中 文 信 息 学 报
第 20卷 第 3期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 20 No. 3

文章编号: 1003 - 0077(2006)03 - 0022 - 07

中文文本分类中基于概念屏蔽层的特征提取方法*

廖莎莎,江铭虎

(清华大学 人文学院计算语言实验室,清华大学 认知科学创新基地,北京 100084)

摘要:本文提出了一种新的基于概念抽取和屏蔽层的特征选择方法。该方法利用 HowNet概念词典中的概念树,通过义原在概念树中的位置信息进行概念抽取,并赋予其适当权值来说明其描述能力。对于权值低于屏蔽层的义原,我们不将其选入特征集,并相应保留原词。具体到每个词,我们计算其 DEF条目中的权值,决定是将原词选入特征集还是进行概念抽取。本文重点研究了如何给义原设定一个合适的权值,如何在选取原词和概念之间取得平衡以及针对非概念词的加权处理。实验证明,设定合适的屏蔽层,不仅可以缩小特征维数,使分类正确率得到一定的提高,而且可以减少不同类别间的分类正确率的差别。

关键词:计算机应用;中文信息处理;文本分类;特征提取;概念抽取;属性特征树;屏蔽层;描述能力

中图分类号: TP391 **文献标识码**: A

A Feature Selection Method in Chinese Text Classification Based on Concept Extraction with a Shielded Level

LIAO Sha-sha, JIANG Ming-hu

(Lab of Computational Linguistics of Chinese Language, Cognitive Sciences Innovation Base
Tsinghua Univ., Beijing 100084, China)

Abstract: In this paper, we propose a novel feature selection method based on concept extraction and shielded level. In this method, we use HowNet as the semantic dictionary to extract concept attributes. Based on their positions in the concept tree, the attributes will get proper weights, which present their description powers. A concept attribute will not be selected as feature if its weight is lower than the shielded level and the original word will be reserved for use. To each word, we calculate all the weights of the concept attributes in its DEF, and decide whether to extract the concept attributes or reserve the word. We focus mainly on how to weight the concept attributes, how to make a balance between concept features and word features, and how to treat the words out of the dictionary. The experiment shows that if a shielded level is set properly, it can not only reduce the feature dimension to a proper scale but also improve the classification precise. Moreover, it can reduce the difference of the classification precise among different categories.

Key words: computer application; Chinese information processing; text classification; feature selection; concept extraction; concept tree; shielded level; description power

1 引言

文本分类(Text Classification)是指依据文本内容,把待定文本分到预先定义好的类别。国

* **收稿日期**: 2005 - 03 - 21 **定稿日期**: 2005 - 10 - 11
基金项目:教育部优秀青年教师资助计划项目(2051);中国科学院模式识别国家重点实验室开放课题基金(10);2003年度清华大学 985 - I期基础研究基金的资助.
作者简介:廖莎莎(1981—),女,硕士研究生,主要研究方向为自然语言处理.

外的文本自动分类研究已经从最初的可行性性基础研究经历了试验性研究进入到了实用化阶段,其中,较为成功的有麻省理工学院(MIT)为白宫开发的邮件分类系统,卡内基集团为路透社开发的 Construe系统等。国内对于文本自动分类的研究起步于 80年代,随着中文信息处理技术特别是中文自动分词技术的日渐成熟,以此为基础的中文文本分类技术的研究得到了快速发展。例如,广东省中山图书馆的计算机辅助图书分类系统、清华大学的自动分类系统、山西大学开发的金融自动分类系统、东北大学图书馆的图书馆分类专家系统。

文本分类大致可以分为三步,文本的向量模型表示,特征选择和分类器构造。训练集的巨大和较高的向量维数是文本分类的两大特点,因此,为了兼顾运算速度和分类精度,我们必需进行合适的特征提取[1]。特征选择通常要寻找包含了原始属性中必要信息的最小特征集,即所谓维数约简,同时根据特征项对分类的贡献度的大小进行特征加权。现行的特征集选择包括词特征选择,字特征选择和概念特征选择等,其中,先行的实际分类系统主要以选择词特征为主,有的在特定领域加入一些人工规则等。但是,由于词语本身存在同义、多义以及对短语和上下文的依赖等现象,因此,单纯基于词形的技术中,把意义可能密切相关的词孤立提取,忽略了词语的语言学特征和相互关系,因此导致这种特征提取存在较大的局限性。例如,传统的向量空间模型最基本的假设是各个分量之间正交,而实际上在真实文本中,作为分量的词特征往往有很大的相关性[3]。

