ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.3, March 2010, pp.427−437
doi: 10.3724/SP.J.1001.2010.03596

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种宋词自动生成的遗传算法及其机器实现*

周昌乐[1,2+], 游 维[1], 丁晓君[3]

[1](厦门大学 智能科学与技术系,福建 厦门 361005)

[2](浙江大学 语言与认知研究中心,浙江 杭州 310028)

[3](厦门大学 英国语言文学系,福建 厦门 361005)

Genetic Algorithm and Its Implementation of Automatic Generation of Chinese SONGCI

ZHOU Chang-Le[1,2+], YOU Wei[1], DING Xiao-Jun[3]

[1](Department of Cogntive Science, Xiamen University, Xiamen 361005, China)

[2](Center for The Study of Language and Cognition, Zhejiang University, Hangzhou 310028, China)

[3](Department of English Language and Literature, Xiamen University, Xiamen 361005, China)

+ Corresponding author: E-mail: dozero@xmu.edu.cn

Zhou CL, You W, Ding XJ. Genetic algorithm and its implementation of automatic generation of Chinese SONGCI. *Journal of Software*, 2010,21(3):427−437. http://www.jos.org.cn/1000-9825/3596.htm

Abstract: Automatic generation of poetry has always been considered a hard nut in natural language generation. This paper reports some pioneering research on a possible generic algorithm and its automatic generation of SONGCI. In light of the characteristics of Chinese ancient poetry, this paper designed the level and oblique tones-based coding method, the syntactic and semantic weighted function of fitness, the elitism and roulette-combined selection operator, and the partially mapped crossover operator and the heuristic mutation operator. As shown by tests, the system constructed on the basis of the computing model designed in this paper is basically capable of generating Chinese SONGCI with some aesthetic merit. This work represents progress in the field of Chinese poetry automatic generation.

Key words: natural language generation; computational poetics; Chinese SONGCI generation; genetic algorithm

摘 要: 主要针对宋词这种特殊的汉语诗歌体裁,开展了有关自动生成算法及其实现方法的探索性研究.研究工作主要根据宋词特点,设计了基于平仄的编码方式、基于句法和语义加权值的适应度函数、基于精英主义和轮盘赌算法的选择策略,采用部分映射和启发式交叉算子和启发式变异算子,从而构建了一种基于遗传算法的宋词生成计算模型并进行了系统实现.实验结果表明,所建立的计算模型及其软件系统,初步实现了机器自动生成宋词的目标,对于给定的主题词和词牌,基本上能够自动生成有一定欣赏价值的宋词.论文的工作也填补了我国在汉语诗歌自动生成研究方面的不足.

关键词: 自然语言生成;计算诗学;宋词生成;遗传算法

中图法分类号: TP391 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant No.60975076 (国家自然科学基金)

Received 2008-12-09; Accepted 2009-02-24

汉语古典诗词的计算化研究始于 20 世纪 90 年代中期,迄今为止,已在语料库建立[1−4]、词汇语义分析[5−7]、创作风格辨析[8,9]、联语应对[9−11]等方面取得了一些初步的成果,但在诗歌的自动生成方面,除了一些民间的自发研究外[12],尚无系统性的学术性研究.相比而言,国外有关机器诗歌自动生成的研究起步较早,目前已尝试了许多方法并积累了一定的经验,从早期的 Word Salada 发展到现在较为成熟的基于进化算法和基于实例推理的方法,机器诗歌生成技术历经了多个阶段的发展,并开发了部分较成型的系统[13−15].

机器诗歌生成主要基于简单的计算程序,采用连接随机生成词汇的方法,生成结果仅是一些词汇的堆砌,形象地被称为 Word Salada.这种方法对诗歌内容、形式和意义的考虑都很少,其作品从严格意义上说并不能称为诗歌.接着便是基于模板的诗歌生成系统应运而生,通过事先定义好的模板来进行“填词”式创作.这类系统的代表有 RACTER 和 PROSE[16],RETURNER,APPI,BORANPO,Masterman 俳句生成系统以及互联网上的 ELUAR,ALFRED 等实用系统[17].基于模板的诗歌生成系统通常有较好的输出,RACTER 和 PROSE 的生成结果还曾被某诗刊杂志录用,但这类系统也存在一些固有的缺陷,比如缺乏灵活性、需要大量人为参与,生成作品的质量取决于模板的设计.

