O2O教育系统中题目标签的智能挖掘算法

张雨思, 杨智, 代亚非

北京大学网络实验室分布式系统组，北京

Introduction

近年来，随着物质文化水平的提高，传统的教育模式已经逐渐不能满足人民的需要，而新兴的互联网+教育则日渐兴起，它可以跨越时间和距离的特点受到广泛的关注和青睐。阿凡题正是在这样的背景下诞生的一款专为中学生作业答疑定制的app，其采用O2O商业模式，将用户提问的题目分配给签约的老师作答。其现在的答题系统设计是用户将自己遇到的题目问题以图片的形式上传，阿凡题已经通过cnn和长短期记忆网络建立了比较成熟的图片OCR识别体系，它可以识别图片中的文字信息并将其以文本的形式呈现出来，命名为OCR_TEXT，然后通过OCR_TEXT对这道题目在题库中进行检索或者分配给在线老师作答。而无论是检索还是老师分配，题目学科都是重要的信息；与此同时，现阶段的答题系统不区分题目的难度，老师回答难题和简单题的报酬相同，导致系统成本上升，老师答题积极性下降。在此基础上，针对现有的答题系统的问题，设计并实现了相应的解决办法，需要对题目的标签信息进行挖掘。一是建立基于神经网络的可以判定题目科目的模型，提高题目的科目标注准确率；二是用词袋模型和逻辑斯蒂回归建立判断数学题目难度的分类器，制定基于题目难度的定价策略，将简单的题目挑出来分配给能力相对较低的老师并降低题目单价。以上改善可以显著的提升用户体验，并降低成本。

算法设计

学科标签挖掘：因为用户提问的题目都是通过图片的形式上传到阿凡题后台，并且但是OCR的识别结果常常不尽如人意，相较于普通文本具有以下的特点，一是OCR_TEXT经常会出现个别的字的错误，虽然不影响大的语义的理解，但是会导致很多关键词语被破坏。二是由于图片切分存在的问题，OCR_TEXT中的字或者词语的前后位置会出现偏差，词语先后的语义也会遭到破坏。在目前的针对短文本分类的方法中，无论是传统的自然语言处理常常采用的n-gram模型、k-means聚类等方法，还是最近业界推崇的word embedding 和LSTM的深度学习方法，基本上都是以文本的词汇为最小分析单元，这在面对完整且无误的文本时往往有较好的效果，但是如果把这些方法应用在OCR_TEXT，首先因为它包含许多错误的字，常常使得一个个词语破坏，所以模型在预测时遇到大量的未登录词，这相当于加剧了信息丢失的程度，其次因为OCR_TEXT中词语的位置关系未必准确，以上算法全都重视序列关系和位置信息，如LSTM网络每个词语的输出都与上个词语有关，因此在无法确保词语位置准确的前提下，算法难免会产生错误。综上可知以上提到的方法都不太适合用以OCR_TEXT作为输入判别学科。针对OCR_TEXT的特点，为了把序列信息丢失和关键词语信息丢失的损失减到最小，不再采用一个个的词语作为输入单元，而是将词语再细分化，拆成一个个的字、符号和标点等元素，并以此作为输入单元，并且忽略字与字之间的位置关系，仅仅以是否出现某个字作为题目的特征，采用基于字的前馈神经网络算法对OCR_TEXT进行文本分类。

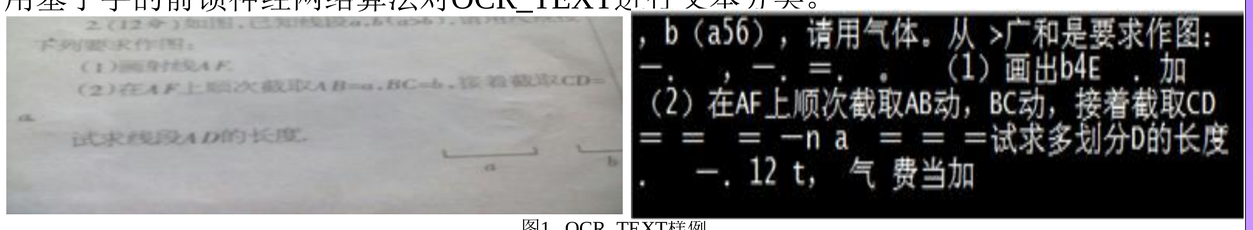

图1 OCR_TEXT样例.

