一种基于免疫遗传算法的网络新词识别方法

丁建立[1,2] 慈 祥[1,2] 黄剑雄[3]
(中国民航大学计算机科学与技术学院 天津 300300)[1] (中国民航信息技术科研基地 天津 300300)[2]
(中国国际航空股份有限公司信息管理部 北京 100071)[3]

摘 要 随着互联网的发展,网络新词不断涌现,但是目前的分词方法很难及时、准确地对其做出识别。对此提出一种应用免疫遗传算法的网络新词识别方法。在分析网络新词特点的基础上,利用汉语词群现象和词位的概念提取出示范抗体,在遗传算法进行的过程中有针对性地注入该抗体。实验表明,该方法对于分词碎片中符合词群现象的新词有着极高的识别率,对于一般网络新词的识别率也基本令人满意。

关键词 免疫遗传算法,汉语词群,词位,抗体,网络新词识别
中图法分类号 TP391.12 **文献标识码** A

Approach of Internet New Word Identification Based on Immune Genetic Algorithm

DING Jian-li[1,2] CI Xiang[1,2] HUANG Jian-xiong[3]
(College of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China)[1]
(Information Technology Research Base, Civil Aviation Administration of China, Tianjin 300300, China)[2]
(Information Management Department in Air China, Beijing 100071, China)[3]

Abstract The development of Internet leads the internet new word coming into being. These unknown words are difficult to identify timely and accurately by the current Word Segmentation Method, therefore internet new word identification method using Immune genetic algorithm was brought forward. This method is based on the analysis of characteristics of internet new word, using the phenomenon of Chinese words and word groups to extract exemplary antibody, and injecting the antibody targeted during the process of genetic algorithm. The experiment results show that the method not only has a higher recognition rates of the new words consistent with the phenomenon of word groups in word fragments but the result of identifying ordinary internet new word is adequate.

Keywords Immune genetic algorithm, Word group, Word position, Antibody, Internet new word identification

1 引言

近些年来,中文分词领域的研究不断取得新的进展,无论是分词的准确率还是分词的速度都有了不小的提升,但是对于新词的识别仍是难点之一。众多研究者从不同的角度应用不同的方法进行了研究,韩洁等人[1]对于互联网环境中的未登录词识别做出了初步性的尝试。郑家恒等人[2]根据汉语构词法建立规则库,通过调用"互斥性字串"过滤规则和构词规则来识别网络新词。Wu 等人[3]提出利用独立词概率来识别被切散为单字串的新词。邹纲等人[4]提出一种以某时间点为界建立背景和前景词串集合,采用评价函数检测 Internet 中新词的方法。崔世起等人[5]针对二元新词、三元新词、四元新词等常见模式,利用多个词典和词性过滤规则、独立词概率等技术检测新词。Li 等人[6]利用 SVM 识别 NW11(单字符+单字符)和 NW21(双字符词+单字符)型的新词。闫蓉等人[7]则尝试使用遗传算法识别未登录词。以上的方法对于新词的识别准确率和效率都不是很理想。网络新词的产生有多种方式。就目前网络新词的产生方式来看,主要有以下3种形式:(1)旧词新用,比如"下课";(2)由热门事件触发产生的新名词,例如"范跑跑"、"猪坚强"等;(3)经汉语词群现象由某个具有高度构词能力的共同语素构造的新名词,例如"艳照门"、"啃老族"等。在这3种产生方式中,第一种虽然语义发生了变化,但是形式并没有发生变化,其利用现有的分词方法基本可以准确识别。第二种新名词的识别较为困难,因为这些名词的产生具有极高的偶然性,并且很多情况下这些名词在现有的汉语构词方式下没有实际的意义。第三种新名词的产生方式是最常见的一种,通过对新名词网[8]上的新名词的统计(2010年1月6号数据),发现其中27%的新名词都是由某些词群产生的(剔除一些非常生僻的新词的话,这一比例会超过30%),属于同一族的词。

到稿日期:2010-02-05 返修日期:2010-06-01 本文受国家高技术研究发展计划(863)(2006AA12A106),国家自然科学基金(60879015,60572167)资助。

丁建立(1963—),男,博士,教授,主要研究方向为智能仿生算法、智能信息处理、信息安全,E-mail:jianliding@yahoo.com.cn;慈 祥(1986—),男,硕士生,主要研究方向为智能仿生算法、网络安全;黄剑雄(1970—),男,高级工程师,主要研究方向为航空公司信息系统、信息处理与信息安全。

