一种新颖的自然语言主题转换精确定位方法*

陈浪舟 黄泰翼

(中国科学院自动化研究所模式识别实验室 北京 100080)

E-mail: huang@nlpr.ia.ac.cn

摘要 自然语言的主题转换是自然语言理解的一个重要线索.语言处理通常是针对不同的主题有不同的数据库和处理方法.因此,如何找到文本中的主题转换点是语言处理中的一个重要内容.该技术在语言理解、文本自动索引以及语言模型的建立等方面都有重要意义.该文以文本主题转换时的词汇突变为表征,提出和定义了反映词汇突变的4个参数,将这4个参数作为输入,利用BP网作为判决工具,建立了一个在不同尺度下文本词汇变化的层次结构模型,实现了一种精确的文本主题转换点的定位方法,其定位精度在一个句子左右.

关键词 自然语言处理,文本切分,文本索引和过滤,BP算法.

中图法分类号 TP18

文本主题转换点的定位是自然语言处理的一个重要内容,具有广泛的应用领域.例如,语言理解、文本主题转换点的定位可以认为是对文本结构的分析,因此,对文本的理解有重要意义.又如,在对话系统的设计中,系统必须密切跟踪话题的转移[1].文本的自动索引[2]也离不开文本主题转换点的确定,只有将文本按领域分割以后才能从中找出我们所需要的内容.与自动索引相对应,文本主题转换定位的另一个用途是文本过滤.现代社会信息高度发达,各种媒体,尤其是在Internet上传送的大量文本数据中,很大一部分是人们所不需要的,如果想将它们自动滤除,也必须进行文本主题定位.因此,研究文本分割具有重要意义.建立适用于语音识别的、领域自适应的统计语言模型也离不开这种按领域分割文本的技术.由于我们所处理的语料都是连续无切分标志的文本,因此首要的任务就是将文本按主题切分为具有明显主题的段落,然后将这些段落按领域聚类,从而形成领域相关的语料,作为建立语言模型的依据.总之,文本主题转换点的定位技术在语言处理的各个应用领域都有重要意义.

文本的主题转换在某些应用场合并不需要太高的精度,如建立领域相关的统计语言模型,需要将几十兆甚至上百兆的语料按领域进行切割.在这种情况下,主题转换的定位精度在几千字节的语料以内都能满足要求.但是,在有些应用场合,人机对话系统,需要随时跟踪用户话题的转移,有时定位必须精确到句子,这种应用为我们的工作提出了更高的要求.本文的任务就是高精度的定位文本的主题转换点.

大家知道,自然语言的任何一段有意义的文本都是围绕一个或者多个主题展开的,跟踪文本主题的变化对文本的理解和其他深入处理都具有重要意义.检测文本的主题变化通常都是以文本中的词汇作为判断依据的.例如,著名的词汇链理论[3],就是依据文本中词汇的重复与搭配等关系,形成一条词汇链,进而确定文本的结构.显然,该方法可以有效地用于文本切分.但是,单一的词汇链有一个缺点,即它对多个主题同时出现的情况比较困难.

文献[4]提出的用领域相关统计语言模型对文本似然度的变化作为判断文本领域变化的方法虽然有效,但需要一个已经建好的领域相关的统计语言模型,而这种模型的建立需要很大的工作量而且本身就需要大量的已

* 本文研究得到国家自然科学基金资助.作者**陈浪舟**,1971年生,博士,主要研究领域为统计语言模型,语音识别.**黄泰翼**,1934年生,研究员,博士生导师,主要研究领域为语言信息处理,智能人机通信.

本文通讯联系人:黄泰翼,北京 100080,中国科学院自动化研究所模式识别实验室

本文1998-09-25收到原稿,1999-01-14收到修改稿

经按领域分类的文本,因此这种方法的应用也受到限制.

文献[5]利用文本主题转换时词汇的突变作为依据,研究文本主题切换的定位.由于自然语言在每个不同的领域都有该领域特有的词汇,这些词汇在领域内频繁出现,而在领域以外出现的概率则很小,因此,文本在两个领域交接处词汇的变化很大.该方法正是利用这一现象来判断文本的主体转换.该方法的主要优点是可以处理多主题并发的现象,缺点是在实际应用中常常因受到局部变化的影响而形成误判.

