基于深度神经网络的语音识别系统研究

李伟林 文 剑 马文凯
(北京林业大学工学院 北京 100083)

摘 要 语音识别是人机交互模式识别领域的一个重要课题,构建了一种基于深度神经网络的语音识别系统,使用了抗噪对比散度法和抗噪最小平方误差法对模型进行无监督训练;使用了均值归一化进行模型优化,提高了网络对训练集的拟合度,并且降低了语音识别的错误率;使用多状态激活函数进行了模型优化,这不仅使得不带噪测试和带噪声测试的语音识别错误率进一步下降,并能在一定程度上减轻过拟合现象;并通过奇异值分解和重构的方法对模型进行了降维。实验结果表明,此系统可以在不影响语音识别错误率的基础上极大地降低系统的复杂性。

关键词 模式识别,深度神经网络,语音识别,隐马尔科夫模型,模型重构
中图法分类号 TP391 **文献标识码** A

Speech Recognition System Based on Deep Neural Network

LI Wei-lin WEN Jian MA Wen-kai
(School of Technology, Beijing Forestry University, Beijing 100083, China)

Abstract Speech recognition is an important subject in the field of human computer interaction pattern recognition. A speech recognition system based on deep neural network was constructed in this paper. The model was trained without supervision by using the method of anti-chirp contrast divergence and anti-chirp least squares error. The model optimization was carried out using the average value normalization. The fitting degree of the network to the training set is improved and the error rate of speech recognition was reduced. The system used the multi-condition activation function for the model optimization, then the error rate of speech recognition without noise and noise measurement was further reduced. So the system can reduce the over fitting phenomenon. The model was reduced by using the method of singular value decomposition and reconstruction. Experimental results show that the system can greatly reduce the complexity of the system without affecting the error rate of speech recognition.

Keywords Pattern recognition, Deep neural network, Speech recognition, Hidden markov model(HMM), Model reconstruction

1 引言

语音识别是将人类所发出的语音转化为文字或符号的技术。从 40 年前开始对声学特征的抽取,到如今使用深度神经网络作为主体的自动语音识别系统,语音识别技术已经逐步完善。但语音识别技术也面临着一些问题,比如在语音识别中单纯地提取出声音频谱作为特征并不能达到很高的识别率,模型具有较高的时间或空间复杂度会限制语音识别技术的应用以及导致环境噪声问题等。

基于深度神经网络的语音识别系统具有很强的非线性处理能力,相比于高斯混合模型(Gaussian Mixture Model, GMM),其可以显著提高系统性能并减少时间和空间复杂度。在性能上可以通过无监督训练的方法提升抗噪性能[1],Hessian Free 优化可以减少训练过程消耗的时间[2];异步随机梯度下降法[3]、随机数据丢弃[4]、基于平均随机梯度下降法的单次迭代算法[5]、奇异值分解[6]、节点修剪(Node-pruning)[7]等方法都可以在一定程度上减少时间和空间的复杂度。深度神经网络在高斯混合模型作为基础上与隐马尔科夫模型(Hidden Markov Model, HMM)相结合,在连续语音识别上得到了很好的实验结果。此后,研究者发现即使没有高斯混合模型作为基础,深度神经网络同样可以取得良好的性能[8]。另外深度神经网络还被应用于自然语言理解[9]。

语音识别系统在不同场合中需要不断提高本身性能,而且需要在保证性能的前提下做到经济合理。深度神经网络复杂度较高,模型复杂,一般需要较高的硬件配置。本文将深度神经网络与特征提取技术和隐马尔科夫模型相结合,通过奇异值分解和重构的方法对模型进行了降维,所构建的自动语音识别系统同样具有较高的性能,识别错误率也较低。