为了避免词特征选择中出现的种种问题,我们利用语义词典中的信息,抽取概念来构成文本向量,由于概念空间比词空间小而且各分量之间相对独立,因此,概念特征比词特征更适合用来表示文本内容。研究表明,通过概念统计和语义层次分析的方法,可以获得更理想的向量空间模型[2]。因此,我们选择概念特征作为本系统中特征提取的主体[3]。中文文本分类中,选取概念作为分类特征的系统相对较少,基本处于研究阶段。例如,厦门大学计算科学系[4],华中师范大学计算科学系[5],清华大学计算机技术与科学系[3]都进行了相关的实验和报告。这些方法利用了概念特征进行分类,并结合了位置,词性等信息抽取关键词,并取得了较好的实验结果。

2 基于屏蔽层的概念抽取方法及评价

本系统中,我们首先采用 HowNet概念词典对文本进行基于词的概念抽取和基于屏蔽层的概念本身的加权,包括如何赋予概念和原词适当的权值,以及对非概念词,多义词的处理;然后利用词频 期望交叉熵实现基于特征的类间分布差异的加权。通过对概念树的层次屏蔽,探讨了概念的描述能力的强弱以及如何在概念和原词之间取得平衡的问题。通过概念加权,我们给予描述能力不同的特征以不同的权值,加强了那些意义明确,对分类作用明显的特征在分类中的作用;而基于期望交叉熵的加权,通过特征在文档和类别间出现频率的差别,消除了文本中出现的稀疏词和模糊词[6],再次加强那些对分类有用的特征。

2 1 基于概念层次的特征提取

本系统所用的概念属性定义来源于 HowNet(知网)[7]。它是一个以概念为描述对象的知识系统,知网中用义原作为概念单位,并利用知识词典为每个词定义一个 DEF词条,DEF词条由一个或者多个义原构成,它是我们进行概念抽取的依据,我们可以通过提取词的 DEF作为文档特征来获得文章内容在属性上的规律,挖掘出原本分散在词语表面中的内在联系,集中特征,突出表现文章的主题,以此来获取更加适合于文本分类的特征选项。

2.1.1 HowNet属性特征加权算法

因为在 HowNet中附带的属性特征表中,是按语义层次结构列出 1505个义原,因此,义原

在选取为特征时应该不同对待,一般的概念特征加权算法是利用节点到树内根节点的距离。计算公式为[3]:

$$W_{ik} = W\,tree_i \cdot [\log(D\,root_{ik} + a) + L] \quad (1)$$

其中,W_{ik}表示第 i棵树中节点 k的权值;$W\,tree_i$用来调整各棵树的总体重要性,即根节点的权值;$D\,root_{ik}$为节点 k到所在树的根节点的距离;L为调和因子。

在本系统中,对该公式进行改进,不仅利用节点到树内根节点的距离,同时利用该节点的子节点的数量作为判定该节点权值的依据。概念数中同层节点的描述能力不完全相同,例如,"时间"就比"万物"的描述能力高,因为它的子节点较少,而"万物"由于拥有较多的子节点而在词条中出现较少,例如"底价"的 DEF为"attribute|属性,price|价格,&thing|万物,commercial|商",而"点"的 DEF为"time|时间,hour|时",明显"时间"比"万物"的描述能力高。

同时,由于不同的树中相同的层次的描述能力并不相同,因此需要对不同的概念树赋予一个总体重要度,概念特征加权的计算公式为:

$$k(m) = W\,tree_i \cdot [\log((Deep_k + 1)/2) + a + \frac{1}{Child_k + b}] \quad (2)$$

其中,$k(m)$表示义原 m的权值,i为 m所在的义原树的编号,k为 m在该义原树中的位置。$W\,tree_i$用来调整各棵树的总体重要性,即根节点的权值,由于"*Attribute*"树,"*Attribute Value*"树"*Event Role && Features*"树,"*Quantity*"树,"*Quantity Value*"树,"*Syntax*"树这六棵树对分类意义较小,它们的 $W\,tree$的重要度为 0.1;"*Entity*"树和"*Event*"树的重要度较高,由于前者大多为名词概念,后者大多为动词概念,因此根据经验,设前者的 $W\,tree$为 1.0,后者为 0.25;$Deep_k$表示该义原的高度,$Child_k$为该义原的下位义原的数目,其值从索引表中获得;a, b为调和因子,是用来控制权值范围,防止权值为负值。实验证明,当 b值取值在 [3, 15]之间时对结果的影响无明显差异,最后实验中,沿用公式 (1)中的 a值 0.15,设定 b值为 5.0。