为了增加机器生成诗歌的灵活性,一些研究人员提出了基于模式的诗歌生成方法.与基于模板的方法一样,系统通过事先设定模式进行诗歌生成.不同的是,模式的灵活性远大于模板.一个典型的例子就是 Kurzweil 开发的 Cybernetic Poet 系统[18],其原理是以人类创作的诗歌为模式,从词汇、词汇结构及排列顺序、韵律模式、诗歌整体结构等方面,对大量的已有诗作进行了基于统计的分析和建模.另一个较为典型的系统是 Rubaud 等人.领导的 ALAMO 小组开发的 Rimbaudelaires 诗歌生成器[19],通过用空格替换 Rimbaud 十四行诗中的名词、动词和形容词来构造诗句模板,然后从 Baudelaire 的诗中选取相应的词进行填充;由于选词算法加入了句法和韵律方面约束,因此能够保证较好作品的产生.

另一种比较常见的诗歌生成方法是基于实例推理(case-based reasoning,简称 CBR)的诗歌生成方法[20−21].采用 CBR 技术的系统通常包括搜索(retrieve)、重用(reuse)、修正(revise)、保留(retain)4 个处理步骤,最有代表性的两个诗歌生成系统分别是 ASPERA 系统[22]和 COLIBRI 系统[23].由于 CBR 方法在知识获取、求解效率、求解质量以及知识积累等方面有着突出的优势,因此对于高质量诗歌的生成非常有利,但如何自动优化修改算法设计则是一个难以突破的瓶颈.于是,将遗传算法引入到了诗歌自动生成领域之中,就形成了诗歌生成较为先进的方法[24−27].

基于遗传算法的诗歌生成模型由生成模块和评价模块两部分组成.生成模块根据词法、句法、概念等信息产生备选诗作,评价模块则依据一定的准则对备选输出给予等级评价.采用遗传算法的实例主要有 Levy 开发的原型系统 POEVOLVE[24],能够生成 Limerick(一种起源于欧洲诗体,五行打油诗);以及 Hisar Maruli Manurung 的 MCGONAGALL 系统,在该系统中将诗歌生成问题看成一个状态空间搜索问题,并提出了语义(meaningfulness),语法(grammaticality)和诗性(poeticness)3 个诗歌必须满足的条件[27].其中,由于 MCGONAGALL 系统在语义表示上采用了词汇化树邻接文法,而在评估函数上采用了编辑距离算法和结构相似度两种度量,使其成为迄今为止最为成熟的一个基于遗传算法的诗歌生成系统.

借鉴上述有关遗传算法诗歌生成系统的主要原理,在我们自己建立的全宋词熟语料库(包括切分、词性、音韵、情感、典故、格律、词牌、句法等内容)的基础上[28−30],针对宋词自身的特点[31−36],按照遗传算法的构造原理[37],具体给出了一种宋词自动生成的遗传算法,并进行了机器实现.希望我们的研究,能够弥补我国在诗歌自动生成学术性研究方面的不足.

1 句法规范性和语义关联度的计算

采用遗传算法进行宋词的自动生成,首先遇到的问题就是要给出衡量宋词优劣与否的量化计算方法,作为适应度函数的构造与计算依据.我们知道,诗歌的质量主要反映在句法和语义两个层次上.一方面,诗歌作为自然语言的一种文学形式的表达,有着严格的句法要求.这里诗歌定义的句法,既包括通常汉语所需遵循的句法,又包括诗歌特有的格律规则,如平仄、押韵等规则.另一方面,诗歌的语义则包括了主题与词句的连贯、风格的

统一、情感与意境的传达等等.语义层次最关键的问题是如何使产生的诗句看起来更有意义,使句与句之间更有连贯性,而不是毫无关联的词汇或句子的堆砌.因此为了使机器能够产生好的宋词,首先要解决宋词句法规律与语义度量的计算问题.

就句法分析而言,宋词作为一种特殊的文体,其句法也有特定的要求.一般每个词牌的词体句法都有固定的总字数、总句数,每一句的字数也是固定的.根据我们全宋词数据库的统计,宋词字句的字数,从一字句到十一字句不等都有出现,但三字句至九字句占的数量最多,其句法分别是:

1) 三字句:上二下一、上一下二;
2) 四字句:上二下二;
3) 五字句:上二下三、上三下二、上一下四;
4) 六字句:上二下四、上四下二、上三下三;
5) 七字句:上三下四、上四下三、上一下六;
6) 八字句:上三下五、上四下四、上一下七、上二下六;
7) 九字句:上三下六、上四下五、上五下四、上六下三.

其中上句若是奇数字句,则首字往往是单字领;句字数多于 2 时,则可进行细分,给出进一步的层次句法分析.