2 语音识别系统的构建

语音识别系统的学习分为高斯混合模型和深度神经网络两个部分。高斯混合模型的学习是较为成熟的技术,并不属

本文受国家级大学生创新创业训练计划资助项目(201510022062)资助。
李伟林(1993-),男,硕士生,主要研究方向为机电系统控制及自动化;文 剑(1981-),男,博士,讲师,主要研究方向为智能控制、机电一体化技术;马文凯(1993-),女,硕士生,主要研究方向为森林工程装备及其自动化。

于深度学习的范畴，但它是构建一个深度神经网络的必要步骤。高斯混合模型是一种线性模型，能够对信号进行线性变换，使得信号被转化为有助于模式识别的形式。经过高斯混合模型的学习过程后，系统会得到每个特征所对应的对齐结果，这些对齐结果将作为标签用于深度神经网络的学习。

2.1 预训练过程

预训练是深度神经网络中一个重要的学习过程。然而，一般的预训练方法并不能很好地解决信号中的噪声问题。为了使得网络能够具有抗噪性能，在预训练中要对受限玻尔兹曼机进行特殊的训练，首先需要按照一定的规则训练它的权值，通过观察各个输入节点和隐含节点的状态进行权值的训练，计算后得到一个公式：

$$\Delta w_{ij}=\varepsilon(\langle v_i h_j\rangle_{clean}-\langle v_i h_j\rangle_{noise}) \tag{1}$$

其中，ε 为学习率，$\langle v_i h_j\rangle_{clean}$ 为纯净信号采样，$\langle v_i h_j\rangle_{noise}$ 为带噪信号采样。

$\langle v_i h_j\rangle_{clean}-\langle v_i h_j\rangle_{noise}$ 和一般的对比散度法中的 $\langle v_i h_j\rangle_{data}-\langle v_i h_j\rangle_{recon}$ 具有非常相似的表达形式。它们之间的不同点在于，前者需要用到纯净和抗噪的信号，并且是基于贪心原理；而后者只需要用到单一的信号，并且是基于能量最小化原理。故前者可命名为抗噪对比散度法。抗噪对比散度法对于受限玻尔兹曼机来说并不是最优的，因为在这种情况下，对于隐含节点来说，在分别经过纯净信号和带噪信号激活后，获得的信息更为接近。这种方法虽然使得受限玻尔兹曼机具有抗噪性能，但是也使得它处理纯净信号的能力变弱。

在使用抗噪对比散度法的同时，本文也使用最小平方误差法对受限玻尔兹曼机进行抗噪约束。这种最小平方误差法的目的是使得在纯净信号和带噪信号的激活下隐含层节点的状态能够取得尽可能小的平方误差值。

使用最小平方误差法，当隐含层节点的状态全部为 0 时，将无法对信号进行重构；而一般的对比散度法的训练目标是使得隐含层节点能够恢复出有用的信号，在训练过程中将会直接避免这种情况的发生。

由于一般的对比散度法、抗噪对比散度法和最小平方误差法都具有各自的优点，因此将这 3 种方法同时用于训练受限玻尔兹曼机。训练过程如下：

(1)使用一般对比散度法，使得模型具有一个较优的初值；

(2)在较优的初值的基础上，使用抗噪对比散度法进行训练；

(3)在此基础上，使用最小平方误差法进行训练。

训练了单个受限玻尔兹曼机后，在上方再堆叠一个新的受限玻尔兹曼机，然后对其进行相同的训练。新的受限玻尔兹曼机的输入应该来自于前一个受限玻尔兹曼机的输出。通过堆叠并训练多个受限玻尔兹曼机，最终可以得到一个深度信任网络。深度信任网络所实现的功能是对输入特征进行逐层转换。由于每一个玻尔兹曼机都具有抗噪功能，这种深度信任网络同样具有抗噪功能。它由 3 个普通的层与 2 个抗噪层堆叠而成。由于纯净信号和噪声信号之间的差异较大，在具有抗噪功能的深度信任网络中，层次较深时参数不容易收敛。因此，实际中的抗噪层一般不超过 2 层。