2.1.2 概念树中的屏蔽层和阈值过滤

由于 HowNet网络中的特征是以树的形式存在,因此存在一些概念比较抽象的义原,如果一个名词的 DEF属性都是抽象义原,则无法体现该词的语义。例如"平等互利"的 DEF为"属性|异同|似|实体",这些义原都没有完整的体现出"平等互利"的概念。如果将原词转化为这样的义原,这些义原极有可能成为对分类没有用处的高频特征,使分类错误率大大提高。如何在原词和概念抽取之间取得平衡,成为使用概念词典首先要解决的问题。为了解决这个问题,我们利用知网信息,构造了一个概念层次树,并通过计算各个义原在树上的层次来确定其描述能力,并将其分为强描述能力义原和弱描述能力义原两种。由于是通过树层来区分义原的描述能力,因此,我们设定了一个阈值,当义原所处的层次低于阈值的时候,将它屏蔽,否则,则加入特征集,这个阈值就是概念树中的屏蔽层。设置屏蔽层的依据是,当描述一个词的若干概念的描述能力都较弱时,可以认为他们并没有提供比原词更多的信息量,因此略过对该词的概念化处理,直接将原词选为特征。计算一个词的 DEF词条的描述能力的公式为:

$$f(c) = \max_{j=0}^{m} k(c_j) \quad (3)$$

其中,$k(c_j)$为词 c的 DEF词条中第 j个义原的权值 (我们先通过概念树表获得 $k(c_j)$的 W_{ik},然后利用公式 (2)获得该义原的权值),m为该词 DEF词条中概念特征的总数目。该公式计算词 c所有义原的描述能力,并通过设定阈值来决定选择进行概念抽取或者保留原词。例如,对于词"程序控制:DEF ="控制|软件",由于概念树中"控制"在 Event树的第 5层,"软件"在 Entity树的第 7层,根据公式 (3)进行计算后,"控制"低于阈值而"软件"高于阈值,于是

将词“程序控制”进行概念抽取,义原“软件”选入特征集,而“控制”不选入特征集。后面的实验部分证明,设定合适的阈值所形成的特征集可以提高分类正确率。

2.1.3 对非概念词的处理

由于概念词典本身的不完整性,对于词典中没有出现的词语,需要进行一定的加权处理。我们在进行原词加权时,考虑到词语的词长信息,认为多字词(字数大于等于4时)成为特定领域的专有名词的可能性较高(例如“市场经济”“人民法院”“经济基础”“民主集中制”等),因此给予其一定程度的补偿,同时,为了防止给多字词赋予过高权重,我们采用取log值的方法进行调和,实验证明,取log值的方法可以防止一些词长过长的词,例如七字词获得过高的权值。公式如下:

$$W_i = a \cdot \log(length(i)) \tag{4}$$

公式中的a值的作用主要是调整非概念词的权重,使之与概念词的平均权重接近,以免使特征集中两类特征的权重差别过大,影响分类结果。实验中,a的取值为0.5。

2.1.4 多义词的处理

对于多义词,有多个DEF定义,由于进行排歧的技术相对比较复杂,系统开销太大,因此本系统不进行排歧处理,只进行一个初步处理方法:首先提取各DEF内的义原并求交集,用新的不为空的交集作为该词新的DEF。如果交集为空,则仍然以词本身作为特征选入。

2.2 基于期望交叉熵的特征加权和约简

在信息论中,IG(Information Gain)信息熵算法是用来度量系统包含信息量的多少或系统有序程度的值。熵越大,系统包含的信息量越大,系统有序程度越低。信息熵被广泛的应用于文本分类研究中,并出现了条件熵,相对熵,联合熵和交叉熵等概念,信息熵是选择效果最好的统计量之一[8,9]。

