

基于HMM与RBF混合模型的情感智能聊天系统

闫丹阳,姜梅,耿秀丽,闫伟

(山东师范大学信息科学与工程学院,山东济南250000)

摘要:当前的智能聊天系统多采用文字进行交流,存在不能准确回答问题、转移话题、答非所问等一系列问题。因此,对情感智能聊天系统进行细致的研究、改进和升级,不仅可以推动系统的发展,还可以为人们提供更为人性、智能化的服务。文中提出了一种基于HMM与RBF的混合模型,创建人类情感语音库,以还原人类最真实的情感感受,并利用Flex技术进行情感语音库的动态更新;同时,运用语料库标注体系,以标注规范、纠错机制、补充学习作为语料库质量监控手段,从而保证语料库的完备性。在该系统下用户既可以采用文字聊天,又能进行语音聊天,并在后台产生文字聊天记录,突破了现有系统只能用文字聊天的局限性。

关键词:隐马尔可夫模型;径向基函数;Flex技术;情感智能;语料库收集

中图分类号:TP302.1

文献标识码:A

文章编号:1673-629X(2018)04-0109-05

doi:10.3969/j.issn.1673-629X.2018.04.023

Emotional Intelligence Chat System Based on HMM and RBF Mixture Model

YAN Dan-yang, JIANG Mei, GENG Xiu-li, YAN Wei

(School of Computer Science and Engineering, Shandong Normal University, Jinan 250000, China)

Abstract: The existing intelligence chat system mainly communicate by text, which cannot accurately answer the question, shift the topic, and obtain all relevant answers. Thus, the further study, improvement and upgrading of emotional intelligence chat system can not only promote its development, but also provide people with a more humanized, intelligent service. In this paper, we propose a HMM and RBF mixture model to build the human emotional speech libraries which is dynamically updated based on Flex technology, catching the most real emotions of human experience. At the same time, this system, which uses the corpus annotation scheme, ensures the completeness of corpus by means of a corpus quality monitoring method of the tagging criterion, error correction mechanism, supplementary learning. Users can chat with text or voice by this system which generates text chat record, breaking the existing limitations of using text chat.

Key words: HMM(hidden Markov model); RBF(radial basis function); Flex technology; emotional intelligence; corpus collection

0 引言

随着信息技术的提高和人类在机器学习领域的研究日益加深,人类对机器系统的智能化和情感化诉求也在扩大。情感智能聊天系统作为具有智能化、情感化特点的聊天工具在信息量与日俱增的今天受到了越来越多的关注。情感智能聊天系统一方面能够作为即时通讯工具完成人机交互,准确向用户传递信息和数据;另一方面也能够在交互通讯的过程中体现出机器人不具有的智能化、情感化等突出特点,从而把通讯过程变成一个有趣味的人机交互过程。

然而当今情感智能聊天系统的发展相对缓慢,普遍存在几个显著的问题:不符合人类的聊天习惯;没有

长时记忆体^[1]的功能;语料库^[2]匮乏;不支持文本和语音双向输入输出^[3-4]。因此,情感智能聊天系统,须加以完善,使其具备更加强大的功能,拥有更加丰富的情感。

针对当前情感聊天系统的不足和匮乏,提出了一种新的情感智能聊天系统的搭建方法。拟通过利用隐马尔可夫模型(hidden Markov model, HMM)和径向基函数(radial basis function, RBF)的混合模型^[5-7]创建人类情感语音库,结合与浏览器已经建立连接的文本数据库,通过Flex技术使系统与浏览器建立连接,使语料库得到扩充和丰富,最后实现对用户的输入做出拟人化的语音或者文本双向输出的目标。

收稿日期:2017-04-23

修回日期:2017-08-29

网络出版时间:2017-12-05

基金项目:教育部留学回国基金;山东省高等学校科技计划项目(J15LN26);山东师范大学大学生创新创业训练计划项目

作者简介:闫丹阳(1995-),男,研究方向为智能信息系统;闫伟,博士,讲师,CCF会员(48076M),研究方向为智能信息系统。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20171205.1432.090.html>