2 相关概念

2.1 免疫遗传算法

遗传算法是一种借鉴了自然界生物进化和自然选择的智能算法，通过对父代个体进行选择、变异、交叉等操作产生中间个体，利用相关评价函数保留较优的 N 个解作为新的一代。这个过程反复迭代进行，就可能得到最优解。遗传算法的主要优点就是对解进行并行搜索和在全局范围内寻找最优解。但是遗传算法容易陷入局部最优解，同时存在着收敛速度较慢的缺点。

为了解决这些问题，相关研究学者开始尝试将免疫算法引入遗传算法，提出了免疫遗传算法。生物免疫系统具有免疫记忆、抗原识别和保持抗体的多样性等特性，利用问题的先验知识提取出免疫算子，通过注射抗体和免疫选择来产生新的个体。这种方法既保留了最佳个体，又避免了算法过早收敛而陷入局部最优，同时加快了算法求解的速度。

图 1 是一个典型的基于示范抗体注射的免疫遗传算法。从图中不难发现，算法的基本步骤与标准的遗传算法相比主要是多了示范抗体的提取和注射，而示范抗体的提取正是算法的核心和难点之一。

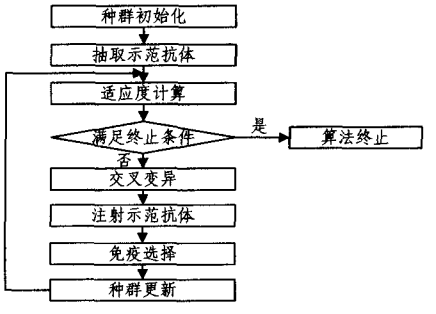

图 1 基于示范抗体的免疫遗传算法

2.2 汉语词群现象

汉语词群现象并不是新近产生的，但是随着网络的发展，词群现象有迅速扩大的趋势，现已成为网络新词产生的一个重要来源。所谓词群，是指具有某一共同特征的一群词的聚合。从心理学的角度来看，这种词群现象是人们趋同心理在语言学上的体现，其产生的根本原因是社会发展的需要。符合词群现象的汉语新词语由共同词素与变动词素构成，其构成模式为共同词素＋变动词素或变动词素＋共同词素。一般来说，由词群现象产生的词有以下几个特点[9]：

(1) 共同语素的位置相对比较固定，绝大多数的共同语素只能出现在词中的特定位置。常见的词群词构词模式有 4 种：1＋1，2＋1，1＋2 和 2＋2，其中 1＋1 和 2＋1 的构词模式占了绝大部分，从数量上来看超过 80%。

(2) 偏正结构的新词语词群数量较多，而其他结构的词群相对较少。其中形名结构的词群，即中心词是名词的词群所占比重较大。

(3) 新词中的变动语素不是单字(串)就是词，而且这些单字(串)往往是构词能力比较强的。也就是说，在大规模的语料库中这些字通常不是单独出现，而是作为某个词的一部分出现。

以上特点对于我们下一步具体的提取规则有很大的启发作用，同时掌握了这些特点对于我们在切分结果中判断是否正确地切分出了一个词群词有着很重要的意义。

2.3 词位的概念

由字构词的分词方法与以往的分词方法不同，它将分词视为字的词位分类问题。关于“由字构词”分词方法的最早一篇论文发表在 2002 年第一届 SIGHAN 研讨会上。目前比较成功的应用由字构词原理的分词系统是由微软亚洲研究院研制的，在其系统中将字分成了 6 类：S(单独成词)、B(词首)、B2(词中第二个字)、B3(词中第三个字)、M(词中)、E(词尾)。例如对以下的分词结果，可以将字标注为以下形式[10]：

/上海/计划/到/本/世纪/末/实现/人均/国内/生产/总值/五千美元/。/

上/B 海/E 计/B 划/E 到/S 本/S 世/B 纪/E 末/S 实/B 现/E 人/B 均/E 国/B 内/E 生/B 产/E 总/B 值/E 五/B 千/B2 美/B3 元/E。/S

在对大规模的语料库进行标注后，最终能够形成一个字库，规定每个字的词位标记中某一个词位超过 50%，那么就认为该词位是这个字的主词位，否则就认为这个字是自由字(Free)。在微软的语料库中总字量为 5147 个，有主词位的字为 3920 个，占据总字量的 76.16%。

由字构词的出发点在于字的位置，而我们的研究出发点也是从连续单字碎片中组合出网络新词，这二者的出发点是一致的。因此，利用好词位的概念将有助于提取出性能更好的示范抗体。