本文提出的方法也是利用词汇的突变来定位主题的切换点,但与文献[5]不同的是,我们不是直接通过检测词汇的变化来作判断.首先,本文从反映词汇变化的相似函数曲线中提取可靠的特征,包括波谷的绝对高度.波谷与波峰的高度差以及相对位置关系等.这些特征更加全面、可靠地反映了文本词汇在主题转换处的特性.其次,本文将这些特征作为输入,训练一个 BP 网来实现对文本主题转换的自动定位.最后,在此基础上,本文提出了一种基于层次结构的定位方法,利用在不同滑动步长下相似函数对词汇变化细节信息的检测能力的不同,构成词汇变化的层次结构,提出了一种基于 BP 算法和层次结构相结合的文本主题变化定位算法,该算法的定位精度在一个句子左右,能满足精度要求很高的需求,同时,这种算法具有极强的鲁棒性,大大减少了由于局部变化影响造成的漏判和误判情况.

1 文本主题转换的词汇突变现象

自然语言在每一领域都有其特殊的关键词,这些关键词具有突发性,即它们在某一特定领域频繁出现,而在其他领域很少出现.这种现象反映在文本的主题变化处,出现了词汇的大量转移.而这种突变正是我们用来进行文本切分的特征.

如图 1 所示,假设无切分连续文本流为…$W_{k-1}W_kW_{k+1}$…,我们用两个宽度相同的相邻滑动窗口 $Wind_1$ 和 $Wind_2$ 按一定的步长依次滑过整个文本流,其中窗口宽度为 n.

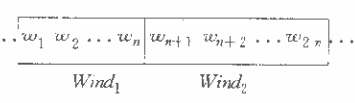


图1

对每个窗口中的文本,我们首先去掉所有的虚词,然后按下式计算两个窗口实词之间的相似度函数.

$$sim(i)=\frac{\sum_{k=1}^{K}Num1(W_k)*Num2(W_k)}{\sqrt{\sum_{k=1}^{K}Num1(W_k)^2*\sum_{k=1}^{K}Num2(W_k)^2}},\tag{1}$$

其中 $Num1(W_k)$为窗口 1 中词 W_k 的出现次数,$Num2(W_k)$为窗口 2 中词 W_k 的出现次数,k 为词表大小,i 为当前位置.当窗口按一定步长沿文本流滑动时,我们将得到一系列的 $sim(i)$,由于词汇突变的原因,这些值在文本的主题转移处会表现为局部最小,如图 2 所示.

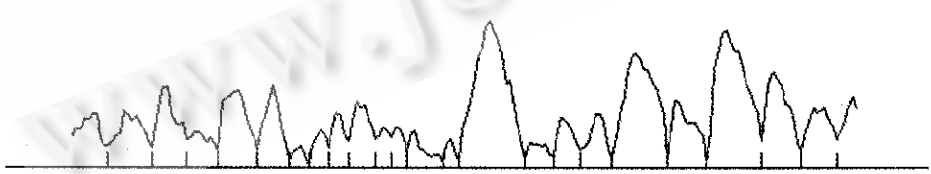
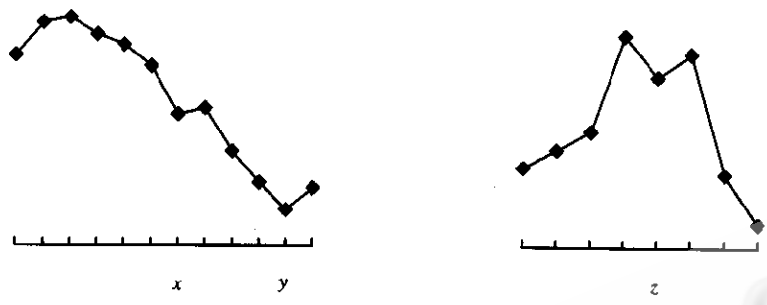


图 2

图 2 为 $sim(i)$的一段波形.图中横轴上垂直线所标记的位置为文本的主题转换处.我们可以看到,主题转换通常发生在波谷所对应的位置.

2 特征参数的选取

虽然主题转换通常发生在波谷所对应的位置,但并不是所有的波谷都对应着主题的转换.由于文本流中局部词汇变化的存在,可能在相似度函数中产生局部扰动,这些扰动通常会造成误判.如图 3 所示.