1 改进方法

1.1 语料库构建

1.1.1 语料的收集

语料收集首先选出合适的语料,进行预处理工作,为下文语料的标注做好准备。文中拟通过运用合适的语料选择方法来提高语料库的覆盖率,进而提高语料库的规模、使用范围和准确性。从情感色彩看,语料库

表 1 部分类语料的详细信息

分类	详细分类	字数	词数	句数
情感类	高兴、悲伤、愤怒、平静、暴躁、愉悦、害怕、惊奇、厌恶	129 486	91 032	4 809
	动画片、当下流行元素、工作问题、结婚生子、婆媳大战、健康养生	6 308 526	4 375 396	237 290
总数		6 438 012	4 466 428	242 099

1.1.2 语料库的标注体系

语料库标注体系表示对话料的加工程度,即把待标注的语料添加到特定的信息集合中。标注体系的类别划分过粗不能准确全面地理解语言,过细导致标注信息过于庞大,会增加标注难度,降低效率,并且会降低模型的健壮性。因此,文中预先标注了语料,参考其他类型语料库标注并结合自身特点制定了特有的标注体系集合,例如:情感模型=(高兴、悲伤、愤怒、平静、暴躁、愉悦、害怕、惊奇、厌恶);生活模型=(动画片、当下流行元素、工作问题、结婚生子、婆媳大战、健康养生)。

1.1.3 语料库的质量监控

语料库监控从标注规范、纠错机制和补充学习三个方面进行。标注规范是在语料标注过程中减少错误操作,提高标注效率和一致性的有效措施^[9];纠错机制则是在语料标注完成后进行错误和一致性检查,防止错误的语料进入语料库^[9];补充学习是为了提高语料库的使用寿命及系统的智能性。现存的语料库在更新学习方面较为缺乏,不能做到及时更新,降低了语料库的寿命,加重了维护人员的工作负担。文中改进的补充学习监控方式通过对用户输入请求的判断规约出表达同类情感语料的补充收录。在用户对一句话、一种意思或情感多次重复描述基础上,把该语料记忆、收录并归类到相应的语料标注体系下,从而实现智能化的提升。

1.2 语音库构建

1.2.1 基于 HMM 和 RBF 的语音库构建

HMM^[10-11]是通过分析语音当前的波形进而推断该波形所对应的最可能的音素,得到该语音信号所对应的文字信息。在训练和识别过程中发现,不同 HMM 模型代表着不同的情感状态。通过同类情感的训练样本多次数据可以得出每个 HMM 的模型参数,

大致分高兴、悲伤、愤怒、平静、暴躁、愉悦、害怕、惊奇、厌恶九大类;从时间轴看,语料库集结了适合各个年龄段人群交流的话题,包括动画片、当下流行元素、工作问题、结婚生子、婆媳大战、健康养生等多类话题^[8]。总的来说,构建语料库更加贴近生活问题,以此达到智能聊天的目的。表 1 列出了部分类语料的详细信息。

后续可以通过修正与该情感相对应 HMM 模型来学习某一种新情感。但是 HMM 的缺陷也比较明显:HMM 训练和识别算法过于依赖强假设,从而造成模式识别性能不尽如人意;虽然充分考虑了特征类内部变化问题,却忽略了类之间的重叠性,仅仅根据各累积概率的最大值作类别判断,导致其分类决策能力欠佳,而且相应的自身适应性、鲁棒性等都产生了一定的限制。

RBF 神经网络是一种典型的局部逼近网络,网络结构分为三层^[12]:

输入层:由一些被称作感知单元的信号源点组成,用来连接网络与外界。

隐藏层:该层的作用是将输入空间传至隐空间并产生局部响应,从而能够实现分类和函数逼近。

输出层:经隐藏层处理,数据在该层进行加权求和后输出。因此,该层节点是一种线性求和单元。

RBF 隐藏层向量维数通常比较高。一般来说,隐藏层向量维数越高,RBF 网络趋近于一个光滑的输入输出映射时就越精确^[12]。RBF 神经网络的特点是能够保持非常高效的自学习。即使输入的维度较高,RBF 也能够保证较强的分类性能和较快的训练速度。

根据模式识别理论,通过非线性映射到高维特征空间可以解决低维空间线性不可分的问题,从而实现线性可分。RBF 网络的输入就是一个原始线性不可分的特征空间,可以使之经过适当的函数变换到达另一线性可分的空间,之后用线性单元解决问题。

为了提高语音情感识别的鲁棒性和识别率,把动态时间建模能力较强的 HMM 和分类学习能力较强的 RBF 这两种方式相结合,提出了基于 HMM 和 RBF 混合语音识别模型的语音库构建,即把 RBF 神经网络计算状态的观察概率结合应用于 HMM 模型,不仅解决了 HMM 模型中鲁棒性不强、训练复杂的问题,而且克服了 RBF 神经网络处理语音动态变化特征序列不尽

如人意的缺点。

语音库构建过程主要包括:

(1)原始语音预处理。

在获取用户的语音信息之后,对声音信息进行预滤波、预加重、短时加窗及端点检测等预处理;对特征参数提取训练、利用CHMM进行声学建模作为识别算法,建立基于CHMM的语音识别算法。另外从信号空间、特征空间、模型空间三方面进行语音补偿,构建一种新的较好结合了维纳滤波、直方图均衡、向量泰勒级数三种算法^[13-14]优点的语音识别算法,确保对声音信息的预处理及初步文本转化更准确,减少计算机的计算量,提高计算机反应速度。

(2)特征提取。

分析每一种语音情感的特点并提取相应特征,为下一步HMM模型的建立做准备。

(3)设计HMM模型及训练。

给每种语音情感都设计了一个HMM模型。所采用的HMM模型训练准则是基于ML的Baum-Welch算法^[15]。训练过程首先是用HMM模型对语音信号进行状态分割并使用Viterbi算法得到最优状态序列^[15],然后为了将变长的最优状态序列转变成固定维数特征向量,采用勒让德系数对正交基函数进行展开^[16]。

(4)RBF模型建立与最终识别。

RBF神经网络将对HMM的状态累计概率进行识别,通过其非线性映射能力进行映射,将RBF神经网络的决策结果认定为最终识别结果。

(5)根据识别结果归类入库。

由以上过程得出语音识别结果,根据不同的情感分类将之分别入库,建立最终的情感语音库。

语音识别过程如图1所示。

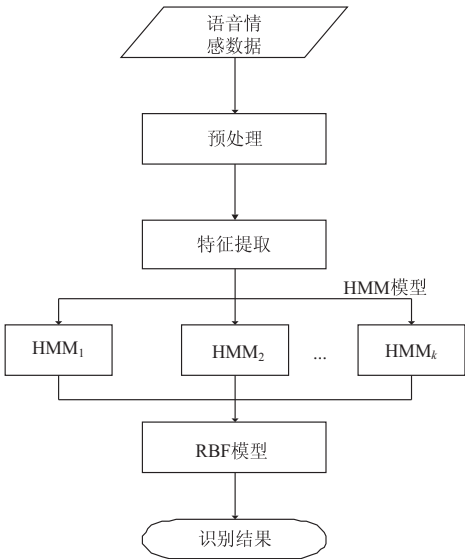


图1 基于HMM和RBF的语音识别过程

1.2.2 基于Flex技术的情感语音库动态更新

使用Flex提供的构建移动应用和传统的基于浏览器应用的基本框架,建立情感语音库与浏览器的连接,定时更新情感语音库中已有的代表某种情感状态的语句。

Flex技术提供构建移动应用和基于浏览器应用的基本架构^[15],其框架是完全开源免费的。使用Flex技术可以减少服务器之间的通信次数,详细展示出数据的细节,从而弥补了许多传统Web应用缺乏的元素,使智能聊天系统具有更良好的反应速度以及更真实的情感表达。