3 网络新词识别算法设计

3.1 基本设计理念

本文方法的处理对象是初分后的连续单字分词碎片。为了保证识别的效果，我们扩大了分词碎片范围，不再简单地将连续单字作为分词碎片，而是将共同语素碎片附近的二字词也作为碎片，一起放入我们的识别系统。在使用现今比较成熟的分词系统(本文采用的是中科院的 ICTCLAS)进行初步切分后，就可以利用本文的方法从切分后的分词碎片(扩大范围后的分词碎片)中提取出新词。受到文献[7]的启发，结合文献[1-6]的核心思想，我们尝试将免疫遗传算法引入网络新词的识别中。在对搜狗互联网语料库统计的基础上结合最新网页数据和新名词网，我们提取了具有代表性的词群 36 个，并归纳出这些词群的相应构词规则。紧接着以遗传算法为基础，将这些词群作为免疫算子引入遗传算法，结合由字构词法中的词位概念提取出示范抗体，对子代个体有针对性地注射示范抗体，这样既有效地避免了遗传算法的局部收敛，也保证了收敛速度。

在实际的分词过程中，还引入了记忆词库和待定新词库，主要目的是加快切分速度。其中初始的记忆词库是由我们收集到的现有符合词群现象的新名词构成，待定新词库是由那些系统虽然已经做出了切分但不是很确定的新名词构成。对于这部分待定词，我们要结合构词规则和最新互联网语料进行检测，一旦发现其符合构词规则且在最新的互联网网页中出现过，就加入记忆词库，这样在下次切分时直接切分即可。

3.2 词群提取与构词规则的表示

每一个符合词群规则的新词都由两个部分组成，即新词＝共同语素＋可变语素。其中词群的核心是共同语素。为了尽可能贴近真实的互联网环境并尽可能多地提取到共同语素，

我们提取共同语素的工作是以搜狗互联网语料库为基础的。考虑到该语料库不是最新的，从包括搜狐、新浪、天涯、猫扑在内的国内著名门户网站和论坛中下载了一批相对较新的网页，再结合新名词网最终提取出了具有代表性、构词能力较强的共同语素36个。这些共同语素以及由此产生的新词均符合上面提及的汉语词群基本特点。

表1列出了我们提取到的非常典型、具有代表性的共同语素。经过数据统计，发现在这些共同语素中，有的语素只出现在新词的左部，有的只出现在右部，少数语素左右都可以出现，语素的构词模式也不尽相同。具体的统计信息见表2和表3。

表1　典型共同语素

共同语素	典型例词
奴	房奴、卡奴
客	黑客、红客
门	艳照门、返航门
化	信息化、全球化
秀	模仿秀、脱口秀
控	拍照控、包包控
族	啃老族、sohu族
虫	书虫、网虫
领	白领、蓝领
模	车模、腿模
鸟	菜鸟、老鸟
民	彩民、股民
爷	倒爷、板爷
星	歌星、笑星
姐	凤姐、励志姐
哥	春哥、励志哥
女	剩女、败犬女
盲	文盲、电脑盲
男	剩男、凤凰男
嫂	军嫂、警嫂
帝	PS帝、技术帝
达人	恋爱达人、灌水达人
群体	强势群体、弱势群体
托儿	房托儿、医托儿
工程	扶贫工程、豆腐渣工程
产业	朝阳产业、低碳产业
打	打黑、打假
亚	亚健康、亚文化
晒	晒账单、晒工资
被	被就业、被离铁
山寨	山寨手机、山寨春晚
吧	吧姐、网吧
炒	炒房、爆炒
网	网恋、域域网
的	的哥、面的
热	热播、汉语热

从表2中很容易发现，出现在新词右部的共同语素占了绝大多数，“客”、“族”等是这类语素的典型代表。

表2　共同语素出现位置

共同语素出现位置	数量
新词的左部	5
新词的右部	26
新词的左右部	5
总计	36

注：此处左右部均可构词的共同语素看成同一个语素。

从表3可以看出1+1和2+1构词模式的共同语素所占比例为80.1%，这和上面提到的词群构词规律是相符的。

表3　共同语素构词模式

构词模式	数量
1+1	19
1+2	4
2+1	8
2+2	4
1+1/2+1	6
总计	41

注：由于某些左右部均可构词的共同语素在构词位置不同时的构词模式不同，因此此处统计将这些共同语素看成两个不同的语素，总的共同语素变为41个。1+1/2+1表示这两种构词法对于该共同语素均适用。

通过对收集到的词群词进行统计，发现这些词的具体构词方式基本都符合以下4条规则：

(1) 如果构词模式中的变动语素数量为1，且经过初步切分后对应变动语素位置的是一个字，计算该字的独立成词概率。如果此时的概率较低，则该字和共同语素很可能结合构成新词，否则不构成新词。典型的例子有“卡奴”。

(2) 如果构词模式中的变动语素数量为1，且经过初步切分后对应变动语素位置的是一个数字或英文单词，则该字很可能和共同语素结合构成新词。典型的例子有“sohu族”。