(a)　　(b)

图3

图3为图2波形中两处局部扰动的放大.在图3(a)中,波形在向最低点下降的过程中出现了一个局部上升点x,而真实的主题转换点显然发生在y处.这种情况通常发生在一个主题尚未结束的时候,图1中的窗口1还处在上一个主题,而窗口2已到达头一个主题与下一个主题交接处.此时,波形向波谷发展,假如在此过程中,文本2和文本1的词汇发生少量重叠,就会产生如图3(a)所示的情况.造成这种扰动的主要原因是,大量与领域无关的词汇平均分布在各主题,因此造成噪声.滤除这些领域无关词汇可以减少这种扰动的发生.

图3(b)所示的是另一种扰动情况,在波形的峰值处出现了一个局部凹陷.事实上,这个局部凹陷显然也不是主题转换位置.这种扰动主要出现在同一主题内部,即窗口1和窗口2都处在一个主题内,同一主题内的局部词汇变动造成了这种扰动.

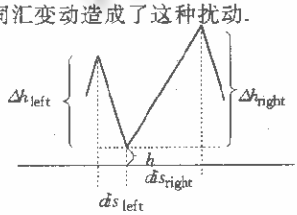


图4

为避免局部扰动的影响,准确地定位主题转换点,我们必须精心设计特征参数.首先,波形的幅度代表了当前位置词汇的绝对相似程度,这是我们必须考虑的因素.其次,我们不但需要考虑绝对相似程度,还需要考虑词汇变化的相对幅度,即波峰与波谷之间的落差.除此之外,波峰和波谷之间的距离通常反映了词汇相似度的衰减速度,这也是一个非常有用的信息.综合考虑上述因素,我们采用如下特征作为BP网的输入参数,如图4所示.

(1) 波谷的绝对高度,图4中用h表示.

(2) 波谷与相邻两波峰之间的平均高度差:

$$\Delta h=\frac{\Delta h_{left}+\Delta h_{right}}{2}. \tag{2}$$

(3) 波谷到左波峰的距离dis_{left}.

(4) 波谷到右波峰的距离dis_{right}.

3 BP算法用于文本切分

BP算法是一种多层前馈网络误差反向传播的学习算法,采用S型函数作为神经元的变换函数.只要有足够多的隐层和隐层节点数,BP网就可以逼近任何非线性映射.这里,我们选择4个输入节点,分别对应于上一节介绍的4种特征参数,隐含层选择10个节点,输出层取一个节点对应0为不切割,1为切割.将预先切割的语料作为训练数据,通过学习使网络收敛.将测试语料通过预处理,进行文本规整并滤除其中的虚词,然后用式(1)计算其相似函数,提取特征,最后送入识别网络,对可疑的主题转换点进行定位.

4 基于层次结构定位方法

上述算法能够成功地定位大部分主题转换点,满足像建立领域相关统计语言模型这一类精度要求不高的应用需求,但依然存在一些问题.

首先，依然存在较多的误判和漏判现象. 在我们的初步实验中，正确定位率只能达到 92%.

其次，定位精度不高，通常 BP 网络所定位的主题转换点与人工确定的主题转换点要有 3～4 个句子的偏差. 这种偏差对于一些精度要求较高的应用，如对话系统是不能忍受的.

上述问题产生的原因在于，计算相似函数时滑动步长的选取对定位性能有极大的影响. 图 5 所示为一段文本在不同滑动步长下的相似函数的波形. 由上至下滑动步长分别选为 1、2、4、8、16 个句子. 我们可以看到，当步长选择较小时，波形含有更多的细节信息，但受局部扰动的影响也更大，随着步长的逐步加大，波形更多地反映了整体包络的变化，而局部信息逐渐丢失. 前者会造成大量的误判，后者会对一些篇幅较小的话题形成漏判，而且步长较大也是造成精度下降的主要原因.

