文章编号：2096-1472(2022)-11-39-05
DOI:10.19644/j.cnki.issn2096-1472.2022.011.009

基于循环神经网络和余弦相似度算法的智能客服机器人研究

蔡发群

(南京科技职业学院，江苏 南京 210048)
✉295463913@qq.com

摘　要：针对客服机器人答非所问的情况，提出一种结合循环神经网络学习算法LSTM(Long Short-Term Memory)、词频-逆文档频率(Term Frequency-Inverse Document Frequency，TF-IDF)算法及余弦相似度算法的客服机器人设计方法。LSTM算法利用长短记忆法更有利于联系上下文进行分词，分词准确率更高。TF-IDF算法可以将非结构化的客户提问和问题库问题用结构化的向量表示出来。通过余弦相似度算法对客户提问标签和问题库标签进行匹配，可以将最优答复提交给客户。试验结果显示，客户提问与问题A的余弦相似度值只有0.52左右，而与问题B的余弦相似度值达0.81，因此可以很好地实现答复推荐。

关键词：循环神经网络；LSTM；TF-IDF；标签；余弦相似度
中图分类号：TP183　**文献标识码**：A

Research on Intelligent Customer Service Robot based on Recurrent Neural Network and Cosine Similarity Algorithm

CAI Faqun

(*Nanjing Polytechnic Institute Nanjing*, *Jiangsu* 210048, *China*)
✉295463913@qq.com

Abstract: In view of the fact that the customer service robot does not always give a relevant answer, this paper proposes a design method of customer service robot that combines recurrent neural network learning algorithm LSTM (Long Short-Term Memory), TF-IDF (Term Frequency-Inverse Document Frequency) algorithm and cosine similarity algorithm. LSTM algorithm with long and short memory method is more conducive to word segmentation from the context, which leads to higher accuracy. TF-IDF algorithm can express unstructured customer questions and question library questions with structured vectors. The cosine similarity algorithm is used to match the customer's question tag and the question library tag, and the optimal reply can be submitted to the customer. Test results show that the cosine similarity value between the customer's question and question A is only about 0.52, while the similarity value with question B is up to 0.81. Therefore, reply recommendation can be well realized.

Keywords: recurrent neural network; LSTM; TF-IDF; tag; cosine similarity

1 引言(Introduction)

随着互联网的发展，购物方式发生了巨变。电子商务已成为主流商务形式，越来越多的商家实施线上交易。电商渠道的商家与客户交流的主要纽带就是客服。为了能够提高客户的满意度，促成交易，商家不得不雇佣大量的客服人员。这就导致市场上客服人员供不应求；另一方面，由于工作压力大，晋升空间有限，所以，客服人员流失严重，服务质量难以保证。人工客服招聘难，流动大，成本高，促使客服机

基金项目：江苏省高校哲学社会科学研究项目“SNS对高职学生社会认知能力影响机制研究——基于‘三元交互决定论’”(2019SJA0631)；江苏省大学生创新创业训练项目“校园优聘”(202112920029Y)；南京科技职业学院人文社科项目“基于SIVORS模型社交媒体信息传播规律研究”(NJPI-2021-ZD-06).

器人应运而生。部分企业也积极上马客服机器人，取代人工客服。然而，客服机器人技术尚不成熟，很多店铺的客服机器人不仅不能很好地解决客户的疑虑，提高客户满意度，反而降低了客户的满意度，甚至答非所问，让客户抓狂。因此，如何构建智能客服机器人，实现精准解答客户疑问和投诉成为客服机器人主攻研究方向。

2 文献综述(Literature review)

1966 年世界上第一个对话机器人ELIZA诞生于麻省理工学院。1995 年Richard S．Wallace博士开发了ALICE系统[1]。2001 年，机器人SmarterChild上线。2008 年，整合了众多网络服务功能的苹果Siri上架。2015 年，京东JIMI正式接入。用户可以通过与JIMI聊天了解商品的具体信息以及反馈购物中存在的问题等。JIMI知道自己不能回答用户的哪些问题，并且知道何时应该转向人工客服。

对话机器人以检索式为主，必须由一个或若干个问答数据库作为支撑。用户提出问题时，系统通过检索、匹配技术从数据库中找出答案来进行回复。因此，数据库的质量对答复的有效性和准确性有很大的影响。然而，数据库规模再大，质量再好，也不可能实现永远完美答复，因为对话的场景日新月异，人们的表达方式也随着时间不断变化，这些变化都需要对数据库进行更新和补充[2]。近年来，深度学习方法在自然语言、语音、图像、视频等领域发展迅猛[3]。使用图神经网络构建推荐算法逐渐成为当今推荐算法的研究热潮[4]。这个方法恰好可以解决客服机器人自动更新和扩充数据库的需求。