(3) 如果构词模式中的变动语素数量为2，且经过初步切分后对应变动语素位置的是一个词，那么此时这个词很可能和共同语素结合构成新词。典型的例子有“凤凰男”。

(4) 如果在构词模式中的变动语素数量为2，且经过初步切分后对应变动语素位置的不是一个词，计算变动语素位置的两个单字的共现概率。如果该概率较小，那么此时这两个单字很可能和共同语素结合构成新词。典型的例子有“飚爱族”。

设规则集 $U=\{u_1,u_2,u_3,u_4\}$，其中 u_1,u_2,u_3,u_4 分别对应上面的4个规则。

为了能在结果检测阶段方便、准确地判断切分出的词是否符合词群规则，我们将每个共同语素的构词方式表示成如下的形式：

$$[c_i,l,g_i,u_i,flag],c_i\in C,g_i\in G,u_i\in U \tag{1}$$

式中，C 是共同语素集，即 $C=\{c_1,c_2,\cdots,c_n\}$，c_i 代表某个具体的共同语素，l 表示该共同语素的位置在新词的左部(相应地，r 表示在新词的右部)，G 是构词模式集，具体来说有如下4种构词模式：

$$G=\begin{cases} g_1:1+1 & \text{模式} \\ g_2:2+1 & \text{模式} \\ g_3:1+2 & \text{模式} \\ g_4:2+2 & \text{模式} \end{cases} \tag{2}$$

式中，g_i 表示某个具体的构词模式。u_i 表示该词群词应满足的规则。需要注意的是，某些共同语素的构词规则可能不止一个，此时这几个规则都需列出。$flag$ 起一个标志位的作用，它的值只能取0或1。0表示该语素仅有一种构词模式，1表示该语素还有另外的构词模式，这主要是针对6个有1+1和2+1两种构词模式的共同语素。以共同语素“客”为例，利用我们定义的方式表示如下：

$$[客,r,g_1,u_1,0] \tag{3}$$

这表示共同语素是“客”，在新词的构成中它只出现在词的右边。新词的构词方式是1+1模式，即“客”字的左边仅有一个单字，并且仅有这一种构词模式。构词规则需满足规则(1)。

典型的代表有黑客、访客、飞客、博客、晒客等。

3.3 算法基本流程

根据以上设计思想，整个网络新词识别流程如图 2 所示。

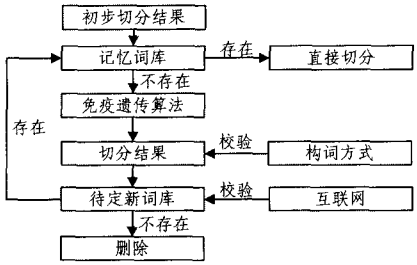

图 2 算法流程图

(1) 首先对待分的句子进行初步切分，得到进一步处理所需的分词碎片。

(2) 查询记忆词库。若在分词碎片中发现记忆词库中的词，则直接切分，否则转到(3)。

(3) 利用免疫遗传算法对分词碎片进行切分，并将切分过程中得到的符合词群现象的新词用构词方式$[c_i, l, g_i, u_i, flag]$进行验证。如果符合，就将其放入待定新词库，否则将得到的新词再次切分成碎片。对于不是由词群现象发现的新词，则直接放入待定新词库，输出切分结果。

(4) 定期地从互联网上下载最新的语料，利用待定新词库中的新词有针对性地在语料中寻找，一旦发现，则将该词加入记忆词库，否则直接删除。

4 算法实现中的几个关键性问题

4.1 编码

算法求解的第一步就是对问题进行某种形式的编码。为了便于示范抗体的提取和注射，我们采用了最常用的二进制编码。具体来说，对于分词碎片中可以组合的单字词编码为“0”，不可组合的单字词编码为“1”。算法求解结束后，解码也非常方便，连续的 0 组合在一起就是识别出的新词，1 则是单字词。

建立位串空间 $W^L=\{w_1, w_2, \cdots, w_K\}$，$w_k=(w_{k1}, w_{k2}, \cdots, w_{kL})$，$w_{kl}\in\{0,1\}$，$k=1,2,\cdots,K$，$l=1,2,\cdots,L$。$L$ 是位串的长度，$K=2^L$，其中 $w_k=(w_{k1}, w_{k2}, \cdots, w_{kL})$ 表示某一个个体。例如，“金融危机影响百姓生活，许多刚毕业就成为房奴的毕婚族也试着用拼婚等在拼客中流行的做法来节省开销”。利用 ICTCLAS 的初步切分结果是“金融/危机/影响/百姓/生活/，/许多/刚/毕业/就/成为/房/奴/的/毕/婚/族/也/试/着/用/拼/婚/等/在/拼/客/中/流行/的/做法/来/节省/开销/”，其中下划线部分是需要我们进一步处理的分词碎片，其编码为“1111111111111111”。