通过对大量宋词语句构成的分析,我们发现,组成句子的有效模式的数目是有限的,并且呈现出了层次化的结构,因此比较适合采用 DFA(deterministic finite automata)或者 NFA(nondeterministic finite automata)来表示,具体策略如下:

(1) 随机组合的词语,在产生大量的备选个体后,逐个进行 DFA 分析测试,通过留下,没通过则剔除.

(2) 发挥 NFA 的优势,在句子产生的初期就运用 NFA,以不同的概率产生不同模式的句子.

显然,这两种策略既相互矛盾又相互补充,理想情况应该结合两种策略.但考虑到采用第 2 种策略需要较大的词库支持,比较耗时,因此本文采用第 1 种策略.图 1 给出的 DFA 判断树,描述的就是字数为 7,分词模式为“2212”的词句.其他字数的词句也可以类似给出对应的 DFA 判断树.

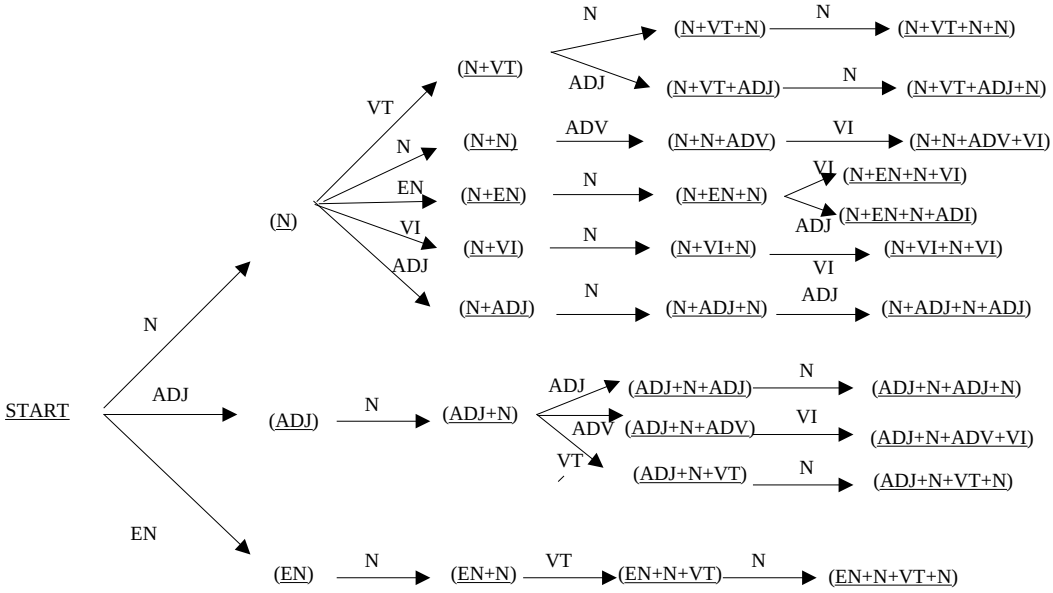

Fig.1 The DFA representation of the right syntax pattern of seven character sentence

图 1 七字词句合法句法模式的 DFA 表示

有了句法的形式规范,接着要解决的是宋词的语义计算问题,包括词义相关度计算、词义相似度计算,以及风格情感一致性计算 3 个方面.

首先,计算词义相关的目的是建立词语间的关联,发掘词语共现和搭配的可能,从而保证生成诗词行文和主题上的连贯.我们可以基于语料库统计来给出利用潜在语义分析和互信息两种方法词义相关度计算方法.

利用潜在语义分析(latent semantic analysis,简称 LSA)计算词义相关度的基本假设是:如果给予大规模的文本语料库,词义相关的词语由于有一定的共现规律,一个词可以用一些有共现规律的词来代表它们的语义.在《全宋词》语料库的基础上,我们可以构造频率矩阵,将所有的待测词(t 个)都用在待测文献(d 句)中的出现频率表示出来,形成 $X=t\times d$ 的矩阵,且其均可以被分解成 3 个矩阵的积,称为 X 的奇异值分解:

$$\underset{t\times d}{X}=\underset{t\times r}{T_0}\,\underset{r\times d}{S_0}\,\underset{r\times d}{D_0'}\,.$$

其中,$T'_0T_0=T_0T'_0=I,D'_0D_0=D_0D'_0=I,S_0=\mathrm{diag}[\sigma_1,\ldots,\sigma_t]$,是单值的对角矩阵,$r$ 是 X 的秩,$\sigma_1\geq\sigma_2\geq\ldots\geq\sigma_r>0$.于是,由于词 t_i 可以表示为 $t_i=(n_1,n_2,\ldots,n_d)$的向量,两个词的相关度就可以用 cosine 距离表示,而 cosine 距离又可以利用任意 $X(=TSD')$的两个相应行向量的点积来求得,由于 T,D 是正规矩阵,S 的对角元素大于 0,我们就有:

$$XX'=TS(TS)'=TS^2T'$$

其中,XX'的第(i,j)个元素是词 t_it_j 向量的点积.