期望交叉熵又称相对熵,也称KL距离。它考虑了文档集合的信息熵和文档中词语的条件熵之间信息量的增益关系,并以此来确定该词语在文本分类中所能提供的信息量,即词语在分类中的重要程度。它和IG算法的不同之处在于它不考虑特征未出现的情况,在对比实验[10]中,用期望交叉熵进行特征选择优于信息熵。期望交叉熵的公式[10]为:

$$RE(t) = P(t) \sum_i P(c_i \mid t) \log \frac{P(c_i \mid t)}{P(c_i)} \tag{5}$$

其中,$P(t)$为特征t出现的概率,$P(c_i \mid t)$为类别c_i在特征t出现情况下的概率,$P(c_i)$为类别c_i的出现概率。

3 实验结果与分析

由于中文文本分类现在还没有一个标准的公认测试集,因此,我们实验室人工整理了《人民日报》95～98年的电子光盘版,并根据人民日报的栏目分类,将文章大致分为六类,1960篇文档。我们从全部文档中随机抽取出训练集和测试集,其中,训练文档2.08MB,测试文档1.11MB,整理后的语料分类如下:

	经济	政治	电脑	体育	教育	法律	合计
训练集篇数	250	175	130	300	150	200	1205
测试集篇数	121	82	55	282	63	152	755

本文的实验系统是在WindowsXP系统下,使用VS.Net作为开发工具进行开发的,其中,使用支持向量机(SVM)构造分类器,该方法被认为是文本分类中表现最好的分类算法之

一[11]。我们使用随即抽取的训练集用来进行特征抽取并构造分类器,并利用测试集进行开放测试,获得分类结果。

实验表明,改进的概念抽取方法在和期望交叉熵加权约简方法结合后,可以通过各自的方式有效的缩减特征维数,而且在缩减过程中,不仅没有降低分类正确性,反而可以去除不必要的噪声,使分类效果更加理想。当特征约简中的阈值设为 3.0时,实验结果如下:

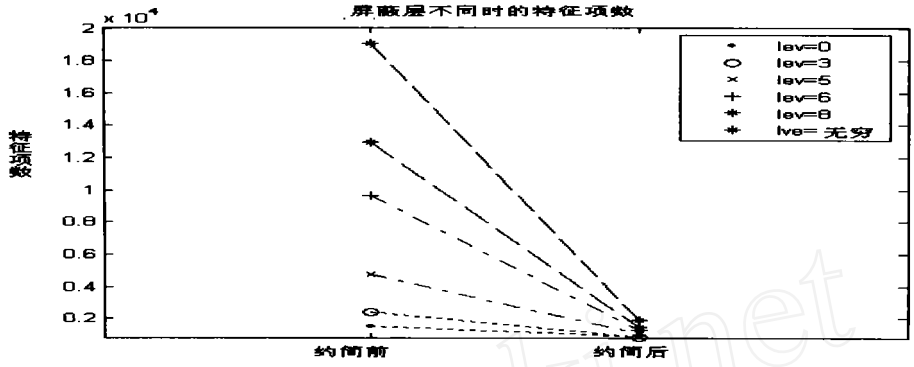

图 1 屏蔽层不同时的特征约简度

从图 1中可以看出,通过将原词映射到概念空间,我们可以获得一个较小的特征空间,对于分类效果最好的 level=6而言,使用概念词典后的概念空间为原词空间的 0.4倍左右,说明在概念空间中,将特征映射到一个相对稳定,数值较大的空间,使得其在每个方向的映射较强,这对文本分类是很有力的。同时,当在特征约简算法中设定一个阈值 ($\theta = 3.0$)时,概念空间的约简度相对较小,也就是说被过滤的概念较少,保留了原特征集中 17.4%的高频特征,而使用原词时,在同样的阈值下,只保留了原特征集中 10.0%的高频特征。可以看出,利用概念可以更多的保留原特征集中的信息。

在分类正确率方面,我们对上述测试集进行测试,使用召回率作为衡量分类质量好坏的标准。某个类别的分类召回率定义为:

$$Recall = \frac{CorrectNum}{TotalNum} \tag{6}$$

其中,$CorrectNum$ 表示通过分类算法正确分类的文档数目;$TotalNum$ 表示该类别中总的文档数目,本实验将六个类别的召回率的平均值作为分类的召回率。