4.2 适应值计算

适应值是评价个体好坏的唯一标准，适应值高的个体将被保留。因此正确地选择适应度函数，将对算法有着决定性的影响。适应度函数的选择并没有单一的标准，关键是要能够准确地反映目标。和一般的优化问题不同，新词识别并没有一个很明确的目标函数，对此我们借鉴参考文献[7]的做法，以所有的基因位独立成词的概率之和来衡量适应度的大小。IWP 表示一个字单独成词的概率，计算方式为 $IWP=\frac{p(word(w_i))}{p(w_i)}$，其中 $p(word(w_i))$ 表示汉字 w_i 在语料库中作为一个单字词出现的次数，$p(w_i)$ 表示在语料库中汉字 w_i 总的出现次数。本文计算 IWP 所用的语料库是搜狗互联网语料库。将 IWP 嵌入到函数中，适应度函数的最终定义为：

$$f(w_i)=\sum_{i=1}^{L}[2^i/(IWP)^2] \tag{4}$$

4.3 交叉与变异

交叉和变异是免疫遗传算法中最核心的步骤之一。交叉是指两个父代个体相互交换部分结构得到新的子代的过程。根据交叉位置的不同，交叉一般可以分为 3 类：单点交叉、两点交叉和均匀交叉。

变异就是对群体中个体串的某些基因位上的基因值以一定的概率进行变动。本文采取的是二进制编码方式，因此本文中的变异实际上就是基因值为 0 的变为 1 或者基因值为 1 的变为 0。

交叉点和变异点的选取都是随机的。

4.4 示范抗体的选取和使用

示范抗体实质上是利用问题的一些先验知识来估计最佳个体的某些基因位上的值，也可以理解为对解做出了一个轮廓性描述，这个轮廓性的描述我们称之为示范抗体。通过对群体中的个体有选择性地注射示范抗体，会很大程度上避免群体出现早熟的现象，同时会极大提升群体的收敛速度。

虽然网络新词的形式看起来和传统的语言模式有所不同，但实质上来说这些新词还是符合语言学的构词规律的。对于初切后的分词碎片，利用词群现象和词位的概念可以在整个基因片段上选取出若干个可以基本确定的基因位，以此来形成一个对于最优解的轮廓性描述。本文示范抗体的提取利用的是共同语素和词位。其中共同语素采取的就是我们提炼出的 36 个共同语素。对于词位我们只使用两个词位，也就是 B(词首)和 E(词尾)。选取这两个词主要是因为这两个词对于判断词与词之间的边界有着很重要的意义，B2，B3 和 M 并不能很明显地界定不同词，而 S 在目前大部分的分词系统初分时都可正确识别。在对分词进行扫描过后，标注词位 B 和 E 出现的基因位。一旦发现分词碎片中包含这些共同语素，则对这些共同语素进行标注。完成以上标注工作，就可以提取出一个最终解的轮廓。这么做是因为一般情况下词首词、词尾词和共同语素的构词能力非常强，本身很少单独使用，出现在分词碎片中最有可能的原因是这些词组成了新词，而该词没有被正确识别，因此通过扫描分词碎片直接标示的方法来提取示范抗体是可行的。

图 3 是以上面的分词碎片为例提取的示范抗体。在对分词碎片扫描一遍以后我们在图中标示出了 5 个基因位。

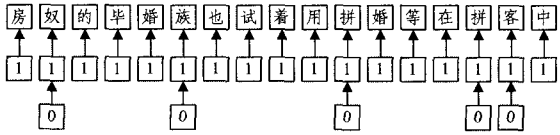

图 3 示范抗体提取

使用示范抗体需要注意，该抗体并不是群体中所有的个体都使用，而是有选择性地针对使用。通常来说我们在每一代的种群中选取若干适应度较低的个体，以一定的概率注射疫苗。选取个体的比例以及注射概率都可以自己设定，但不宜过高，因为过多的示范抗体可能会导致整个群体出现某种程度上的“同化现象”。

4.5 算法详细实现步骤

仍沿用上面的例子，详细阐述算法的核心步骤如下：

(1) 利用 ICTCLAS 对句子进行粗切分，得到分词碎片“房/奴/的/毕/婚/族/也/试/着/用/拼/婚/等/在/拼/客/中/”。分词碎片的初始编码为全 1 二进制序列，长度 $L=17$。设置迭代计数器 $t=0$，最大迭代次数 gen 设为 50，初始群体规模 $n=10$，交叉概率设为 0.6，变异概率为 0.01，这一步主要是完成初始化的工作。以上这些参数的选取并没有一个固定的要求，本文参数选择跟大多数文献中的参数设置方法一致。