2.2 反向传播算法

2.2.1 有监督训练

无监督训练深度信任网络之后，可以对网络进行有监督训练。在使用随机梯度下降法进行训练的过程中，初期为了加快速度，一般采用较大的学习率。而在训练中期之后，因为模型的性能逐渐变好，对参数的更新也需要越来越精确，所以需要较小的学习率。因此，需要一种调整学习率的规则，使得学习率的大小能根据不同阶段的需求进行调整。一般来说，最常用的方法是交叉熵准则(Cross Entropy Criterion，CEC)[10]。它使用交叉熵损失函数评价一个模型的精确度。在输出层，假设网络有 m 个节点，每一个节点的输出是 o_j，对应的期望输出是 t_j，那么这个网络的交叉熵损失函数为：

$$loss=-\frac{1}{m}\sum_j t_j\cdot\log(o_j) \tag{2}$$

一般来说，在训练时需要进行交叉验证(Cross-validation，CV)。如果交叉熵损失的降低程度小于一定数值，说明此时的学习率已经太大，应该减少学习率再继续训练。如果交叉熵损失在学习率很低的情况下也没有明显降低，说明模型已经训练得足够充分，可以终止训练。

2.2.2 多状态激活函数

Logistic 函数可以很好地处理二分类问题，而在 ReLU 上实现多分类将取得比二分类更好的识别效果。但是 ReLU 存在线性部分，参数在训练过程中会面临较为严重的过拟合问题。为了实现多分类，且减轻过拟合问题带来的危害，本文使用一种由多个 Logistic 函数叠加而成的多状态激活函数，其表达式为：

$$M_N(x)=\sum_{i=1}^{N}\frac{1}{1+e^{-x+x_i}} \tag{3}$$

其中，N 为正整数，x_i 为每个 Logistic 函数的位移。

2.2.3 均值归一化

对输入层节点的均值进行归一化较为容易，可以直接根据输入的特征值进行归一化处理；而对隐含层节点的归一化处理比较困难，训练过程中神经网络参数不断变化，会造成相应输入向量的变化，这就使得无法使用与处理输入层相同的方法对隐含层进行归一化。本文使用一种基于运行均值(Running Average)的二阶收敛算法对隐含层进行归一化。

均值归一化的目的是使得 $a_{(n)}$ 逐成为零向量，所以在训练结束时，所有节点的运行均值都被尽可能地约束至零，这种约束可以使模型性能得到提升。

2.3 模型重构

为了减少模型的内存消耗并且提高解码模型的效率，可以采用一系列方法对训练后的模型进行重构。其中，一种较为有效的方法是基于奇异值分解的模型重构[6]。

深度神经网络的参数主要集中在权值矩阵。一个模型占用的内存以及识别所消耗的时间主要取决于权值矩阵的维度，为了提高效率应当尽可能地减少权值矩阵的维度。然而如果在训练时减少矩阵的维度，会使得模型的性能变差。为了在保证模型性能不变差的情况下减少参数数量，需要在完成训练之后再对模型进行重构。奇异值分解可以将一个权值矩阵分解成为 3 个不同矩阵，包括一个左特征矩阵、一个奇异值矩阵和一个右特征矩阵。在奇异值矩阵中，奇异值从大到小排列，有一些奇异值甚至可以被忽略。通过舍弃一部分奇

异值，并且删除特征矩阵中相应的行或列，可以使得模型的参数数量明显减少，如图 1 所示。

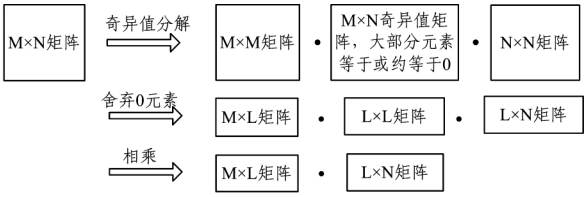

图 1　奇异值分解与重构

然而当矩阵被重构之后，如果参数数量太少，可能会导致有用信息的丢失，使得模型的性能下降。这时可以使用一个重训练的过程对模型进行微调，使得模型的性能重新得到提升。重训练可以通过随机梯度下降法完成。但由于一个权值矩阵被重构为两个，原有的训练规则不再适用，需要使用一种新的训练规则。对于每一层这种规则具体为：