3 基于循环神经网络学习算法的智能机器人答复推荐系统(Intelligent robot reply recommendation system based on recurrent neural network learning algorithm)

客服机器人给出准确答案的过程主要包括三个步骤。第一步，数据采集。数据采集包括客户基本信息、客户行为信息、客户提问信息以及常见客户历史问题及对应回复信息等。第二步，数据处理。数据处理主要是进行客户提问标签提取。首先，对客户当前提问进行分词处理，抽取出可以代表其含义的属性词，用具体的词向量来表示客户的抽象提问，每个词对应的权重由信息检索算法TF−IDF来决定，然后利用客户购物、浏览、咨询及评价等历史信息的特征数据，以及客户的年龄、性别、地域等来分析出此用户的提问特征，最终提取出客户提问标签。第三步，答复推荐。通过对比客户提问标签和数据库中问题标签，为用户推荐相关性最大的回复信息。具体步骤如图1所示。

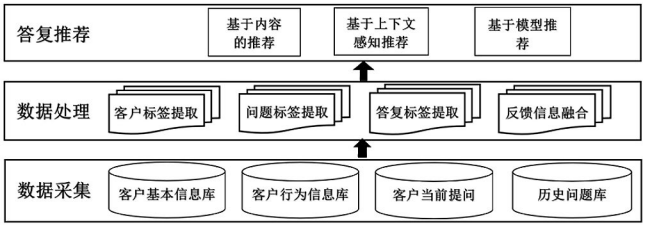

图1 基于循环神经网络学习算法的智能机器人答复推荐系统

Fig.1 Intelligent robot reply recommendation system based on recurrent neural network learning algorithm

3.1 数据采集

客服机器人回复客户提问的主要依据，除了客户当下提问内容外，还要考虑客户年龄、性别、偏好、地域、受教育程度、人口统计学信息，用户的浏览、购买、评价行为信息等。因此，客服机器人必须先采集相关信息。

客户提问可以通过相关入口直接录入。而客户年龄、性别、偏好、地域、受教育程度、人口统计学信息则可以在客户注册时录入到客户基本信息库，需要时再在客户基本信息库中提取。用户的浏览、购买、评价等行为信息可以在客户购物过程中录入到客户行为信息库中。

3.2 数据处理

对于客户的提问，首先必须进行分词处理，将非结构化的客户提问分成一个个词语，并对每个词语赋予权重，然后结合客户的基本信息和行为信息，构成客户提问向量，从而获取客户提问标签。问题库中的问题也用同样的方法设置标签。最后将客户提问标签与问题库中的问题标签进行对比，从而确定如何答复客户。

3.2.1 基于LSTM模型客户提问分词处理

要想实现精准回答，智能客服机器人必须能够分析客户提问和问题库中问题的语义。国内客服机器人主要服务于中国人，因此，必须要先学会中文分词。中文分词作为中文自然语言处理领域的重要基础研究，近些年来很多专家学者致力于该领域的研究，研究方法主要分为三种：(a)基于规则的方法；(b)基于传统机器学习模型的方法；(c)基于深度神经网络模型的方法[5]。深度神经网络模型基础算法是自然语言处理算法(Natural Language Processing，NLP)，包括词法分析、句法分析、语义分析、文档分析等。

中文分词是一个经典的预测序列问题，方法主要有基于隐含马尔柯夫模型(Hidden Markov Model，HMM)分词法、基于条件随机场(Conditional Random Field，CRF)分词法和基于LSTM的分词方法。Hochreiter等人在1977 年又

提出LSTM模型，LSTM是循环神经网络(Recurrent Neural Network，RNN)的一个变种，它克服了RNN模型训练过程中梯度消失和梯度爆炸的缺陷。LSTM模型的总体结构如图2所示。