(2) 对碎片的搜索过程中发现共同语素“奴”、“族”和“客”，并且发现主词位是句首的“拼”字。将这些字所在的基因位记录下来，组成示范抗体。

(3) 利用适应度函数计算当前种群所有个体的适应度，进行停机条件的判断。一旦满足条件，则算法结束，否则继续。

(4) 对当前种群以(1)中的概率进行交叉、变异操作。这一步是整个算法的核心。根据本文问题的特点，我们选择的是单点交叉的方式。例如在初始种群中有如下的两个个体：

ω_1:1111111111111111 ω_{10}:0000000000000000

在第一次迭代的过程中，随机产生的交叉位为第 9 和第 10 基因位之间，即个体 ω_1^0 基因的 10－17 位和个体 ω_{10}^0 基因的 10－17 位进行交叉互换，得到新的两个个体为：

ω_1':1111111100000000 ω_{10}':0000000011111111

为了保证较优个体不被破坏，变异概率一般选择得很小，但也可能发生。例如在第 8 次迭代中 ω_6 发生了变异操作，变异位为 11，ω_6 由 10001100011111111 变为 1000110001011 1111。

(5) 对当前种群中适应度靠后的 30%的个体以 0.5 的概率注射示范抗体。在第一次迭代之后，经计算，适应度 ω'_1 处于群体中靠后的 30%被选择进行抗体的注射，也即对 ω_1' 中的第 2,6,11,15 和 16 位注射抗体，将这些基因位统一变为 0，注射抗体后 ω_1' 变为 10111011100000000。

(6) 进行免疫选择，选取适应度最高的前 10 个个体更新种群，转至步骤(3)。

在上例中，经过 18 次迭代后，得到了最优的结果“00100011110011001”。将连续的 0 组合起来，即得到识别出的 4 个网络新词：“房奴”、“毕婚族”、“拼婚”、“拼客”。利用相应的构词模式进行验证，发现“房奴”、“毕婚族”和“拼客”均符合其共同语素的构词模式，“拼婚”是利用词位识别出的新词，因此将这 4 个词全部输出至待定新词库。最后我们从互联网上下载的最新语料可以发现这 4 个词均在网页中出现过，所以这 4 个词最终将全部存入记忆词库，待下次切分用。

5 实验及结果分析

算法的实际效果采用 863 测评的 3 个指标：正确率、召回率和 F 值来评价。其中：

$$\text{正确率}=\frac{\text{识别出的词语出现在标准结果中的词语数}}{\text{识别出的词语总数}}\times 100\% \quad (5)$$

$$\text{召回率}=\frac{\text{识别出的词语出现在标准结果中的词语数}}{\text{标准结果中的词语总数}}\times 100\% \quad (6)$$

$$F\text{值}=\frac{2\times\text{正确率}\times\text{召回率}}{\text{正确率}+\text{召回率}}\times 100\% \quad (7)$$

从人民网、搜狐和猫扑上我们一共选取了 300 个含有网络新词的句子来做实验，每个网站各选取 100 个，一共包含新词 576 个。

从表 4 中可以发现，利用改进后的算法不论是准确率还是召回率相比单纯地使用遗传算法都有了很大的提高，这主要是因为抗体的加入使得种群的进化方向更有针对性。

表 4 算法识别效果对比

网站类型	识别正确	识别错误	未识别	准确率	召回率	F 值
遗传算法	416	87	73	82.7%	72.2%	77.1%
免疫遗传算法	458	65	53	87.6%	79.5%	83.4%

表 5 的结果证明了我们的算法对于词群词的识别效果非常好，但是对于一般的网络新词相对较差，但是总的来讲还是比较令人满意的。

表 5 不同类型网络新词识别效果对比

新词种类	识别正确	识别错误	未识别	准确率	召回率	F 值
词群词	306	25	11	92.5%	89.5%	91.0%
非词群词	152	40	42	79.2%	65.0%	71.4%

从表 6 中可以发现，对于新词识别的各项指标，3 种不同类型的网站有着 3 种不同的表现。其中以生活娱乐类的识别效果最好，门户网站其次，政治主题类识别效果相对最差。出现这种情况我们分析主要是跟网站的定位有关。生活娱乐类的网站受众主要是年轻群体，用语比较灵活，且常利用语群现象编造新词。大部分的语群类新词都出于这类网站，因此识别率非常高。政治主题类的网站用词比较谨慎，出现的语群新词相对较少，取而代之的是政治新词有不小的比例，因此识别的效果相对差些。门户网站的定位居于二者之间，因此效果居中。