1.3 分词算法

现有的分词算法分为三大类:基于字符串匹配的分词算法、基于统计语言模型的分词算法和基于理解的分词算法^[17]。但由于基于统计语言模型的分词算法对常用词的敏感度低,基于理解的分词算法尚处于试验阶段等局限性,采用基于字符串匹配的分词算法,其中的双向最大匹配法,即把正向最大匹配法和逆向最大匹配法相结合,能够确保最精准的匹配度。

由于汉语词的长度差异大,有的多字词,长度为十几个汉字,而单字成词长度为1。最大匹配算法的初始切分长度常为词典最长词条的汉字数 M ,如此切分和匹配影响了算法效率。另外,二字词和三字词在汉语词中占有相当大的比例,而以词首字开始的二字词、三字词和多字词的数量能够反映出词首字开始的词为二字词、三字词和多字词的可能性。因此,在最大匹配算法中引进随机数得到最大匹配的概率算法,并以词首字最长词长 L_{\max} 为最大切分限界值^[18-20]。设待切分的语料汉字串为 $\text{Str} = S_1S_2\cdots S_n$,基于最大匹配的概率算法描述如下:

(1)取 S_1 ,通过hash映射,找到词首字索引项,获取相关数据。

(2)若 $\text{maxlen} = 1$,则 S_1 为词首字的词表为空,将 S_1 切分出来。然后令 $\text{Str} = S_2S_3\cdots S_n$,继续下一次切分;若 $\text{maxlen} > 1$,则计算:

$$S_{\text{No}} = N_{\text{tw}} + N_{\text{th}} + N_{\text{mlt}}$$

其中, N_{tw} 表示二字词数量; N_{th} 表示三字词数量; N_{mlt} 表示多字词数量。

(3)产生 $1 \sim S_{\text{No}}$ 范围内的随机数: $X = \text{Random}(S_{\text{No}})$ 。

Case $X \leq N_{\text{tw}}$,取 $K = 2$;

Case $X \leq N_{\text{tw}} + N_{\text{th}}$,取 $K = 3$;

Case $X \leq N_{\text{tw}} + N_{\text{th}} + N_{\text{mlt}}$,则取 $K = \text{maxlen}$ 。

(4)取 $\text{Str}_1 = S_1S_2\cdots S_k$,在字典中查找 Str_1 。

①若 Str_1 不是词,重新产生随机数,获取余下的 K 值,继续在字典中查找,直到查找成功。若所有 K 值查

找都不成功,则 S_1 在此处可视为 1 个单字词,得到切分 $S_1/S_2S_3\cdots S_n$ 。同时可通过人工干预方式,将词首字为 S_1 的一个子串作为新词,将其插入到多字词链表。

②若 Str_1 是词,则增加一个字 $Str_1 = Str_1 + S_{k+1}$,再查找,若 Str_1 是词,继续增加一个字,直到 L_{\max} ,并记录词的最后一个字的位置 p 。则可暂时获得切分词: $Stmp_1 = S_1S_2\cdots S_p$ 。

③取 S_2 为首字词,重复以上操作,则可获得另一切分词 $Stmp_2$,若 $Length(Stmp_1) > Length(Stmp_2)$,则得到切分词: $Stmp_1$,否则,得到切分词: $S_1/Stmp_2$ 。

(5)移动汉字串指针,进行下一次切分,直到整个串切分完成。

例如:“当中国人民站起来的那一天”。
词首字为“当”,若 $Stmp_1 = \text{“当中”}$,而词首字为“中”, $Stmp_2 = \text{“中国人民”}$ 。
可切分为:当/中国人民。
词首字:“站”,则 $Stmp_1 = \text{“站起来”}$,词首字为“来”, $Stmp_2 = \text{“来”}$ 。