先前向计算输出矩阵，令

$$X_{(m\times r)}=W_{1(m\times l)}\cdot(W_{2(l\times n)}\cdot I_{(n\times r)}) \tag{4}$$

那么

$$O_{(m\times r)}=\varphi(X_{(m\times r)}) \tag{5}$$

在反向计算中，得到误差信号 $E_{(m\times r)}$ 之后，需要用一种特殊的方法来更新权值矩阵和传递参数。权值矩阵的更新规则为：

$$\Delta W_{1(m\times l)}=\varepsilon\cdot\varphi'(X_{(m\times r)})*E_{(m\times r)}\cdot(W_{2(l\times n)}\cdot I_{(n\times r)})^{\mathrm{T}} \tag{6}$$

$$\Delta W_{2(l\times n)}=\varepsilon\cdot W_{1(m\times l)}^{\mathrm{T}}\cdot(\varphi'(X_{(m\times r)})*E_{(m\times r)})\cdot I_{(n\times r)}^{\mathrm{T}} \tag{7}$$

偏置值向量的更新规则为：

$$\Delta b_{(m)}=\varepsilon\cdot addcols(\varphi'(X_{(m\times r)})) \tag{8}$$

传递给下一层的误差为：

$$E_{(n\times r)}^{\downarrow}=W_{2(l\times n)}^{\mathrm{T}}\cdot[W_{1(m\times l)}^{\mathrm{T}}\cdot(\varphi'(X_{(m\times r)})*E_{(m\times r)})] \tag{9}$$

使用这种方法可以同时更新两个重构的矩阵，并且也能够将误差传递至下一层。

实际上，这种方法最重要的意义并不在于训练过程，而在于识别过程。当神经网络被用于识别时，需要根据输入的特征进行一次前向计算，从而将特征转化为对应的标签。此时，神经网络模型的时间消耗和内存消耗就成为了需要关注的方面，特别是在硬件有限的设备（如一些嵌入式设备）上。如果消耗的运行时间过长或消耗的内存过多，都会影响实际应用时的效果。而这种基于奇异值分解的模型重构方法不仅能够减少参数数量，而且可以减少在前向计算过程中实数乘法的次数。也就是说，这种方法不仅减少了内存消耗，而且减少了运行的时间。这就说明，可以在硬件有限的情况下提高识别的效率，或是在保证识别效率的情况下减少硬件所需要的成本。此外这种方法还可以提高模型的性能，在一定程度上解决模型的过拟合问题。

2.4　系统结构与识别过程

语音识别器分为声学模型部分和语言模型部分两个主要部分。声学模型部分主要由梅尔频率倒谱生成部分和深度神经网络部分组成，其中深度神经网络是通过之前所述方法训练而成，它很大程度上决定了系统的性能。语言模型部分由连接概率模型构成，它是一个有向有环图，在给定马尔科夫链之后通过在该有向有环图上寻找一条概率最大的路径，可以得到一组输出，将其作为最终的识别结果。图 2 表明了该系统中进行语音识别的过程。声音信号先通过深度神经网络生成音素标签，音素标签再经由语言模型生成文字。