表 6 不同类型网站识别效果对比

网站类型	识别正确	识别错误	未识别	准确率	召回率	F 值
政治主题类	113	20	15	85.0%	76.4%	80.5%
门户网站类	152	22	17	87.4%	79.6%	83.3%
生活娱乐类	193	23	21	89.4%	81.4%	85.2%

注：政治主题类以人民网为代表，门户网站类以搜狐为代表，生活娱乐类以猫扑为代表。

结束语 本文提出的方法主要是用来从现有系统切分后的连续单字分词碎片中识别出网络新词，在已有的利用遗传算法识别未登录词的基础上，在算法中加入了示范抗体，不但提高了收敛速度，也增加了识别的准确率。从实验结果来看，整体的实验效果令人满意。在实际的应用中，本文方法更适合用来识别生活娱乐类的语料。下一步的工作主要是改进示范抗体的选取，提高其他种类语料的识别准确率。

参 考 文 献

[1] 韩洁，周勇，刘少辉，等. 基于 WWW 的未登录词识别研究[J]. 计算机科学，2002，29(12)：155-156

[2] 郑家恒，李文花. 基于构词法的网络新词自动识别初探[J]. 山西大学学报：自然科版，2002，25(2)：115-119

[3] Wu An-di，Jiang Zi-xin. Statistically-enhanced new word identification in a rule-based Chinese system[C]// The 2nd Chinese Language Processing Workshop. Hong Kong，2000

[4] 邹纲，刘洋，刘群，等. 面向 Internet 的中文新词语检测[J]. 中文

信息学报,2004,18(6):1-9
[5] 崔世起,刘群,孟遥,等.基于大规模语料库的新词检测[J].计算机研究与发展,2006,43(5):927-932
[6] Li Hong-qiao, Huang Chang-ning, Gao Jian-feng, et al. The use of SVM for Chinese new word identification[C]//Processing of 2004 International Joint Conference on Natural Language. China,2004:723-732
[7] 闫蓉,张蕾.基于遗传算法的汉语未登录词识别[J].计算机应用与软件,2008,25(7):88-90
[8] 新名词网[EB/OL]. http://www.xinmingci.com/,2010-01-06
[9] 刘吉艳.汉语新词语词群现象研究[D].上海:上海外国语大学,2008
[10] 黄昌宁,赵海.由字构词——中文分词新方法[C]//中文信息处理前沿进展——中国中文信息学会二十五周年学术会议.2006

(上接第 228 页)

$$(U\times V)/(R_1\prod R_2)=\{\{(x_1,y_1),(x_1,y_2),(x_2,y_3),(x_2,y_4),(x_1,y_3),(x_1,y_4),(x_2,y_1),(x_2,y_2)\},\{(x_3,y_1),(x_3,y_2),(x_3,y_3),(x_3,y_4)\}\}$$
$$=\{\{x_1,x_2\}\times\{y_1,y_2,y_3,y_4\},\{x_3\}\times\{y_1,y_2,y_3,y_4\}\}$$

所以 $R\subseteq R_1\prod R_2$,但是 $R\neq R_1\prod R_2$。

定理 10 R 是 $U\times V$ 上的等价关系,则 R 是积可分解,即 $R=R_1\prod R_2$ 充分必要条件是 $[(x,y)]_R=[x]_{R_1}\times[y]_{R_2}$,$\forall x\in U,\forall y\in V$。

证明:(1)根据定理 3(2),必要性显然成立。

(2)由定理 9,只要证明 $R_1\prod R_2\subseteq R$,从而充分性成立。以下是证明过程。

对任意 $((x,y),(x',y'))\in R_1\prod R_2$,有 $(x,x')\in R_1$,$(y,y')\in R_2$。所以 $x,x'\in[x]_{R_1}$,$y,y'\in[y]_{R_2}$,则 $(x,y),(x',y),(x,y'),(x',y')\in[x]_{R_1}\times[y]_{R_2}$。

由条件 $[(x,y)]_R=[x]_{R_1}\times[y]_{R_2}$,我们有

$(x,y),(x',y),(x,y'),(x',y')\in[(x,y)]_R$

即 $((x,y),(x',y'))\in R$,故 $R_1\prod R_2=R$。

定理 11 R 是 $U\times V$ 上的等价关系,则 R 不是积可分解的充分必要条件是存在 $x_1,x_2,x_{11},x_{22}\in U,y_1,y_2,y_{11},y_{22}\in V$,使得 $((x_1,y_1),(x_2,y_2))\notin R$,但 $((x_1,y_{11}),(x_2,y_{22}))\in R$,$((x_{11},y_1),(x_{22},y_2))\in R$。