可切分为:当/中国人民/站起来。
最后可切分为:当/中国人民/站起来/的/那一天。
尽管正向最大匹配法和逆向最大匹配法都是比较常用的分词算法,但并不代表它们能准确无误地完成用户所需要的切分任务。统计结果表明^[21],正向最大匹配算法的错误率为 1/169,逆向最大匹配算法的错误率为 1/245。事实上,只能最大限度地追求低失误差,文中采用将两者结合的手段,能在一定程度上提高分词的正确性,以期达到更加智能的切分效果。

1.4 情感语言的输出

利用语音合成技术将查询到的文本结果转化为语音输出,并利用 TTS 技术朗读预先未知的任何语句,将文字信息的实时动态转化为语音形式输出到用户端,从而实现聊天系统与用户之间更富情感的对话。
在文本信息转化为语音信息输出时,系统会把语音预处理之后的文本和经过情感语音库匹配后输出的文本以聊天记录形式保存下来,实行保密机制,用于验证登陆查看聊天记录。具体实现过程将在下文阐述。

2 实验及结果分析

2.1 情感语料的收集

2.1.1 录制语料

采用 Cool Edit Pro 高质量地完成录音、编辑、合成等多项任务。在录音时采用采样频率为 11.025 kHz、采样精度为 32 位、单声道的录制方式,录制语言保存为 PCM 编码的 WAV 格式。选择 30 名 18~22 岁在校大学生,要求口齿清楚、听力正常、能较好表达自身情感。录制者按情感提示朗读相应的情景文本,录制有

关的语音数据以供后续研究。
2.1.2 分析评估语料库
从情感识别率 (EIR) 以及情感强度 (ES) 两方面对所获得语料库进行分析评估^[22-23]。具体规则如下:

- (1)情感识别率:从有限的情感种类集(如高兴、愤怒、惊讶、悲伤、恐惧等)识别给定情感句子的目标情感,测试其识别率大小。
- (2)情感强度:评估给定情感句子的情感强度,可以设计 5 个打分标准:非常弱、弱、一般、强、非常强。

通过上述标准评测出的语料库有效性发现,上文录制的情感语料能够满足研究需求,因此可以继续下一阶段实验。

2.2 HMM 和 RBF 相结合的情感语音识别

2.2.1 实验样本的选择

表 2 为上一节选择出的典型的实验录音脚本(部分)。

表 2 实验录音脚本(部分)

序号	语句	序号	语句
1	我们去看电影吧	6	这天气真热啊
2	考试成绩出来了	7	有辆汽车向我们开过来
3	快要下雨了	8	明天是周末
4	怎么会这样	9	我去接你
5	咱们说走就走	10	他快来了

2.2.2 实验过程及结果

实验信号采用汉明窗分帧,其标准为窗长 256、帧移 128,6 个 HMM 状态,每个状态给以 5 个高斯概率密度函数,24 维 RBF 输入。提取特征向量并且结合 HMM 和 RBF 混合模型识别,采用单一的 HMM 仿真进行识别率对比。结果表明,单一 HMM 识别效果较差,平均识别率仅为 60.1%,而 HMM 和 RBF 混合模型平均识别率为 66.1%,整体效果较单一 HMM 更好。从实验结果可以看出,混合模型对提高识别率有较好的效果,因此选择该技术与情感语音聊天相结合来改善目前语音聊天系统的部分问题。

由于信息采集是以完全模拟现实环境为准则,所以其抗噪音性能的验证也是不言而喻的,此处给出相关实验数据(见表 3),并对实验结果进行直观化处理,如图 2 所示,以方便对实验结果的进一步分析。