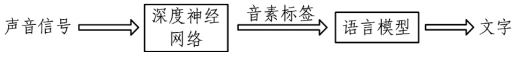

图 2　语音识别过程

具体来说，识别过程包括如下几个步骤：

（1）以 0.1 秒为单位时间，对声音进行分帧；

（2）计算每一帧声音的梅尔频率倒谱系数，得到一个输入特征矩阵 $I_{(n\times r)}$，其中 r 为帧数；

（3）使用深度神经网络进行逐层的前向计算，得到输出矩阵 $O_{(m\times r)}$；

（4）在输出矩阵中，找到每一个列向量中的最大输出概率值，构成一个具有 r 个元素的马尔科夫链 $H_r=(h_1,h_2,h_3,\cdots,h_r)$，其中每一个元素分别代表每一帧声音的音素；

（5）将 H_r 作为语言模型的输入，在有向有环图中求得最大概率所在的路径，得到相应的文字信息，输出识别结果。

3　实验结果与分析

采用本文所述的语音识别系统对小规模连续语音做了识别实验。在小规模连续语音识别实验中，对各种深度神经网络的训练和优化方法进行了测试，包括抗噪模型、均值归一化、多状态激活函数和奇异值分解与重构法，得到了一种具有抗噪性能且规模较小的模型。

通过使用语音数据库对模型进行训练和测试，可以评价一个语音识别系统的性能。一般来说，需要先使用训练集对模型进行训练；然后再通过测试集进行错误率的测试。语音数据库包括如下几个部分。

（1）训练集：包括用于训练的录音以及录音所对应的文本。训练所使用的录音来自于在自然条件下采集的声音，也就是在可能具有各种背景噪声和扰动的环境下录制的声音，而且一段录音应当是一句或一段完整的语音。声源也应当覆盖不同地区，包含各种不同的音色、口音和说话方式。录音所对应的文本是每一段录音所对应的最原始的文字信息。

（2）测试集：包括用于测试的录音以及录音所对应的文本。

（3）词典：用于描述每个文字的发音，可直接使用汉语拼音进行标注。

（4）语言模型：用于描述每个词的出现概率以及与下一个词的连接概率，是一种概率模型，主要为 N-gram 模型。

测试语音识别系统时，一般采用识别的词的错误率来评价系统的性能。错误率的计算公式为：

$$\text{错误率}=\frac{\text{错误词数}}{\text{总词数}}\times 100\% \tag{10}$$

测试深度神经网络的性能时，一般采用交叉熵损失进行评价，即假设输出层网络有 m 个节点，每一个节点的输出是 o_j，对应的期望输出是 t_j，那么交叉熵损失是：

$$loss=-\frac{1}{m}\sum_j t_j\cdot\log(o_j) \tag{11}$$

小规模测试所使用的训练集为大约 20 小时的录音及其对应文本。集内测试时所使用的数据为训练集中随机取出的 2 小时录音；集外测试中不带噪声测试所使用的数据为 2 小时录音及其对应文本，带噪声测试所使用的数据为在这 2 小

时录音中加入随机噪声所构成的数据。所使用的字典带有7232个常用汉字，语言模型为一个5-gram模型。

实验过程分为如下几个步骤：1)提取梅尔频率倒谱系数作为特征；2)训练单音素模型；3)训练三音素模型；4)用线性判别分析进行优化；5)预训练一个深度信任网络；6)在深度信任网络的基础上，使用快速训练算法和均值归一化的随机梯度下降法训练一个深度神经网络；7)进行奇异值分解和重构，降低模型规模，再使用均值归一化的随机梯度下降法重新训练网络；8)测试错误率。

在实验步骤1)至步骤4)中，对高斯混合模型所用的训练和优化技术都已经基本完善。在高斯混合模型的基础上，继续进行步骤5)至步骤8)的实验。在步骤5)至步骤7)中为了加快深度神经网络的训练速度，将使用一个NVIDIA GeForce Titan Black显卡进行计算，该显卡具有2880个流处理器以及6GB显存。

