据集时，对于爬虫得到的相似问句需要人工进行判断，费时费力，且具有一定的主观性。下一步的工作会围绕如何自动或半自动构建知识库；或者对基于知识图谱的问答系统进行设计。

参考文献

[1] 黄际洲，孙雅铭，王海峰，等. 面向搜索引擎的实体推荐综述[J]. 计算机学报，2019，42(7)：1467-1494.

[2] 王瑛，何启涛. 智能问答系统研究[J]. 电子技术与软件工程，2019(5)：174-175.

[3] 韩程程，李磊，刘婷婷，等. 语义文本相似度计算方法[J]. 华东师范大学学报(自然科学版)，2020(5)：95-112.

[4] 钱丽萍，汪立东. 基于中心短语及权值的相似度计算[J]. 郑州大学学报(理学版)，2007(2)：149-152.

[5] 杨思春. 一种改进的句子相似度计算模型[J]. 电子科技大学学报，2006(6)：956-959.

[6] SALTON G，WONG A，YANG C S，et al. A vector space model for automatic indexing [J]. Communications of The ACM，1975，18(11)：613-620.

[7] LANDAUER T K，DUMAIS S T. A solution to Plato′s problem：The latent semantic analysis theory of acquisition，induction，and representation of knowledge [J]. Psychological Review，1997，104(2)：211-240.

[8] HOFMANN T. Probabilistic latent semantic analysis [J]. Uncertainty in Artificial Intelligence，1999，15(6)：289-296.

[9] BLEI D M，NG A Y，JORDAN M I，et al. Latent dirichlet allocation [J]. Journal of Machine Learning Research，2012(3)：993-1022.

[10] 石凤贵. 基于自然语言处理的 Word2Vec 词向量应用[J]. 黑河学院学报，2020，11(7)：173-177.

[11] 吉久明，施陈炜，李楠，等. 基于 GloVe 词向量的“技术——应用”发现研究[J]. 现代情报，2019，39(4)：13-22.

[12] 阴爱英，吴运兵，郑一江，等. 基于 fastText 模型的词向量表示改进算法[J]. 福州大学学报(自然科学版)，2019，47(3)：314-319.

[13] 梁敬东，崔丙剑，姜海燕，等. 基于 word2vec 和 LSTM 的句子相似度计算及其在水稻 FAQ 问答系统中的应用[J]. 南京农业大学学报，2018，41(5)：946-953.

[14] 张秀华，云红艳，贺英，等. 基于注意力机制的新闻事件检测研究与应用[J]. 计算机与数字工程，2021，49(6)：1143-1147，1280.

[15] WANG S，JIANG J. A compare-aggregate model for matching text sequences[J]. arXiv preprint arXiv：1611.01747，2016.