第 2 种计算词义相关度的方法是基于互信息(mutual information,简称 MI)的方法[38].如果 s 为句子,w 为候选词语,$F_s(w)$是候选词 w 出现在句子 s 中的频率,$F_i(w)$是词语 w 在矩阵 i 列出现的频率,$F_s(j)$是 s 句子在 j 行出现的频率,而

$$N=\sum_i\sum_j F_i(j)$$

是矩阵所有项的统计,θ是为避免出现分母为 0 的情况而设的辅助数,暂设为 1;那么 $mi_{w,s}$ 就表示 w 和 s 之间的互信息,我们有:

$$mi_{w,s}=\frac{\dfrac{F_s(w)}{N}}{\dfrac{\sum_i F_i(w)}{N}\times\dfrac{\sum_j F_s(j)}{N}+\theta}.$$

为防止互信息在遇到词稀疏时的偏差,还可以引入纠偏因子[39]:

$$\frac{F_s(w)}{F_s(w)+1}\times\frac{\min\left(\sum_i F_i(w),\sum_j F_s(j)\right)}{\min\left(\sum_i F_i(w),\sum_j F_s(j)\right)+1},$$

然后计算 cosine 词义相关度:

$$cos\_sim(w_i,w_j)=\frac{\sum_s mi_{w_i,s}\times mi_{w_j,s}}{\sqrt{\sum_s mi_{w_i,s}{}^2\times\sum_c mi_{w_j,s}{}^2}+\varepsilon}$$

得到词语间语义关系的度量,其中,ε是为避免出现分母为 0 而设的辅助数.

对 MI 与 LSA 两种方法的计算结果进行对比,我们发现两种方法计算出的相关词有相当多的重叠但又有不同.因此,对于最终的计算结果,我们首先选取两种算法的重叠部分,相关度则用两者各占 50%的加权和表示;其次对于不重叠的部分,我们按相关度从高到低进行排列,并保留相关度大于 10^{-3} 的词.

现在来看词义相似度的计算问题.词语相似度主要用于衡量文本中词语的可替换程度.计算词义相似度,目的是在保证所选词紧扣主题的前提下,尽量使生成诗词的语言更丰富多变.目前自然语言的词义相似度有两类常见的计算方法,一种是利用大规模的语料库进行统计,另一种是根据本体知识来计算[39−41].

大规模语料统计方法利用词语的相关性来计算词语的相似度,基于的假设是:凡是语义相近的词,它们的上下文也应该相似.具体计算的策略是,事先选择一组特征词,然后计算这一组特征词与每一个词的相关性(一般用这组词在实际的大规模语料中在该词的上下文中出现的频率来度量),于是,对于每一个词都可以得到一个相

关性的特征词向量,然后利用这些向量之间的相似度(一般用向量的夹角余弦来计算)作为这两个词的相似度.

设在给定的语料库Ω和词表δ中,特定词语 x 在Ω上的语义 S_x 定义为如下五元组:

$$S_x = \{L_x, R_x, C_x, \delta, \Omega\},$$

其中,L_x 为 x 的左同现词汇特征向量,R_x 为 x 的右同现词汇特征向量,C_x 为对仗词汇特征向量.特征向量的元素为特征词与特征值组成的二元组(y,V_{xy}),有:

$$V_{xy} = \frac{\log f(xy)}{\log f(x)\log f(y)}.$$

其中,$f(xy)$为 y 在对应的 x 的相对位置上出现的频度(同一句的左边、右边或对仗位置上);$x,y\in\delta$,$f(x)$,$f(y)$分别是 x,y 在语料库Ω中出现的频度.那么,两个词语之间的语义相似度 Sim(x,y)可以通过计算其在 3 个不同的词汇特征空间(L_x,R_x,C_x)中的距离来得到(距离越小,相似度越大):

$$Sim(x,y) = \frac{1}{k_1\cdot\Delta L_{xy} + k_2\cdot\Delta R_{xy} + k_3\cdot\Delta C_{xy}},$$

其中 k_1,k_2,k_3 是可以根据语料库实际情况进行调整的加权参数,Δ为欧氏向量距离算子,其计算公式为

$$\Delta(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}.$$

在给定的语料库中,当 $Sim(x,y)$超过特定的阈值 R 时,就定义这两个词语 x,y 在该语料库中具有相似关系,x 的所有相似词组成的集合为 x 的相似词集 L_x.考虑到计算的复杂性和词义相似度在应用中较强的针对性,在实际计算相似度时,我们仅对词库中高频名词 545 个和形容词 367 个近义词集进行计算.