表 3 识别结果 %

情感	单一 HMM 的识别率	混合模型的识别率
惊奇	57.6	68.8
愤怒	60.0	64.0
高兴	63.2	70.3
悲伤	59.6	61.3

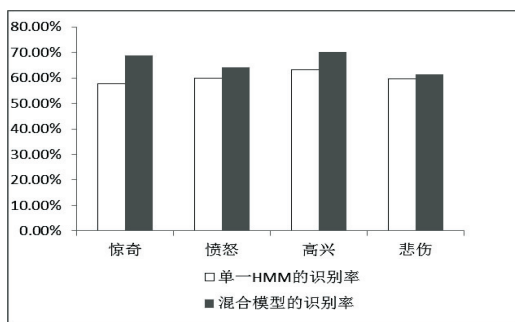


图2 混合模型和单一 HMM 模型的实验数据分析图

2.2.3 相关实验数据结果分析

人的语速变化与其所处的情感状态有关。实验结果表明,语音信号的振幅特征与各种情感信息具有较强的相关性:喜、怒、惊等情感,信号的振幅往往较大,悲伤情感的幅度值则较低。而且可以看出这些幅度值的差异越大,体现出的情感变化也越大。基于此,可以利用语音信号中的语速以及语音持续时间等参数来判断语言的情感。对于利用这一特点来判断人的语音情感,从而使该系统感知出人的说话情感变化具有重要的现实意义。

3 系统实现过程概述

系统实现过程共有三个步骤:首先将语音输入转化为文字,这一步可以采用 HMM 和 RBF 技术对语音输入进行识别,并转化为可供后台处理的文本格式。这是至关重要的一步,因为语音到文本的转化容易产生差错,所以采用基于 HMM 和 RBF 的混合模型来更好地解决这一问题;其次是将转化的文字进行分词处理并与语料库中的词组进行匹配,可采用多种分词算法来达到更加智能化的匹配,如贪心算法、双向最大匹配法等^[16],以便为下一步的输出回复做好准备;最后需要将匹配出来的回答以文字方式直接输出或者转化为语音进行输出。文中使用基于 HMM 和 RBF 的混合模型进行语音文本转化,具体过程如图3所示。

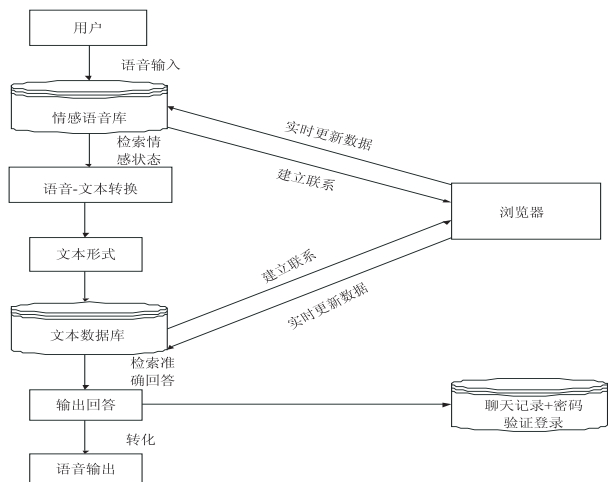


图3 系统实现示意图

4 结束语

提出了一种基于 HMM 和 RBF 的情感智能聊天系统搭建方法。该方法利用 HMM 和 RBF 的混合模型创建人类情感语音库,利用 HMM 生成最佳语音状态序列,用函数逼近技术产生对最佳状态序列进行时间归正,RBF 神经网络进行分类识别。再通过 Flex 技术建立数据库与浏览器之间的连接,保证系统拥有足够丰富的数据库和语料库。结合双向最大匹配算法,完成对中文分词和分析归类。将各个模块组建在一起实现更加智能化情感化的聊天系统。尽管如此,对于所构想的真正智能化还有一定的差异,特别是在聊天系统的自主学习方面仍然需要很大的改进。

参考文献:

[1] 罗毅. 一种基于 HMM 和 ANN 的语音情感识别分类器[J]. 微计算机信息, 2007, 23(12-1): 218-219.