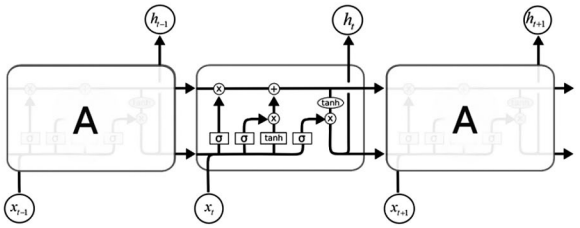

图2 LSTM模型总体结构

Fig.2 Overall structure of LSTM model

LSTM在循环单元内部引入了门限(gate)结构[6]。LSTM模型中有三个“门”，分别是“遗忘门”“输入门”和“输出门”，分别用于决定哪些信息要舍弃，哪些信息要留下，最终输出哪些信息。如图3所示，t时刻的相关信息受$t-1$时刻的信息影响，同时又影响着$t+1$时刻，从而保持持续性。

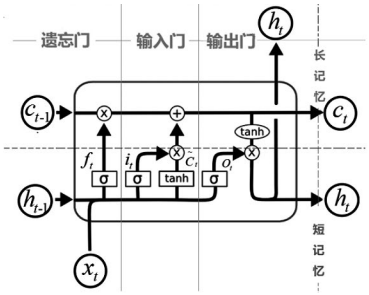

图3 LSTM模型详解图

Fig.3 LSTM model details

LSTM的第一个门是遗忘门(forget gate)，确定丢弃哪些信息。如图3所示，f_t是一个概率值，$f_t=\sigma(W_f\times[h_{t-1},x_t]+b_f)$，决定哪些信息应该被丢弃。其中，$h_{t-1}$表示历史隐藏状态值，$C$即cell state,表示细胞状态值，$C_{t-1}$表示细胞历史状态值，用来刻画神经元记忆中不容易衰减的部分，构造长期记忆。x_t是当前时刻的输入值，σ是激活函数，f_t接近0时表示细胞历史状态值C_{t-1}被遗忘，f_t在0−1时C_{t-1}被部分保留，f_t等于1时表示C_{t-1}被完整保留下来。$f_t\times C_{t-1}$表示长记忆保留部分。因此，遗忘门又叫“长期记忆加权模块”，其实就是对长期记忆的选择性记忆，如图3所示。

第二个门是输入门(input gate)，用来确定存储哪些新信息。如图3所示，输入门主要包括两个层，首先是σ(sigmoid)层，用来决定保留哪些原有的值，由i_t来决定，$i_t=\sigma(W_i[h_{t-1},x_t]+b_i)$；其次是tanh层，用来决定添加哪些新信息，由$\tilde{C}_t$决定，$\tilde{C}_t=\tanh(W_C\cdot[h_{t-1},x_t]+b_C)$。$i_t\times\tilde{C}_t$表示短记忆更新部分，如图3所示。

长记忆保留部分和短记忆更新部分共同作用，将原细胞状态C_{t-1}更新为新细胞状态C_t。$C_t=f_t\times C_{t-1}+i_t\times\tilde{C}_t$，就是原单元状态$C_{t-1}$乘以$f_t$，再加上$i_t$乘以$\tilde{C}_t$，从而实现记忆的连续性。

第三个是输出门(output gate)，确定输出的内容。如图3所示，历史隐藏状态值h_{t-1}和当前时刻输入值x_t通过sigmoid函数计算出o_t，$o_t=\sigma(W_O[h_{t-1},x_t]+b_o)$，决定输出哪些部分。然后，进入tanh层，将$\tanh(C_t)$乘以$o_t$，计算输出值$h_t$，$h_t=o_t\times\tanh(C_t)$。

为了将中文分词问题转化成序列标注问题，就需要将分词中的每一个字进行标注。在深度学习分词研究工作中，常用的标注集是四词位(B、M、E、S)，分别表示一个分词的开始词位、中间词位、结束词位以及以单独一个字构成的分词[7]。中文分词中，LSTM记忆单元就是文本的上下文，再结合相关函数，推理出文本的分词构成。

3.2.2 问题向量空间模型

客服机器人研究的对象主要是客户的提问和问题库问题，都属于非结构化数据。一般用向量空间模型(Vector Space Model，VSM)将非结构化的问题结构化。

记问题库问题集合为$Q=\{q_1,q_2,\cdots,q_N\}$，将问题库中所有问题中出现过的词用集合$T=\{t_1,t_2,\cdots,t_n\}$表示，也就是问题库中有N个问题，而这些问题里包含了n个不同的词，构成词典库。这样我们就可以使用向量$q_j=(\omega_{1j},\omega_{2j},\cdots,\omega_{nj})$来表示第$j$个问题，其中$\omega_{kj}(k=1,2,\cdots,n)$表示第$k$个词$t_k$在问题$q_j$中的权重，$\omega_{kj}$值越大表示越重要，$\omega_{kj}$为0，表示词$t_k$没有出现在第$j$个问题$q_j$中。所以，为了表示第$j$个问题$q_j$，关键就是如何计算$q_j$各分量$\omega_{kj}$的值。