最后我们来讨论宋词的风格与情感一致性计算问题.应该看到,这是一个十分复杂的问题,目前还不可能作比较深入的研究.因此,为了尽可能简化这个问题,我们主要是在全宋词风格与情感标注的基础上,对词语做简单风格与情感分类统计来作为计算依据.比如,对词语进行风格上的量化,主要通过将词语集分为柔和、中性、强烈 3 个子集,然后递归地对各个子集进行相应的操作,最后将词语集分为 7 个不同意味的子集,用数字分别表示为–3、–2、–1、0、+1、+2、+3 这 7 种水平,分数越高,代表该词语能够体现某种风格的贡献度越强.类似地,对于词语情感意义的量化也参照风格量化的标准,从悲哀到快乐分为(–3,3)的 7 个等级.

总之,通过上述各个方面的形式分析和量化计算,我们可以为遗传算法自动生成宋词提供基本的句法和语义形式与量化计算手段,为利用遗传算法来进行宋词的自动生成铺平技术上的道路.

2 宋词生成遗传算法的构造原理

通过分析宋词词语的构成规律,可以发现每一首宋词都是词语库中某些词语的一种排列组合形式.从这个角度出发,我们可以认为,诗词生成问题在本质上是一个解空间中寻求最优化的问题,而解决这类问题正是遗传算法的优势所在.于是运用遗传算法来自动生成宋词,就是把宋词的自动生成看作一个状态空间搜索问题,这样就可以将遗传算法的优化机制引入到宋词的自动生成模型中.一般遗传算法包括求解问题编码、初始种群生成、适应度函数设计、遗传操作确定等 4 个方面的内容.下面给出宋词自动生成对应遗传算法的构造原理.

(1) 问题编码方案:对于宋词生成的编码问题是一个难点,为了避免烦琐,考虑到宋词的特点,我们提出了将“平、仄”与“0、1”编码相对应的编码方案.比如词牌《清平乐》平仄分布如下:

⊙平⊙仄,⊙仄平平仄.⊙仄⊙平平仄仄,⊙仄⊙平⊙仄.

⊙平⊙仄平平,⊙平⊙仄平平.⊙仄⊙平⊙仄,⊙平⊙仄平平.

其中⊙表示可平可仄.根据我们的编码方案可得如下编码串:

*0*1,*1001.*1*0011,*1*0*1.

*0 *100,*0*100.*1*0*1,*0*100.

其中,通配符*表示⊙的编码.在实际运用中,为缩小问题的解空间,我们以概率最大分词模式作为首选分词依据,即上述编码的宋词分词为如下词串:

*0 / *1,*1 / 0 / 01.*1 / *0 / 0 / 11,*1 / *0 / *1.

*0 / *1 / 00,*0 / *1 / 00.*1 / *0 / *1,*0 / *1 / 00.

相应地,词语库中的单字词和双字词也用 0、1 来编码,因此单字词分为平、仄两类,对应编码 0、1;双字词分为平平、平仄、仄平、仄仄 4 类,对应编码 00、01、10、11.

(2) 初始种群的生成:考虑到宋词严格的格律要求,在求解该优化问题过程中,我们始终将格律要求作为必须满足的约束条件,这样,种群初始化操作包括:

1) 随机生成满足词牌要求的韵部.如《清平乐》的上阙要求压仄韵,下阙转平韵,则随机生成一个平声韵部和一个仄声韵部.

2) 根据给定的主题词,从词语库中挑选与主题词相关度大于 $k1$ 的词语,构成一级候选词语空间.再从一级候选词语中挑选相关度高的一部分词语,再分别查找与这些词语相关度高的词语,构成二级候选词语空间……直到递归形成候选词语空间中的词语数量大于 $n1$.

3) 从候选词语空间随机选择满足押韵要求的词语,首先填充每个需要押韵的位置,然后在满足平仄要求的基础上,随机选词填充剩余的位置.重复上述操作,生成含 N 个个体的初始种群.

(3) 适应度函数:是问题约束条件反映,因此其设计不但与遗传算法中选择操作直接相关,而且还直接影响遗传算法的迭代终止条件.针对宋词生成问题,我们对个体适应性的评判主要依据以下 4 个指标:

1) 句法合法性 G:通过 DFA 检验的得分为 1,否则为 0.

2) 主题相关性 R:为所有语词与主题词的相关度之和.

3) 词句搭配的适当性 P:为所有两个连续语词的相关度之和.