证明:(1)充分性,只需证明 $R\subset R_1\prod R_2$ 即可。R_1,R_2 的定义及由条件可知 $(x_1,x_2)\in R_1$,$(y_1,y_2)\in R_2$,从而 $x_1,x_2\in[x_1]_{R_1}$,$y_1,y_2\in[y_1]_{R_2}$,故 $(x_2,y_2)\in[x_1]_{R_1}\times[y_1]_{R_2}$,但 $(x_2,y_2)\notin[(x_1,y_1)]_R$。由定理 10 知,$R$ 不是积可分解的。

(2) 由 R 不是积可分解的,则 $R\subset R_1\prod R_2$。故存在 $x_1,x_2\in U,y_1,y_2\in V$,使 $((x_1,y_1),(x_2,y_2))\in R_1\prod R_2$,但是 $((x_1,y_1),(x_2,y_2))\notin R$,由 R_1,R_2 的定义可知,存在 $x_{11},x_{22}\in U,y_{11},y_{22}\in V$,使得 $((x_1,y_{11}),(x_2,y_{22}))\in R$,$((x_{11},y_1),(x_{22},y_2))\in R$。因此结论成立。

推论 1 R 是 $U\times V$ 上的等价关系,则 R 是积可分解的充分必要条件是如果 $x_1,x_2,x_{11},x_{22}\in U,y_1,y_2,y_{11},y_{22}\in V$,使得 $((x_1,y_{11}),(x_2,y_{22}))\in R$,$((x_{11},y_1),(x_{22},y_2))\in R$,那么 $((x_1,y_1),(x_2,y_2))\in R$。

结束语 粗糙集的理论研究有两条途径:一条是从论域上的二元关系出发构造各种粗糙集模型,称之为构造性方法;另一条途径是把上、下近似作为基本概念,利用一个公理集来刻画上、下近似算子,揭示公理集性方法。本文对不同论域上的近似空间进行笛卡尔合成,得到有限个近似空间的积近似空间,研究了它的一些基本性质,特别对可分解子集的上、下近似进行了刻画,研究了它的近似精度和粗糙度与分解式的关系。进一步地考察了一个笛卡尔积上的近似空间的可分解性。后期的研究可以从以下几方面展开,继续完善本粗糙集模型及其扩展:研究笛卡尔积粗糙集模型的公理化刻画及具体算法实现,结合文献[13,14]等研究这种推广模型的应用。由于我们的讨论是基于两(有限)个互不相同的论域,因而使得构造性方法在实际中具有更广泛的应用领域。

参 考 文 献

[1] Pawlak Z. Rough Set[J]. International Journal of Computer and Information Science,1982(5):341-356
[2] 张文修,吴志伟,梁吉业,等.粗糙集理论与方法[M].北京:科学出版社,2001
[3] Yao Y Y. Two views of the theory of rough sets in finite universe[J]. Internat. J. Approx. Reason,1996,15:291-317
[4] Yao Y Y, Lingras P J. Interpretations of Belief Functions in the Theory of Rough Sets[J]. Information Sciences,1998,104:81-106
[5] Li T J. Rough approximation operators on two universes of discourse and their fuzzy extensions[J]. Fuzzy Sets and Systems,2008,159:3033-3050
[6] Wu W Z, Leung Y, Mi J S. On characterizations of (I,T)-fuzzy rough approximation operators[J]. Fuzzy Sets and Systems,2005,154:76-102
[7] 邱雅竹,付蓉.两个论域上的粗糙集模型及其应用[J].四川师范大学学报:自然科学版,2005,28(1):15-18
[8] 高明,王继成.双论域下粗糙集数据约简方法研究[J].计算机工程与应用,2009,45(2):144-146
[9] 刘贵龙.基于两个集合上粗集模型的算法实现[J].计算机科学,2006,33(3):181-184
[10] 余扬.双论域的粗糙集模型[J].科学技术与工程,2005,5(10):661-662
[11] 何薇薇.多论域粗糙集模型及其应用[J].科学技术与工程,2006,6(19):3013-3016
[12] 杨勇,李廉.双论域上粗糙集的矩阵定义[J].计算机工程与应用,2007,43(25):1-3
[13] 闫林,王全蕊,刘延.Rough 逻辑公式的语义分析及基于语义分析推理的研究[J].模式识别与人工智能,2006,19(4):433-438
[14] 陈卫东,张维明.笛卡尔积运算对数据库数据质量的传递影响[J].计算机科学,2008,35(6):210-213