首先，提取50万道题目的OCR_TEXT，统计所有出现的字以及它们出现的频次，共得到了7716个字，然后去除出现频次太少的字，保留4000个字作为评判单元，可以得到如图5所示的字表。用4000个神经元作为输入层，每一个神经元和一个字对应，若这个OCR_TEXT中包含这个字，则相应的神经元输入为1，否则为0，因此输入向量是一个4000维的0、1向量。

构建的网络里面共包含三个隐层，每一个隐层有400个节点，输入层与隐层之间以及隐层与隐层之间都采用全连接的连接方式。第三层隐层的激活函数采用sigmoid函数：

$$f(x)=\frac{1}{1+e^{-x}}$$

之后为了避免梯度消失问题，提高训练速度，第一层隐层和第二层隐层的激活函数采用rectifier函数：

$$f(x)=\begin{cases}0, & x<0\\ x, & x\geq 0\end{cases}$$

同时为了避免过拟合，隐层和隐层之间添加一个droupout层，这个dropout层在训练模型时以50%的概率丢弃相连的神经元之间的连接。可以将droupout看作是模型平均的一种形式，对于每次训练针对的样本，其对应的网络结构都是不同的，但是这些不同的网络结构层与层之间连接的权值都是共享的，可以说是一种极端的bagging，因此能够有效的防止过拟合。

输出层有9个节点，和最后一层隐层是全连接的关系，采用softmax函数作为其激活函数，可以将softmax函数看作是sigmoid函数在多维情况的扩展。对于m个标量x1，x2，…，xm来说，softmax函数有如下定义：

$$Z_k=\mathrm{softmax}(x)=\frac{e^{x_k}}{\sum_{i=1}^{m}e^{x_i}}$$

输出层接收来自最后一层隐层的输入，输出概率为：

$$h_\theta\left(x^{(i)}\right)=\begin{Bmatrix}p\left(y^{(i)}=1\middle|x^{(i)};\theta\right)\\ p\left(y^{(i)}=2\middle|x^{(i)};\theta\right)\\ \cdots\end{Bmatrix}=\frac{1}{\sum_{j=1}^{k}e^{\theta_j^T x^{(i)}}}\begin{Bmatrix}e^{\theta_1^T x^{(i)}}\\ e^{\theta_2^T x^{(i)}}\\ \cdots\end{Bmatrix}$$

最后的神经网络模型架构为：

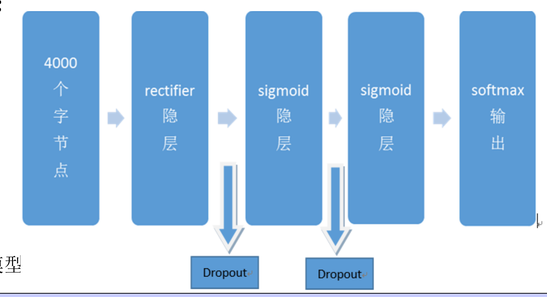

图2 神经网络科目判别模型

题目难度标注：给题库中的题目进行难度标注之后，用户问的题目的难度可以用题库中最相似的题目难度近似，因此和科目判别不同的是，难度判断的对象是完整的题目文本，不是有关键字丢失的OCR_TEXT，出于这个考虑，在分类方法上的选择有所不同，不再是以字为基本单元，而是以词语为基本单元，采用词袋模型和逻辑斯蒂回归的方法构造模型。