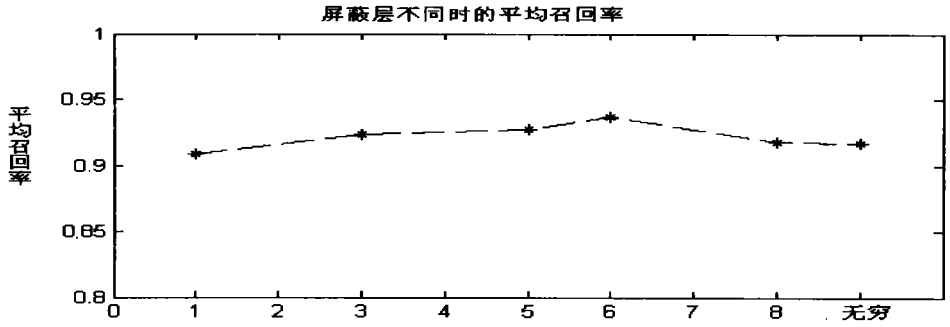

图 2 屏蔽层不同时的平均召回率

从图 2中可以看出,单纯利用词频 (即 lev=无穷)和概念的分类效果都不是最好的。实验证明,完全使用 HowNet中的概念作为特征,分类的召回率反而有所下降,仅为 90.9%,低于使用原词作为特征时的 91.7%。而通过设定概念屏蔽层,随着阈值的改变,分类效果有明显的

改变,最高可以达到 93.7%。此时屏蔽层为 6,这个屏蔽层滤去了大多数描述能力较弱的抽象概念,同时将原词加入特征集。

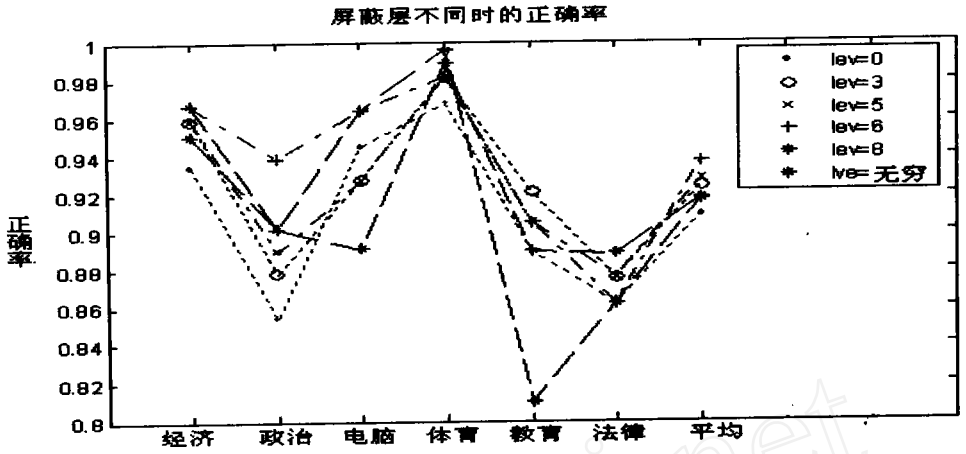

图 3 基于不同屏蔽层时各类召回率

对于不同类别的正确率,图 3显示的结果表明,基于词的特征选择对类别中的文档本身要求较高,不同类别的召回率差别较大,可以从 81.0%变化到 96.7%,表现为图中的曲线波动较大;而基于概念的特征选择在不同类别间的召回率差别相对较小,曲线相对较为平滑。类别间分类率出现差别的最可能原因是在基于词的特征选择中,不同类别的关键词的数量和出现频率是不同的,如果某一个类专用的关键词越多,出现的频率越高,那么在向量中所占的比例就越大,该类的召回率就越好。例如体育类中由于包含了较多的专业词汇,因此召回率是最好的。而利用概念之后,差别变小,其原因可能是语义词典将大量专业词汇转化为概念特征,融合了那些语义相关的特征,使各类的关键词的数量和频率上的差别变小。

4 结论

使用概念作为文本分类中的特征项,可以有效的缩减特征维数,并将原词映射到一个相对稳定的维数空间,但由于弱概念词的存在,使得原词转化为概念时可能会丢失一些重要的分类信息,因此,通过屏蔽概念树中的上层节点,可以将概念描述能力不足的弱概念词保留,这样虽然使得维数空间有一定的增加,但变化并不是很大,尤其是设定了一个较低的阈值滤去那些出现次数太小的特征之后。在分类正确率方面,如果选择合适的屏蔽层,分类效果也会有明显的提高,同时可以减少不同类别之间分类正确率的差别,消除分类器对某些类的偏好。实验表明,通过基于屏蔽层的概念提取方法,可以将原词和概念较好的结合在一起,达到较高的分类率。将来的工作主要集中在将该方法应用到更大规模的语料库中,并将 HowNet中其他的信息运用到概念抽取中来。

参 考 文 献:

[1] 周茜,赵明生,扈雯. 中文文本分类中的特征选择研究 [J]. 中文信息学报, 2004, 18(3): 17 - 23.