4) 风格和情感统一性 S:追求高的风格和情感统一性,就是要求同一首宋词中出现词汇的风格和情感得分都趋于一致.因此,S 等于所有词语情感得分的方差与风格得分的方差之和的倒数.

适应度函数 F 定义为以上 4 个量归一化的加权和,即

$$F=\lambda 1G+\lambda 2\ R+\lambda 3P+\lambda 4\ S.$$

其中,G,R,P 与 S 均已归一化,$\lambda 1,\lambda 2,\lambda 3,\lambda 4$ 为相应的加权系数.

(4) 选择操作:选择操作就是从群体中按个体的适应度函数值选择出较适应环境的个体.考虑到宋词作品的优化是一个主观性较强的问题,目前尚无固定、量化的标准可以借鉴,我们采用精英主义和轮盘赌算法相结合的模型作为选择个体的依据.精英主义方法在每一次产生新的一代时,首先把当前最优解原封不动地复制到新的一代中,其他选择步骤不变.这样任何时刻产生的一个最优解都可以存活到遗传算法结束.在保留了当前最优解后,采用轮盘赌算法完成对剩余个体的选择,即按照个体适应度值所占全部个体适应值总和的比例作为被选概率来选择个体.

(5) 交叉算子:交叉操作是遗传算法中最重要的操作,是决定算法收敛性能的关键,因此必须慎重选择交叉算子的策略.通过对宋词编码特点的分析与实验,我们采用包括可以跨句进行的部分映射和启发式两种交叉操作.这里,部分映射交叉可看作二进制串的两点或多点交叉在换位表达中的扩展,用特别的修复程序来解决简单的两点或多点交叉引起的非法性.可以跨句部分映射交叉方法虽然操作简单快速,且由于交叉点可以在整首宋词范围内随机选取,因而产生的子代与父代有较大的相异性,能够有效避免种群单一化的过早出现;但是由于可能破坏句子的句法有效性的问题,因此必须通过启发式交叉策略加以补救.所谓启发式交叉方法,就是将父代以句为单位切分,比较每个句子的适应度,选择在适应度低的句子之间进行交换操作,而保留适应度高的句子.不同于部分映射交叉,在启发式交叉中,我们只允许等位基因的交换.显然采用启发式交叉可以保留个体内部适应度高的基因片段,而对适应度低的基因片段加以改变,这样无疑增强了子代的进化能力.

(6) 变异算子:变异操作是按一定概率,对个体编码串上的某个或某些基因位的值进行改变.针对宋词生成,我们采用启发式变异操作.步骤如下:

步骤 1:对于要进行变异的个体,比较每句的适应度,选出适应度值最小的句子.

步骤 2:若所选句不符合句法规范,找出与原句句法组合最接近的一种合法组合,利用词义相关,替换原句某

个或某些基因位.

步骤 3:否则,随机选取句中一个基因位 W_n,获取其邻位基因 W_{n-1} 的词性 P,查找与 W_n 相关度最大且词性为 P 的词,替换 W_{n-1}(若 n=1,改对 W_{n+1} 进行操作).

到此为止,我们完成了宋词自动生成的遗传算法的全部核心部分的设计.于是我们就可以进一步给出完整的具体算法并进行系统实现.

3 宋词生成的系统实现与实验结果分析

当我们完成了宋词自动生成遗传算法中主要原理各部分的设计构造,还要完整给出对应的实现算法,只需在此基础上具体给出宋词自动生成的遗传算法流程及主要参数确定即可.算法主要由初始种群生成、适应度计算、选择、交叉、变异 5 个主要步骤组成.算法主要流程如下:

算法. 生成初始种群,大小为 k_1.

置代数 gen=0,若 $gen<k_2$ 或进化停止,则反复执行以下操作:

计算种群中各个个体的适应度;

将适应度最大的个体复制到子代;

置 n=0,循环次数 $n<N/2$,则反复执行以下操作:

进行选择操作,选出两个父代个体;

产生一个随机概率 p

若 $p<k3$,则执行交叉操作,产生子代;

否则,保持,将父代复制到子代;

对新产生的子代执行下列操作:

计算子代的适应度;

置 m=0,若 $m<k4$,反复执行以下操作:

对子代执行变异操作,

若新的适应度比原来的小,将适应度置为新的适应度,更新子代;

否则,以概率 $k5$ 接受;

m++;

计算当前子代的适应度;

若适应度大于相应的父代,则将子代代替父代;

否则,以概率 $k6$ 接受父代;

n++;

gen++;

算法结束.