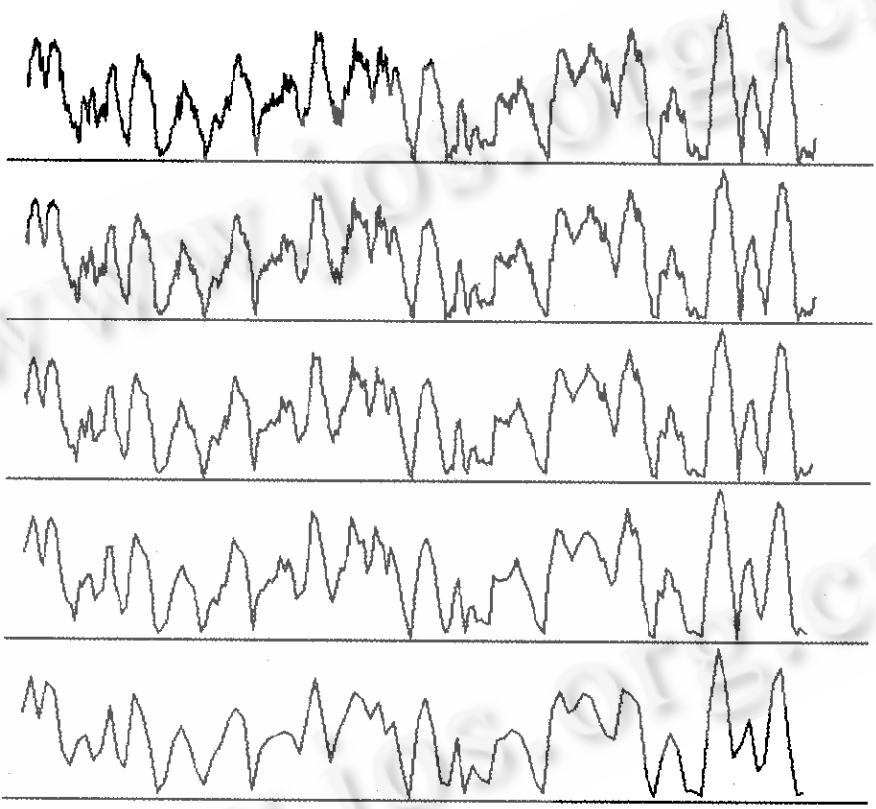


图 5

在单一步长的情况下，误判问题和精度不高是一对矛盾，不可能同时得到解决. 因此，我们提出利用不同步长下的不同分辨率建立层次结构的方法来解决上述问题. 建立层次结构的主要思想是，真实的主题转换点在不同分辨率下的判别中有连续性，而局部扰动不存在这种连续性.

层次结构的定位方法分以下几个步骤.

(1) 分别在不同的步长下，计算不同分辨率的相似函数，步长越小，分辨率越高. 然后对不同分辨率的相似函数分别提取特征，训练出不同分辨率下的 BP 网络. 在高分辨率下的 BP 网络能够成功地发现所有的词汇突变现象，这些突变现象有些对应着主题转换点，但也有很多对应着局部扰动. 低分辨率下的 BP 网受局部扰动的影响很小，但是漏判现象比较严重. 而且步长较大带来的另一个问题是定位精度较低.

(2) 对测试语料，利用不同分辨率下 BP 网的定位结果形成层次结构. 首先定义两个集合，即主题转换点的可靠候选集合 Cred 和不可靠候选集合 No_Cred，然后作如下处理：

(a) 从分辨率最低的定位结果开始，令 level＝0，将 BP 网定位的所有主题转换点归入 Cred 集，No_Cred 集

为空.

(b) 对当前层 Cred 集中所有的元素在下一层的定位点中寻找其最近邻.如果该元素与其最近邻之间的距离小于某一事先预定的阈值,则将该元素的最近邻规入下一层的 Cred 集合,否则将该元素归入下一层的 No_Cred 集合.对当前层 No_Cred 集合中的所有元素在下一层的定位点中寻找其最近邻,如果该元素与其最近邻之间的距离小于某一事先预定的阈值,则将其最近邻归入下一层的 No_Cred 集合中.否则将该元素从 No_Cred 集合中删除.将下一层中的定位点中所有不是上一层定位点的最近邻的元素归入下一层的 No_Cred 集合中,作为上一层可能漏判的转换定位点的候选.

(c) level+1,如果未达到最底层(即分辨率最高的 BP 网所形成的判决结果),转(b),否则,将当前 Cred 集和 No_Cred 集中的所有元素合并,作为最终的定位结果.其中 Cred 集中保留的元素为从最上层到最底层始终保持了连续性的判决点,No_Cred 集中的元素为被上层漏判但在下层判决中保持了连续性的元素.

层次结构的判决方法有以下几个优点:

首先,层次结构的判决结果都是在所有层次或至少在多个层次上保持了连续性的结果,绝大多数的误判在这种连续性判决中被滤除.