在训练深度信任网络时，采用的特征为每0.1秒的梅尔频率倒谱系数，对该特征求delta特征，即与前0.1秒特征和后0.1秒特征求差值，之后将连续11个0.1秒的delta特征合并，这样就得到了一个429维的特征向量，这种特征向量将作为网络的输入。为了与特征向量对应，深度信任网络的输入层节点数为429，设置5个隐含层，每个隐含层的节点数取1024，并且取Logistic函数作为激活函数。预训练第一层网络使用0.2作为学习率，循环迭代次数为2；预训练其它层网络使用0.4作为学习率，循环迭代次数为1。

在得到的深度信任网络上加上一层具有2096个节点的Softmax输出层，构成一个完整的深度神经网络。在隐含层中，对于一般的模型，取Logistic函数作为激活函数；对于使用多状态激活函数的模型，取Logistic函数作为第一和第二层的激活函数，取二阶的多状态激活函数作为第三层的激活函数，取三阶的多状态激活函数作为第四层的激活函数，取四阶的多状态激活函数作为第五层的激活函数。训练使用均值归一化的随机梯度下降法，初始学习率为0.008，同时使用纯净和带噪声的测试集进行交叉验证。当交叉熵损失的减少量小于0.01时，将开始对学习率进行减半；当交叉熵损失的减少量小于0.001时，将自动停止训练。为了进行对照，还将通过最一般的随机梯度下降法训练一个相同结构的网络。

训练结束之后，需要对模型进行奇异值分解、重构和再训练，以降低模型的规模。首先对除了第一层网络以外的所有层进行分解和重构，分别保留128，256和384个奇异值；然后通过均值归一化的随机梯度下降法对重构的网络进行再训练，再训练的初始学习率设置为0.000001，并且同样使用纯净和不带噪声的测试集进行交叉验证。当交叉熵损失的减少量小于0.01时，将开始对学习率进行减半；当交叉熵损失的减少量小于0.001时，将自动停止训练。

训练深度神经网络时，训练集上的交叉熵损失变化曲线如图3所示。可以看出，抗噪的均值归一化模型在整体上交叉熵损失最低，而在加入了多状态激活函数之后，交叉熵损失有较明显的提升，一般模型同样拥有较低的交叉熵损失。从下降趋势来说，多状态激活函数在第一次迭代后，交叉熵损失明显高于其它三者，而在之后下降较快，但最后依然无法到达更低的数值。

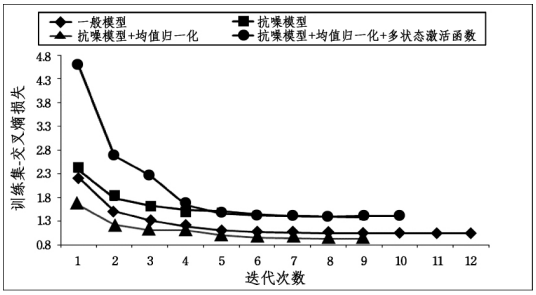

图3　训练集上的交叉熵损失变化曲线

训练深度神经网络时，测试集上的交叉熵损失变化曲线如图4所示。在测试集上，抗噪模型取得了相对更低的交叉熵损失值。最后3个抗噪模型的交叉熵损失值较为接近，其中加入均值归一化和多状态激活函数的抗噪模型得到了最低的损失值。与图3相比较可以得知，虽然多状态激活函数在训练集上表现不佳，但是在测试集上有较好的表现，具有一定的抵抗过拟合的能力。