第j个问题中与词典里第k个词对应的TF−IDF值计算方法如下：

$$TF(t_k,q_j)=\frac{n_{kj}}{\sum n_{ij}} \tag{1}$$

$$IDF(t_k,q_j)=\lg\frac{N}{n_k+1} \tag{2}$$

$$TF-IDF(t_k,q_j)=\frac{n_{kj}}{\sum n_{ij}}\times\lg\frac{N}{n_k+1} \tag{3}$$

其中，n_{kj}是第k个词在问题j中出现的次数，$\sum n_{ij}$表示问题j中出现的词的总数，n_k表示所有问题中包括第k个词的问题数量。最终第k个词在问题j中的权重可由式(4)获得。

$$\omega_{kj}=\frac{TF-IDF(t_k,q_j)}{\sqrt{\sum_{s=1}^{n}TF-IDF(t_s,q_j)}} \tag{4}$$

这样就可以实现对所有客户可能提出问题归一化。可以让不同问题统一用向量表示，这些向量就可以作为对应答复的标签。

3.2.3 问题标签提取

客服机器人除了要考虑客户的当前提问，还需要考虑客户的购物经历、评价、浏览、以往提问等信息，以及年龄、性别、地域、受教育程度、地域等信息，还有人口统计学特征或行为特征等，也将作为向量维度。因此客户提问标签可以表示为

$$\boldsymbol{q}_j=\left(\omega_{1j},\omega_{2j},\cdots,\omega_{nj},\ \psi_{1j},\psi_{2j},\cdots,\psi_{mj},\xi_{1j},\xi_{2j},\cdots,\xi_{kj},\cdots\right) \quad (5)$$

3.3 答复推荐算法

目前搜索引擎、社交媒体等平台使用的推荐算法主要有Rocchio算法、决策树算法(Decision Tree，DT)、线性分类算法(Linear Classifer，LC)、朴素贝叶斯算法(Naive Bayes，NB)、余弦相似度算法(Cosine Similarity，CS)等。

Rocchio算法是一种匹配推荐算法，是处理反馈的著名算法，经常用于搜索引擎。往往会先向客户呈现各种可能结果。然后根据客户的点击情况，再缩小结果范围，慢慢精准化。这种方法显然不适合客服机器人。

决策树算法是一种逼近离散函数值法，是一种典型的分类方法。首先利用归纳算法生成可读的规则和决策树，然后对数据进行分析。决策树算法其实是通过一系列规则对数据进行分类。当项目属性较少而且是结构化属性时，决策树一般会是个好的选择。但是如果项目的属性较多，且都来源于非结构化数据，比如文章、问题等，那么决策树算法的效果就不理想了。

线性分类算法指用一个线性方程把待分类数据分开。对于二维的情况，就是用一条直线将数据分开。对于三维的情况，则是用一个超平面将数据区分开来。它的分类算法基于一个线性预测函数，决策的边界是平的，比如直线和平面。这种方法显然也不适合客服机器人。

朴素贝叶斯方法是在贝叶斯算法的基础上进行了简化。基于朴素贝叶斯算法进行分类的基本原理是假设特征之间互相独立[8]，既没有哪个属性对于决策结果影响比较大，也没有哪个属性对于决策结果影响比较小。这种方法简化了贝叶斯方法的复杂性，但是，在一定程度上降低了贝叶斯分类算法的分类效果。

余弦相似度算法是通过向量空间中两个向量夹角的余弦值作为衡量两个项目间相似性的方法。余弦值越接近1，角越接近0 度，表明两个向量越相似性高，等于1时表明两个向量完全相同，如图4所示。余弦值越接近−1，角越接近180 度，表明两个向量基本不相似，或相似度很低。两个向量之间的角度的余弦值确定两个向量的方向的一致性。余弦相似度通常用于正空间，因此给出的值为−1到1。