首先用信息增益来挑选可以区分难题和简单题目的词语构成词袋，信息增益是一种挑选特征的方法，它依据某项特征能为整个分类提供的信息量多少来衡量这个特征的重要程度，而信息量的多少由熵衡量。熵就是信息的混乱程度，其定义如下：若一个变量X取值可以使x1, x2, …, xn,每一种取到的概率分别是p1, p2,…p4,则熵为：

$$H(X)=-\sum P_i\cdot\log_2 P_i$$

在难度分类中，有词汇t和没有词汇t的时整个系统的熵的差值就是这个词汇的信息量。设类别为C，根据定义，词汇t的条件熵为：

$$H(C|T)=P_t*H(C|t)+P_{\bar{t}}*H(C|\bar{t})$$

于是词汇t的信息增益计算公式为：

$$IG(T)=H(C)-H(C|T)$$

具体到训练集的题目和词语中，先统计难题和简单题目的数目N1和N2,然后再统计每个词语在难题中相应字段出现的频率A，不出现的频率C，在简单题目中出现的频率B，不出现的频率D，则这个词语对于难度判断的信息增益为：

$$\begin{aligned}ig=&-\left(\frac{N1}{N1+N2}\log\frac{N1}{N1+N2}+\frac{N2}{N1+N2}\log\frac{N2}{N1+N2}\right)\\&+\frac{A+B}{N1+N2}\left(\frac{A}{A+B}\log\frac{A}{A+B}+\frac{B}{A+B}\log\frac{B}{A+B}\right)\\&+\frac{C+D}{N1+N2}\left(\frac{C}{C+D}\log\frac{C}{C+D}+\frac{D}{C+D}\log\frac{D}{C+D}\right)\end{aligned}$$

选择信息增益大的词语作为区分难题和简单题目的词袋。

除了文本词语维度，还计算了其他非NLP维度的皮尔逊相关系数，选择相关性较大维度作为特征加入模型

维度	相关性系数
题目长度	0.391979
分析长度	0.275219
解答长度	0.298809
题目中公式数量	0.046394
题目中公式的平均长度	0.046337
题目中最长的公式长度	0.031157
分析中公式的数量	0.158465
分析中公式的平均长度	0.089105
分析中最长的公式长度	0.038333
解答中公式的数量	0.102454
解答中公式的平均长度	0.045487
解答中最长的公式长度	0.030341

模型评测

对于科目判别的模型：随机抽选9个学科的用户提交题目的OCR_TEXT各50道，人工确定其其学科之后用模型进行测试，结果如下表：

模型 科目	语文	数学	英语	物理	化学	生物	政治	历史	地理	召回率
语文	46	1	1	1	0	0	1	0	0	92.00%
数学	0	49	0	0	0	0	1	0	0	98.00%
英语	0	0	50	0	0	0	0	0	0	100.00%
物理	0	0	0	49	1	0	0	0	0	98.00%
化学	0	0	0	1	48	0	1	0	0	96.00%
生物	0	0	0	0	1	49	0	0	0	98.00%
政治	2	1	0	0	0	0	46	1	0	92.00%
历史	1	0	0	0	0	0	1	48	0	96.00%
地理	2	0	0	1	1	1	0	2	43	86.00%
准确率	90.2%	96.1%	98.1%	94.6%	93.8%	98.0%	92.0%	94.1%	100%	

由评测结果可以看出，模型对于数学、外语、物理、生物、化学等学科的分类准确率较好，错误集中语文、历史、政治、地理等文科科目互相之间的误判，具体的错误原因在下文会有更详细的分析。对于科目判别的平均准确率为95.19%，平均召回率为95.11%。考虑到实际流量中数学占比很大，地理、历史等较少，系统中实际流量加权的正确率和召回率要高于平均数值，分别为96.39%和96.27%，比之前用户标注的学科分类提升了30%以上。

对于题目难度标注的模型：由于要将简单题目的价格降低，分配给大学生老师作答，以此来降低成本，考虑到大学生老师的比例和降低价格之后的风险，系统应更关注简单题目的识别，尤其要保证简单题目的正确率。因此在最后的逻辑斯蒂回归中将阈值定为0.3，用某一天的用户流量中随机抽取了27000道用户提问的数学题，通过模型评测，共得到9396道简单题目，人工抽样100道题目进行分析评测，发现判定出来的简单题准确率极高，可以达到99%以上，召回率在50%，判别出来的简单题数目基本上和大学生老师答题数目相当。