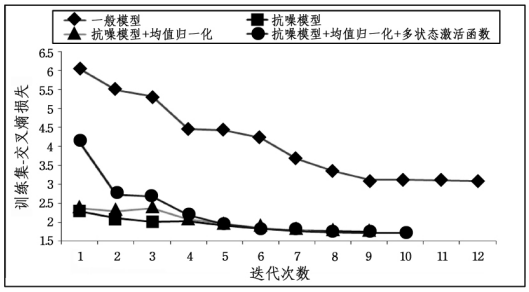

图4　测试集上的交叉熵损失变化曲线

最后，进行错误率的测试。先使用随机抽取的训练集的一部分进行集内测试；然后分别使用不带噪声和带噪声的测试集进行集外测试，以此得到各个模型的错误率。不同方法的深度神经网络模型的语音识别错误率如表1所列。与一般模型相比，抗噪模型在集内测试和不带噪声测试中表现相对较差；而在带噪声测试中表现较好。这是因为在抗噪训练中，纯净信号的输出和带噪声信号的输出之间的差异被缩小，纯净信号受到了噪声的干扰，而带噪声信号中的噪声被减弱。在使用了均值归一化进行优化后，集内测试、不带噪声测试和带噪声测试中模型的语音识别错误率都略微下降。而在使用了多状态激活函数进行优化之后，不带噪声测试和带噪声测试的错误率又略微下降。

表1　深度神经网络的错误率(%)

模型	集内测试错误率	集外-不带噪声测试错误率	集外-带噪声测试错误率
一般模型	8.43	10.59	41.54
抗噪模型	10.83	11.90	25.35
抗噪模型＋均值归一化	9.42	11.33	24.77
抗噪模型＋均值归一化＋多状态激活函数	9.42	11.10	24.36

结束语　本文构建了基于深度神经网络的语音识别系统，并使用小规模语音测试集进行了实验，实验结果表明：

(1)抗噪对比散度法和最小平方误差法的结合使深度信任网络具有一定的抗噪性能。在小规模连续语音识别实验中，经过这两种方法训练的模型对纯净语音的识别能力减弱，而对带噪声语音的识别能力增强。

(2)在有监督训练阶段，多状态激活函数和均值归一化的

随机梯度下降法能够对模型进行优化。在小规模连续语音识别实验中，均值归一化的随机梯度下降法能够在一定程度上降低语音识别系统的识别错误率，在与多状态激活函数相结合后，系统的识别错误率再一次降低。

(3)使用奇异值分解与重构法可以对网络进行降维。在小规模连续语音识别实验中，使用这种方法能够使得网络的参数数量减少为原有的 0.49 倍，并且性能仅受到轻微的影响。

本文主要研究了语音识别系统中的声学模型，而语音识别系统的性能不仅受到声学模型的影响还受到语言模型的影响。为了进一步提升系统性能并且降低系统的硬件资源消耗，未来将考虑对语言模型的解码算法进行优化。

参 考 文 献

[1] Vincent P, Larocheiie H, Lajoie I, et al. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion[J]. Journal of Machine Learning Research, 2010, 11: 3371-3408

[2] Martens J. Deep learning via hessian-free optimization[C]// Proceedings of the 27th International Conference on Machine Learning (ICML-10). Israel: Haifa, 2010: 735-742

[3] Dean J, Corrado G, Monga R, et al. Large scale distributed deep networks[C]// Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Lake Tahoe, Nevada, United States: Proceedings of a meeting held. 2012: 1232-1240

[4] Deng W, Qian Y, Fan Y, et al. Stochastic data sweeping for fast DNN training[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: ICASSP 2014, 2014: 240-244

[5] You Zhan, Wang Xiao-rui, Xu Bo. Exploring one pass learning for deep neural network training with averaged stochastic gradient descent[C]// ICASSP 2014. 2014: 6854-6858

[6] Xue J, Li J, Gong Y. Restructuring of deep neural network acoustic models with singular value decomposition[C]// INTERSPEECH 2013. Lyon, France: 14th Annual Conference of the International Speech Communication Association, 2013: 2365-2369

[7] He Y, Qian F Y, et al. Reshaping deep neural network for fast decoding by node-pruning[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: ICASSP 2014, 2014: 245-249

[8] Graves A, Mohamed A R, Geoffrey E. HINTON. Speech recognition with deep recurrent neural networks[C]// ICASSP 2013. 2013: 6645-6649