其次,层次结构的判决在 No_Cred 集中保留了在上层低分辨率下不能判决,但在下层高分辨率判决中具有连续性的结果,非常成功地解决了单一步长判决中难以解决的漏判问题.

第 3,层次结构判决的最终结果都是由最底层,即分辨率最高的定位点中选出的,其定位精度远远高于单一层次下所得的结果.用层次结构定位的精度通常在一个句子左右,可以满足精度要求很高的应用需求.

第 4,与单一步长的判决相比,层次结构的判决并不需要更多的训练数据.我们只需将相同的训练语料在不同的移动步长下计算相似函数,提取特征作为不同层次 BP 网的训练数据.这种方法只增加训练多个 BP 网所需的离线计算量,除此之外不增加任何负担.

5 实验结果和结论

我们的实验以《人民日报》作为训练和测试语料,题材涉及政治、文化、体育等多个方面.我们以人工切割的语料约 320K 字节作为训练数据,具体分布见表 1.

表1

题材	所占比例(%)
政治	28
经济	43
文化	24
体育	5

滑动窗口的宽度为 40 个汉语文本句子.附录为用基于 BP 网和层次结构的方法对一篇文章的主体转换点的定位结果,在转换点处用"*"标出.本文的方法把文章分为 3 段:第 1 段笼统地介绍了安塞妇女擅长剪纸艺术;第 2 段介绍了两个例子加以说明;第 3 段介绍这种现象的成因和文化意义.可以看出,这种切分基本符合人的主观判断.与此同时,我们也将单一的 BP 网的切分结果给出,在转换点处用"·"标出.可以看出,单一的 BP 网也能找出主题转换点,但在第 1 个转换点处偏移了一个句子,其精度不如多层判别.

由于在大规模的测试中,分割结果的准确性没有一个明确的客观标准,所以我们采用以下评判方法.重新组织测试语料,挑选长度较短的文章组成连续文本流,由于组成文本流的每篇文章篇幅不长,我们可以假设每篇文章内部不发生主题转换,因此文本主题转换点的寻找就有了一个客观的标准,即如果算法找出的分割点恰好是每篇文章的交接处,我们就认为分割正确,否则为错误.在组织测试语料时,为考虑测试难度,可令题材相近的文章相邻,这样做的主要目的是为了使不同主题文章之间的界限模糊,增加测试难度.

为了验证本文提出的方法,我们进行了较大规模的实验.我们按上述原则选择了 200 篇短文作为测试语料,实验结果见表 2.

表 2

方法	正确定位率(%)	平均定位精度
单一的 BP 网	92	4·3 句
BP 网+层次判决	100	0·8 句

在表 2 中，正确定位率为被正确切分的文章在总的测试语料中的比率，如果切分点落在离正确文章边界 8 句之内的地方，则被任务切分正确. 平均定位精度则是实验所得主题转换点与文章实际边界之间的平均距离. 从实验结果可以看出，BP 网与层次结构结合的方法很好地解决了局部扰动带来的误判和漏判问题，同时，定位精度也大大提高，可以满足精度需求较高的应用.

自然语言主题转换点的自动定位是自然语言处理的一个重要内容，具有广泛的实用意义. 以往的主题转换方法在精确性和实现复杂度上都存在着一些不足. 本文针对以往方法的不足，试图用神经网络来处理自然语言中不确定的现象，提出了基于 BP 算法的文本主题转换定位方法. 本文的另一个主要工作是发现在不同步长下文本的词汇变化曲线对词汇变化的局部细节的分辨率不同，而正确的主题转换点在不同分辨率下具有连续性. 利用此性质，本文建立了词汇变化曲线在不同分辨率下的层次结构，并将神经网络的判决和这种层次结构相结合，建立了一种定位精度很高且具有较强鲁棒性的文本主题转换定位算法. 实验证明，该算法能满足精度要求较高的应用需求.

参考文献

1 Smith R W, Hipp R D, Biermann A W. An architecture for voice dialog systems based on prolog-style theorem proving. Computational Linguistics, 1995,21(3):280～320

2 Berry M W, Dumais S T, O'Brien G W. Using linear algebra for intelligent information retrieval. Society for Industrial Applied Mathematics Review, 1995,37(4):573～595

3 Morris J, Hirsy G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 1991,17(1):21～48

4 Lin Sung-Chien, Tsai Chi-Lung. Chinese language model adaptation based on document classification and multiple domain-specific language models. In: Proceedings of European Conference on Speech Communication and Technology. Rhodes, Greece: European Speech Communication Association, 1997. 1463～1466

5 Marti A. Hearst texttiling: segmenting text into multi-paragraph subtopic passages. Computational Linguistics, 1997,23(1):33～64

附录.