分析上述算法,我们不难看出算法终止的两个条件:(1) 完成了预先给定的最大进化代数;(2) 种群中的最优个体在连续若干代没有改进或平均适应度在连续若干代基本没有改进.算法中 k_1~k_6 为 6 个可调参数:k_1 为种群大小,一般取值在 30~200 之间;k_2 为设定的最大进化代数,取经验值 5 000;k_3 为交叉概率,一般取值在 0.3~0.9 之间;k_4 为变异操作次数,取值 3 000;k_5 为变异概率,一般取值范围在 0.001~0.2 之间;k_6 为父代接受概率,取值 0.3.

有了具体的宋词生成算法,就可以构建宋词自动生成系统,按用户输入的关键词(要求输入 1~3 个关键词)和词牌名自动生成宋词.实际系统共分数据库建立、句法语义处理、基于遗传算法的生成 3 个基本模块.实际系统是在普通微机的 Windows 平台上采用 VisualC++ 6.0 开发实现的,测试机器基本参数为:CPU 1.83GHz,内存 512 MB.目前系统仅支持 10 个常见词牌的宋词生成,这 10 个词牌分别是《蝶恋花》、《青玉案》、《清平乐》、《浣溪纱》、《西江月》、《点绛唇》、《鹧鸪天》、《江城子》、《长相思》、《浪淘沙》.

例如,取种群大小 $k1$ 为 100,最大进化代数 $k2$ 为 5 000,交叉概率 $k3$ 为 0.8,变异操作次数 $k4$ 为 3 000,变异概

率 $k5$ 为 0.15,父代接受概率 $k6$ 为 0.3.当输入主题关键词为“菊”,词牌名为《清平乐》时,系统经过如下运行过程.

首先系统提取主题关键词“菊”,在词义相似和词义相关库中进行查找,形成表 1 所示的计算结果.接着,系统根据《清平乐》词牌的要求随机生成两个韵部.上阙仄韵“小”,下阙转平韵“魂”,即随机生成了一个平声韵部和一个仄声韵部.规定每个个体中至少出现一个与主题词的词义相似词.生成的初始种群个体举例如下(之一):

登临多少,入夜催秋草.憔悴田园添缠绕,携手光阴欢笑.

金菊零落离魂,春风相近黄昏.为我悲秋斜倚,此生天气重门.

Table 1 Computational results of synonyms and correlated words of “JU” (chrysanthemum)

表 1 “菊”的词义相似和词义相关计算结果

Synonyms of “JU”	黄菊 紫菊 嫩菊 槛菊 兰菊 菊花 金菊 菊蕊 野菊 松菊 晚菊 庭菊 细菊 篱菊 赏菊 丛菊 新菊 菊香 白菊	
Correlated words of “JU” (excerpt)	Firstly correlation	轻寒 登高 秋色 重阳 晓寒 离恨 雁 黄 管弦 香 秋 晚秋 微雨 萧疏 零乱 凄然 黯淡 凄楚 憔悴 萦绊 愁颜 梦 影 夜 西风 零落 幽怨 微凉 斜日 馨香 鸿雁 金 祝寿 紫 中秋 新酿 东篱 高歌 醉 残 良辰 庭院
	Secondly correlation	情 舞 携手 竟 金尊 忆 轻轻 朱阑 残 难忘 红烛 朦胧 寒 烛影 无端 明镜 雁 梧桐 燕 吹 扁舟 故国 潇湘 残荷 露 叠翠 晨星 浩渺 清泪 回首 遥看 人间 笙歌 共舞 冷艳 长亭 相逢 双桨 红颜 暮云 吟 幽愫

最后,经过选择、交叉、变异等操作,系统最后生成的结果为:

相逢缥缈,窗外又拂晓.长忆清弦弄浅笑,只恨人间花少.

黄菊不待清尊,相思飘落无痕.风雨重阳又过,登高多少黄昏.

分析生成结果,可以看出:

(1) 在音韵方面,该词满足《清平乐》的平仄、押韵等要求;

(2) 在句法方面,没有出现明显的句法错误,除了个别句子,如“黄菊不待清尊”稍嫌别扭外,其他句子都较为通顺;

(3) 在语义表达方面,充分体现词义相关计算的优势,全词具有较好的主题一致性和叙述连贯性;

(4) 在风格和情感表达方面,一些关键语词既具有比较统一的婉约派风格,又表达了一种较为统一的悲伤情感.

总体而言,这首词是一首较为成功的机器诗作.主要的问题是句子间逻辑关系安排有所不当.换言之,句间逻辑关系是我们需要进一步考虑的问题.