[2] 季妲,罗振声,万敏,高小云. 基于概念统计和语义层次分析的英文自动文摘研究 [J]. 中文信息学报, 2003, 17(2): 14 - 20.

[3] 李蕴,罗振声,厉宇航. 基于语义相关和概念相关的自动分类方法研究 [J]. 计算机工程与应用, 2003, (12): 106 - 109.

[4] 苏伟峰,李绍滋,李堂秋. 一个基于概念的中文文本分类模型 [J]. 计算机工程与应用, 2002, (6): 193 -

195.
[5] 王萌,何婷婷,姬东鸿,王晓荣. 基于 HowNet概念获取的中文自动文摘系统 [J]. 中文信息学报, 2005, 19(3): 440 - 446.
[6] 钱铁云,王元珍,冯小年. 结合类频率的关联中文文本分类 [J]. 中文信息学报, 2004, 18(6): 30 - 36.
[7] Dong Zhengdong, Dong Qiang the download of Hownet[EB/OL], http://www.keenage.com.
[8] Yang Yimin, and Pedersen J O. A comparative study on feature selection in text categorization[A]. In: proceedings of the 14th International Conference on Mahine Learning(ICML - 97) [C], 1997.
[9] 代六玲,黄河燕,陈肇雄. 中文文本分类中特征抽取方法的比较研究 [J]. 中文信息学报, 2004, 18(1): 26 - 32.
[10] 李凡,鲁明羽,陆玉昌. 关于文本特征抽取新方法的研究 [J]. 清华大学学报 (自然科学版), 2001, (7): 99 - 102.
[11] FABRIZIO SEBASTIANI machine learning in automated text categorization[C]. ACM computing surveys, Vol 34, No 1, March 2002, P1.

(上接第 5页)

用的训练语料规模,要比其他方法少的多。同时应用了二叉树剪枝方法,也提高了系统的运行效率。

5 结束语

本文采用了最大熵方法识别输入句子的韵律短语边界,在搜索最佳识别结果时,引入了二叉树方法作为剪枝策略,对不满足韵律短语不完全覆盖二叉树子树的韵律短语对应的弧进行剪枝,大大缩小了搜索空间,提高了识别 F-Score近 35%,在小训练语料上,平均 F-Score达到了 80.8%。

参 考 文 献:

[1] 赵晟,陶建华,蔡莲红. 基于规则学习的韵律结构预测 [J]. 中文信息学报, 2002, 16(5): 30 - 37.
[2] 牛正雨,柴佩琪. 基于边界点词性特征统计的韵律短语切分 [J]. 中文信息学报, 2001, 15(5): 19 - 25.
[3] 应宏,蔡莲红. 基于结构助词驱动韵律短语界定的研究 [J]. 中文信息学报, 1999, 13(6): 41 - 46.
[4] 曹剑芬. 基于语法信息的汉语韵律结构预测 [J]. 中文信息学报, 2003, 17(3): 41 - 46.
[5] 李剑锋,胡国平,王仁华. 基于最大熵模型的韵律短语边界预测 [J]. 中文信息学报, 2004, 18(5): 56 - 63.
[6] 叶竹钧. 朗读中的停顿探析 [J]. 语文教学通讯, 1995, (Zl): 78 - 79, 1995, (7): 30 - 31.
[7] 汪国胜. 标点符号概说 [J]. 高等函授学报 (哲学社会科学版), 1996, (6): 19 - 23.
[8] 中华人民共和国国标《标点符号用法》, 1996, 6.
[9] Min Chu, Yao Qian, Locating Boundaries for Prosodic Constituents in Unrestricted Mandarin Texts[J], 2001, Computational Linguistics and Chinese Language processing, Vol 6, No. 1, 61 - 83.
[10] 赵永贞,刘挺,王志伟,陈惠鹏,邵艳秋. 汉语文语转换系统中停顿指数的自动标注 [J]. 中文信息学报, 2004, 18(5): 48 - 55.
[11] 聂鑫,王作英. 汉语语句中短语间停顿的自动预测方法 [J]. 中文信息学报, 2003, 17(4): 39 - 44.
[12] 吴志勇,蔡莲红. 语音合成中韵律关联模型 [J]. 中文信息学报, 2004, 18(2): 44 - 50.