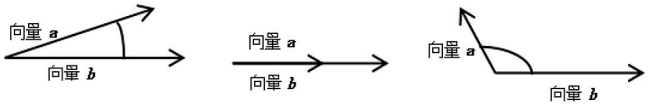

图4 余弦相似度算法原理图

Fig.4 Schematic diagram of cosine similarity algorithm

向量空间余弦相似度理论就是基于上述原理来计算个体相似度的。假设 $\boldsymbol{a}$ 向量(a_1,a_2)、$\boldsymbol{b}$ 向量(b_1,b_2)是二维向量，如图5所示，那么余弦定理的表达形式，即向量 $\boldsymbol{a}$ 和向量 $\boldsymbol{b}$ 夹角的余弦为

$$\cos\theta=\frac{\boldsymbol{a}\times\boldsymbol{b}}{\|\boldsymbol{a}\|\times\|\boldsymbol{b}\|}=\frac{(a_1,a_2)\times(b_1,b_2)}{\sqrt{a_1^2+a_2^2}\times\sqrt{b_1^2+b_2^2}}=\frac{a_1b_1+a_2b_2}{\sqrt{a_1^2+a_2^2}\times\sqrt{b_1^2+b_2^2}} \quad (6)$$

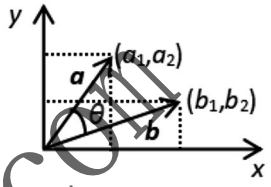

图5 二维向量夹角示意图

Fig.5 Schematic diagram of included angle of two-dimensional vector

当向量 $\boldsymbol{a}$ 和向量 $\boldsymbol{b}$ 为n 维向量时，则向量 $\boldsymbol{a}$ 与向量 $\boldsymbol{b}$ 夹角的余弦为

$$\cos\theta=\frac{\boldsymbol{a}\times\boldsymbol{b}}{\|\boldsymbol{a}\|\times\|\boldsymbol{b}\|}=\frac{(a_1,a_2,\cdots,a_n)\times(b_1,b_2,\cdots,b_n)}{\sqrt{a_1^2+a_2^2+\cdots+a_n^2}\times\sqrt{b_1^2+b_2^2+\cdots+b_n^2}}=\frac{\sum_{i=1}^{n}a_ib_i}{\sqrt{\sum_{i=1}^{n}a_i^2}\times\sqrt{\sum_{i=1}^{n}b_i^2}} \quad (7)$$

余弦相似度算法通常用于数据挖掘中的文本比较。智能客服机器人在回答客户问题时主要用到的就是文本比较。因此，余弦相似度算法比较适合客服机器人答复推荐。首先，基于LSTM网络学习将客户提问和问题库中的问题向量化，然后就可以利用余弦相似度算法比较客户提问和问题库中问题向量的相似度，从而确定提交给客户的答复信息。

假设某服装店铺的客户提问：皮肤黑选哪个颜色比较合适呢?

假设问题库中有9 个问题，其中只有问题A和问题B与客户提问有共有词。

问题A：皮肤白哪个颜色合适?

对应答复：亲，皮肤白穿什么颜色都好看哦，只是不同的颜色彰显的气质不同。粉色温柔、甜美，黄色明快、纯洁，红色热烈、有朝气，深蓝色低沉、神秘，紫色优美高雅、雍容华贵。亲可以根据自己的需要选哦。

问题B：皮肤黑哪个颜色合适?

对应答复：亲，您好，根据我们的经验，淡黄色、豆沙粉、米白色比较衬皮肤，显白哦!

在暂不考虑客户基本信息和行为信息的情况下，到底应该把哪个答复推荐给客户呢？基本思路是要看客户提问与问题库中问题的相似性，而两者之间的相似性取决于词向量的关系，因此，先对它们进行分词，然后计算权重，最后得到词向量，最后再利用余弦相似度算法计算客户提问与问题库中各问题的相似程度，取值最大的那个就是问题的答复。

第一步，对客户提问和问题库问题进行分词处理。

客户提问：皮肤/黑/选/哪个/颜色/比较/合适/呢?

问题A：皮肤/白/哪个/颜色/合适

问题B：皮肤/黑/哪个/颜色/合适

第二步，通过对人工客服语料库训练可知，语气词“呢”，形容词“比较”，以及“?”等在联系上下文的情况下可以忽略不计。因此，可以列出所有词：皮肤,黑,白,选,哪个,颜色,合适。

第三步，计算客户提问和问题库问题词频。

客户提问：皮肤1，黑1，白0，选1，哪个1，颜色1，合适1

问题A：皮肤1，黑0，白1，选0，哪个1，颜色1，合适1

问题B：皮肤1，黑1，白0，选0，哪个1，颜色1，合适1

第四步，给出客户提问和问题库问题词频向量，即它们的标签。

客户提问向量：[1,1,0,1,1,1,1]