当安塞县的一位农村妇女在法国图鲁斯市剪纸表演时
法国观众为她的精湛技艺所叹服
市长不相信是用一把普通的剪刀剪出来的
他拿起剪刀翻来覆去地看
怀疑是不是安装了电脑
安塞农村劳动妇女大都会剪窗花
全县会剪花的妇女达一万多人
安塞民谚生女子要巧的
石榴牡丹冒铰的
剪纸是安塞品评巧手妇女的标准
有的人一巧百巧
剪花能手也是绣花能手面花能手
姑娘们从小就跟着妈妈铰花绣花
一辈子几十年
有着中国民间艺术体系的深厚的造型功底色彩功底和独特的艺术风格
代代传承发展
安塞的剪纸是根据特定的历史地理条件创作出来的
传统的题材有牛马鸡羊喜鹊凤凰蔬菜花果等等
从生活的各个角度寄托了劳动人民热爱劳动向往幸福的情感
展示了一个红花绿叶鸟飞鱼翔人欢马嘶万物争荣欢乐如意的艺术世界
*
一些农村妇女是真正的艺术家
•
记得暖水泉村一位叫延喜芳的八十一岁的老人
临终前
把窑洞门窗全部换成自己亲手剪的新窗花
在炕壁周围贴上华丽的炕围花
然后将留下的一包窗花花样交给儿媳妇说我没有给你留下甚么东西
我一辈子就爱剪花
就把这些窗花样子留给你吧

我再也没有甚么遗念了
说完就在满窑剪花的花丛中安详谢世
安塞妇女既是剪花能手
也是劳动能手和勤俭持家能手
一九八六年中央美术学院到陕北山沟请几位老人到中央美院任教一个月
小脚老太进入最高美术学府
这可是开天辟地的新鲜事
她们热忱地手把手教
同学们以极大的兴趣日夜守在她们身旁
老人们的教学引爆同学深入生活的热浪
后来全国各地到安塞采风的人越来越多
有两位美院的同学进沟到胡凤莲老人家里看她剪花
老人疼爱年轻人
把家里温暖的住窑让出来给同学
自己搬到隔壁当仓库的寒窑里
夜里着了凉
从此一病不起离开人世
* *
安塞地处中华始祖黄帝氏族文化发祥地的陕北黄土高原
这里遍布仰韶文化和龙山文化的彩陶遗址
秦始皇汉武帝又在这里修长城开直道北御匈奴
出现过灿烂的秦汉文化高潮
明代以后由于天灾人祸
生态平衡遭到破坏
形成历史上的交通封闭和文化封闭
使其它地域早已失传的古老的民族文化传统得以在民间较完整地保存下来
成为历史文化的活化石
由于农业经济的社会分工
主要从事家务劳动和艺术活动的劳动妇女
一把剪刀一根针代代相传
把极为丰富的文化传统继承下来
一个窑洞就是一幢活的民族民俗和民间艺术博物馆
它的文化内涵是我们民族从原始社会至今长达六七千年的历史文化积淀
因此
安塞劳动妇女创造的民间美术的价值
远远超越了民间美术本身
具有极为丰富的哲学美学考古学历史学民族学社会学和人类文化学内涵
是民族文化的凝聚和结晶
安塞劳动妇女为民族文化的继承和发展作出了历史性的贡献

A Method to Position the Natural Language Topic Change Accurately Based on Neural Network and Hierarchies of Word Change

CHEN Lang-zhou HUANG Tai-yi

(*National Laboratory of Pattern Recognition Institute of Automation The Chinese Academy of Sciences Beijing* 100080)

Abstract The topic change of natural language is a very important clue to natural language understanding. Since different database and method should be used when different topic text is processed generally, it is important to find the topic change point in text. This technology is very useful in natural language understanding, text indexing and language model building, etc. In this paper, using the burst character of vocabulary in the change of topic, the authors present four parameters to reflect this character. They propose a method of text segmenting based on BP algorithm and hierarchical structure of word change. The accuracy of this method is about one sentence.

Key words Natural language processing, text segmenting, text index and filter, BP algorithm.