[9] Sarikaya R, Hinton G E, Deoras A. Application of Deep Belief Networks for Natural Language Understanding[J]. IEEE/ACM Transactions on Audio, Speech & Language Processing, 2014, 22(4): 778-784

[10] Mohamed A, Dahl G E, Hinton G E. Acoustic modeling using deep belief networks[J]. IEEE Transactions on Audio, Speech & Language Processing, 2012, 20(1): 14-22

(上接第 34 页)

5.3 分析与讨论

从图 2 可以看出，WClipper 在新华网、新浪网、凤凰网、农博网、金农网上的查全率表现跟 Evernote 工具和 Readability 很接近，且查全率接近 1。在搜狐网、环球网上的查全率表现不如 Evernote 工具和 Readability，是因为 WClipper 遗漏掉了主题内容的一些附加信息，比如文章来源。这是受限于本文的算法特性，文章来源一般以链接的形式存在，会被识别为链接型节点，而直接忽略掉。从图 3 看出，WClipper 在搜狐网、凤凰网、环球网上的查准率低于其他工具，是因为网站的部分页面含有评论区域并且评论区域含有比较多的文本内容。评论区的内容不属于主题信息的一部分，但由于 NTA 算法并未对文本节点做出进一步的区分，因此最后都整合到一起，影响了查准率。从图 4 中 F1 指标来看，WClipper 工具的综合提取效果比 Evernote 工具高出 0.3%，比 Ynote 工具高出 5.01%，这在一定程度上验证了本文方法的有效性。

结束语 在前人工作的基础上结合对网页噪声特点以及网页性质的观察和统计，提出了一种基于 DOM 节点类型标注的主题信息抽取方法。将 DOM 节点划分为 4 种类型并依据节点内聚度、节点文本密度、阈值等统计信息实现网页主题内容的抽取。初步试验验证了本文方法的有效性，显示了本文方法相比其他同类工具表现出较好的主题信息抽取效果。由于该方法不依赖特定标签且只规定了较少的启发式规则，因此具有较好的通用性和算法效率。但是也存在一些不足，进一步的研究和改进包括：识别和去除网页中不是超链接形式的噪声，如评论文字、版权声明等，可以借助网页中的视觉特征来加以甄别。

参 考 文 献

[1] Gibson, David, Punera K, et al. The volume and evolution of Web page templates[C]// Special Interest Tracks and Posters of the 14th International Conference on World Wide Web. ACM, 2005

[2] Wang Ji-ying, Lochovsky F H. Data-rich section extraction from html pages[C]// Proceedings of the Third International Conference on Web Information Systems Engineering, 2002 (WISE 2002). IEEE, 2002: 313-322

[3] Yi L, Liu B, Li X. Eliminating noisy information in web pages for data mining[C]// Proceedings of the 9th ACM SIGKDD Int Conference on Knowledge Discovery and Data Mining. New York: ACM, 2003: 296-305

[4] 欧健文，董守斌，蔡斌. 模板化网页主题信息的提取方法[J]. 清华大学学报(自然科学版)，2008(S1)：1743-1747

[5] Bauer, Daniel, et al. FIASCO: Filtering the Internet by Automatic Subtree Classification, Osnabruck. Building and Exploring Web Corpora[C]// Proceedings of the 3rd Web as Corpus Workshop, Incorporating Cleaneval. Vol. 4. 2007

[6] Lin S H, Ho J M. Discovering informative content blocks from Web documents[C]// Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2002

[7] 时达明，林鸿飞，杨志豪. 基于网页框架和规则的网页噪音去除方法[J]. 计算机工程，2007，33(19)：276-278

[8] Cai Deng, et al. VIPS: a vision based page segmentation algorithm. Microsoft technical report[R]. MSR-TR-2003-79, 2003

[9] 邹永强，钟志农. 一种高效的新闻网页噪声过滤方法[J]. 微型机与应用，2011，30(16)：64-67