[2] 胡瑞敏, 薛东辉, 姚天任, 等. 神经网络方法及其在语音识别中的应用[J]. 高技术通讯, 1995(6): 11-15.

[3] COWIE R. Emotion recognition in human-computer interaction[J]. Signal Processing Magazine, 2001, 18(1): 32-80.

[4] 韩文静, 李海峰, 阮华斌, 等. 语音情感识别研究进展综述[J]. 软件学报, 2014, 25(1): 37-50.

[5] 马晓梅, 李雪耀, 王洋. 基于 HMM 的连续语音中的关键词检测[J]. 黑龙江科技信息, 2008(32): 91.

[6] 赵力, 钱向民, 邹采荣, 等. 语音信号中的情感识别研究[J]. 软件学报, 2001, 12(7): 1050-1055.

[7] 闻彬, 何婷婷, 罗乐, 等. 基于语义理解的文本情感分类方法研究[J]. 计算机科学, 2010, 37(6): 261-264.

[8] LEE C M, NARAYANAN S. Toward detecting emotions in spoken dialogs[J]. IEEE Transactions on Speech and Audio Processing, 2005, 13(2): 293-303.

[9] 孙晋文, 肖建国. 基于 SVM 文本分类中的关键词学习研究[J]. 计算机科学, 2006, 33(11): 182-184.

[10] NEFIAN A V, HAYES M H. Face detection and recognition using hidden Markov models[C]//Proceedings of the international conference on image processing. [S. l.]: IEEE, 2002: 141-145.

[11] CHEN S H, CHEN W Y. Generalized minimal distortion segmentation for ANN-based speech recognition[J]. IEEE Transactions on Speech and Audio Processing, 1995, 3(2): 141-145.

[12] ORR M J L. Introduction to radial basis function networks[J]. Internationale Zeitschrift für Vitaminforschung, 2003, 37(3): 97-101.

[13] 郝杰, 李星. 汉语连续语音识别中经典 HMM 的实验评测[J]. 计算机工程与应用, 2001, 37(13): 1-4.

[14] 蒋丹宁, 蔡莲红. 基于语音声学特征的情感信息识别[J]. 清华大学学报: 自然科学版, 2006, 46(1): 86-89.

低体现不出 SQL 代码的结构,如果比重过高,难以体现细节之处。经过分析衡量,临界点大约在 0.62 ~ 0.69 之间,为便于计算,将关键词的影响因子定为总代码的 2/3。

经过实验与数据统计,该模型可以对较长的 SQL 代码进行相似度匹配。采用同义库以及最长公共子串的算法,有效提高了自动评分的正确性。但是经过分析发现,自动评分也存在语义正确但是评分低的情况,如关键词错误。因此,对于关键词错误的情况,应该另行分析。

3 结束语

针对数据库 SQL 代码的评分,提出了新的 SQL 代码的自动评分模型。该模型借鉴了代码的静态分析方法,对相似语义构建了 SQL 代码的同义库,运用最长公共子串的算法对代码片段进行相似度匹配,又根据人工评判数据进行多项式拟合,制定了贴合实际的评分策略,并为此开发了原型系统。实验结果表明,该模型有效提高了 SQL 代码的评分效率及正确率,因此可以定量地衡量 SQL 代码的水平。但是,该模型存在一些问题,比如关键词出错的处理,尚需进一步深入,可作为未来工作的研究点。

参考文献:

- [1] 解 萍.《数据库》课程在学生软件能力培养中的作用分析[J]. 科技视界,2012(15):101-102.
- [2] 杨鹤标,刘 玲,杨立凡. 基于结构相似匹配的 SQL 程序自动评估模型研究[J]. 计算机工程与科学,2010,32(11):92-96.
- [3] 郑燕娥. Java 编程题自动评分技术的研究与实现[D]. 泉州:华侨大学,2013.
- [4] 段汉周,凌 捷,郑衍衡. VB 程序设计考核自动评阅系统中若干问题的研究[J]. 计算机工程,2001,27(4):167-168.
- [5] SAMARIA F, YOUNG S. HMM based architecture for face identification[J]. Image and Vision Computing, 1994, 12(8):537-543.
- [6] FREITAG D, MCCALLUM A. Information extraction with HMM structures learned by stochastic optimization[C]//Proceedings of the eighteenth conference on artificial intelligence. [s. l.]:[s. n.], 2002:584-589.
- [7] 杨晓恕,蒋 维,郝文宁. 基于本体和句法分析的领域分词的实现[J]. 计算机工程,2008,34(23):26-28.
- [8] 欧振猛,余顺争. 中文分词算法在搜索引擎应用中的研究[J]. 计算机工程与应用,2000,36(8):80-82.
- [9] 马玉芳,李方数据. Web 中文文本分词技术研究[J]. 计算机

- [5] ŠTAJDUHAR I, MAUŠA G. Using string similarity metrics for automated grading of SQL statements[C]//International convention on information & communication technology, electronics & microelectronics. Washington DC, USA: IEEE Computer Society, 2015:1250-1255.
- [6] 王甜甜. 基于语义相似度的编程题自动评分方法的研究[D]. 哈尔滨:哈尔滨工业大学,2005.
- [7] 马培军,王甜甜,苏小红. 基于程序理解的编程题自动评分方法[J]. 计算机研究与发展,2009,46(7):1136-1142.
- [8] 牛永洁,张晓光. 关于程序设计题自动评分方法的研究[J]. 信息技术,2010(11):155-156.
- [9] CHEN Yaofei, CHEN Huantong. Research of automatic marking on SQL server skill assessment based on XML[C]//International conference on web information systems and mining. Washington DC, USA: IEEE Computer Society, 2010:8-12.
- [10] HINKKA M, LEHTOAND T, HELJANKO K. Assessing big data SQL frameworks for analyzing event logs[C]//24th Euromicro international conference on parallel, distributed, and network-based processing. Washington DC, USA: IEEE Computer Society, 2016:101-108.
- [11] 王小凤,周明全,耿国华,等. 一种基于字符距离的特征字符串近似匹配算法[C]//图像图形技术与应用学术会议. 北京:北京师范大学出版社,2008.
- [12] 徐黎明. 基于 GST 字符串近似匹配算法的研究[J]. 内蒙古科技与经济,2016(7):87-89.
- [13] KLEINER C, TEBBE C, HEINE F. Automated grading and tutoring of SQL statements to improve student learning[C]//Proceedings of the 13th Koli calling international conference on computing education research. New York, NY, USA: ACM, 2013:161-168.
- [14] 冯君远,赖明钦,李启良. C 语言源代码抄袭识别的研究[J]. 福建电脑,2013,29(5):34-36.
- [15] POHUBA D, DULIK T, JANKU P. Automatic evaluation of correctness and originality of source codes[C]//10th European workshop on microelectronics education. Washington DC, USA: IEEE Computer Society, 2014:49-52.
- [16] 应用,2004,24(4):134-135.
- [17] ZHANG M Y, LU Z D, ZOU C Y. A Chinese word segmentation based on language situation in processing ambiguous words[J]. Information Sciences, 2004, 162(3-4):275-285.
- [18] 何国斌,赵晶璐. 基于最大匹配的中文分词概率算法研究[J]. 计算机工程,2010,36(5):173-175.
- [19] 王洪伟,郑丽娟,尹 裴,等. 基于句子级情感的中文网络评论的情感极性分类[J]. 管理科学学报,2013,16(9):64-74.
- [20] YE Q, ZHANG Z, LAW R. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches[J]. Expert Systems with Applications, 2009, 36(3):6527-6535.

(上接第 113 页)