问题A向量：[1,0,1,0,1,1,1]

问题B向量：[1,1,0,0,1,1,1]

根据TF−IDF算法可知，客户提问标签如下式：

$$\left[\frac{\frac{1}{6}\times(1-2\lg 2)}{\sqrt{\frac{1}{6}(6-9\lg 2-\lg 3)}},\frac{\frac{1}{6}\times(1-\lg 3)}{\sqrt{\frac{1}{6}(6-9\lg 2-\lg 3)}},0,\frac{\frac{1}{6}\times(1-\lg 2)}{\sqrt{\frac{1}{6}(6-9\lg 2-\lg 3)}},\right.$$
$$\left.\frac{\frac{1}{6}\times(1-2\lg 2)}{\sqrt{\frac{1}{6}(6-9\lg 2-\lg 3)}},\frac{\frac{1}{6}\times(1-2\lg 2)}{\sqrt{\frac{1}{6}(6-9\lg 2-\lg 3)}},\frac{\frac{1}{6}\times(1-2\lg 2)}{\sqrt{\frac{1}{6}(6-9\lg 2-\lg 3)}}\right] \tag{8}$$

因此，客户提问标签为[0.1,0.13,0,0.17,0.1,0.1,0.1]。

同理可知，问题A标签为[0.12,0,0.21,0,0.12,0.12,0.12]，问题B标签为[0.12,0.16,0,0,0.12,0.12,0.12]。

现在，问题比较就变成如何计算向量夹角余弦值的问题了。三个问题就像空间中的三条向量，都是从原点(0,0,0,0,0,0,0,0)出发，指向不同的方向。每两个向量之间形成一个夹角，夹角余弦值代表着两个向量方向的一致性，也就是提问和问题间的相似性程度。

根据余弦相似度算法，客户提问和问题A两个向量之间夹角θ的余弦值，如式(9)所示。

$$\cos\theta=\frac{\sum_{i=1}^{n}a_ib_i}{\sqrt{\sum_{i=1}^{n}a_i^2}\times\sqrt{\sum_{i=1}^{n}b_i^2}}\approx 0.52 \tag{9}$$

同理，客户提问和问题B两个向量之间夹角α的余弦值约为0.81。计算结果表明客户提问与问题B两个向量间的夹角的余弦值为0.81，远大于客户提问与问题A间的夹角的余弦值0.52。所以客户提问和问题B更相似，因此应该将问题B对应的答复提交给客户。当然从语义表达上可以看出，虽然问题A和问题B中没有“选”这个词，但加上有这两个词后语义是没有什么改变的。考虑到这一点，两个余弦值应该更高，相似性也会更明显些。

4　结论(Conclusion)

中国语言博大精深，有的一词多义，有的则是一义多词，因此，不同语境下分词结果可能会不同，不同的分词结果，向量表达也有可能相同。要想精确表达，还需进一步优化算法。本研究采用LSTM算法进行分词，还不能完全解决分词歧义，序列长度超过一定限度后，LSTM算法也会出现梯度消失的情况。另外，利用余弦相似度算法进行答复推荐也存在一定的局限性。余弦相似度算法只考虑向量的方向，不考虑其大小。因此，推荐答复时，可能还不够准确。智能机器人技术在不断发展中，相关算法在不断升级，后期还要进一步研究相关算法，以实现客服机器人精准答复。

参考文献(References)

[1] 田星.面向客服聊天机器人的领域本体构建的研究与应用[D].成都:电子科技大学,2018.

[2] 陈勤.在线教育客服数据挖掘与对话机器人设计[D].北京:北京交通大学,2018.

[3] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521(7553):436−444.

[4] 冯兴杰,生晓宇.基于图神经网络与深度学习的商品推荐算法[J].计算机应用研究,2021,38(12):3617−3622.

[5] 任智慧,徐浩煜,封松林,等.基于LSTM网络的序列标注中文分词法[J].计算机应用研究,2017,34(05):1321−1324,1341.

[6] 邢明磊.智能客服对话系统的设计与实现[D].北京:北京邮电大学,2020.

[7] 张子睿,刘云清.基于BI−LSTM−CRF模型的中文分词法[J].长春理工大学学报(自然科学版),2017,40(04):87−92.

[8] 马文,陈庚,李昕洁,等.基于朴素贝叶斯算法的中文评论分类[J].计算机应用,2021,41(S2):31−35.

作者简介：

蔡发群(1976−)，女，硕士，讲师.研究领域：电子商务，网络营销.