表 2 进一步给出了一些典型的系统生成的示例.考虑到目前对于机器艺术作品质量的评测主要通过图灵测验性方式进行,因此我们也采用评判专家组来进行宋词生成结果的评测.表 3 则给出了对 50 次生成实验进行测评的结果分析,其中用户对所生成宋词的满意率是衡量该系统性能的最重要指标,尽管测试具有较大的主观性,我们还是针对主题相关度评判、风格情感一致性评判和总体质量评判这 3 个指标进行评测.评判专家组由 5 名中文系本科生组成,评判采用 5 分制.

Table 2 Three typical examples of the automatic generation of SONGCI poetry system

表 2 宋词自动生成系统 3 个典型示例

Input		Output	Style
Key word	Ci Pai		
菊	清平乐	相逢缥缈,窗外又拂晓.长忆清弦弄浅笑,只恨人间花少. 黄菊不待清尊,相思飘落无痕.风雨重阳又过,登高多少黄昏.	风格婉约
饮酒	西江月	饮酒开怀酣畅,洞箫笑语尊前.欲看尽岁岁年年,悠然轻云一片. 赏美景开新酿,人生堪笑欢颜.故人何处向天边,醉里时光渐渐.	风格豪放
佳人	点绛唇	人静风清,兰心蕙性盼如许.夜寒疏雨,临水闻娇语. 佳人多情,千里独回首.别离后,泪痕衣袖,惜梦回依旧.	风格婉约

Table 3 Systemic testing results

表 3 系统性能测评结果

Duration on the average	Scale of satisfaction (0~5)		
23m 37s	Theme relevance	Consistency in style and tone	Overall satisfaction
	4.13	3.42	3.86

实验结果表明,系统基本实现了自动生成宋词目的,生成作品的质量大部分是可接受的,偶尔也有较为出色的诗作生成(如实例分析中的《清平乐》).但在系统的运行效率和风格情感计算方面还有待改进.

4 结束语

在对机器自动生成诗歌的现有方法进行总结和分析的基础上,本文提出了基于遗传算法的宋词生成模型并进行了系统实现.我们主要是借鉴汉语古诗词计算语言学研究在词汇语义分析方面已取得的成果,建立宋词切分和音韵标注语料库.然后在此基础上,针对古诗词与现代汉语的区别,提出了基于 DFA 的句法判定规范,以及基于词义相似度、词义相关度、词汇风格和情感特征的语义度量.最后,我们根据宋词特点,设计了包括基因编码,适应度计算,选择、交叉、变异等遗传操作在内的具体遗传算法,并构建系统加以实现.实验结果说明了遗传算法模型的有效性和较好的通用性,基本上达到了我们的研究目标,填补了我国在诗歌自动生成方面研究的不足.但我们的研究工作也存在着一些问题,比如在诗词生成的研究中,缺乏系统的自学习能力,缺乏生成宋词的自动评价体系等,这些都是将来需要进一步研究的问题.

References:

[1] Liu YB, Yu SW, Sun QS. The implementation of a computer-aided environment for ancient poetry researches. Journal of Chinese Information Processing, 1996,11(1):27−35 (in Chinese with English abstract).

[2] Sui ZF, Yu SW, Lo FJ. The research on automatic pinyin-tagging for the famous Song poems and its implementation. Journal of Chinese Information Processing, 1998,12(2):44−53 (in Chinese with English abstract).

[3] Lo FJ, Lee YP, Tsao WC. The format auto-checking and database indexing teaching system of Chinese poetry and lyrics. Journal of Chinese Information Processing, 1999,13(1):35−42 (in Chinese with English abstract).

[4] Lo FJ. The design and application of the system for poetic language segmentation and semantic classification tagging. In: Proc. of the 4th Symp. on digital reservation. 2005 (in Chinese with English abstract).

[5] Yu SW, Hu JF. Word-Based statistical analysis of Chinese ancient poetry. Language and Linguistics, 2000,4(3):631−647 (in Chinese with English abstract).

[6] Hu JF. The lexicon meaning analysis-based computer aided research work of Chinese ancient poems [Ph.D. Thesis]. Beijing: Peking University, 2001 (in Chinese with English abstract).

[7] Li LY. A study on term connection oriented NLP technique and its applications [Ph.D. Thesis]. Chongqing: Chongqing University, 2004 (in Chinese with English abstract).

[8] Li LY, He ZS, Yi Y. Poetry stylistic analysis technique based on term connections. Journal of Chinese Information Processing, 2005,19(6):98−104 (in Chinese with English abstract).

[9] Yi Y. A study on style identification and Chinese couplet responses oriented computer aided poetry composing [Ph.D. Thesis]. Chongqing: Chongqing University, 2005 (in Chinese with English abstract).

[10] Fei Y. Research on multi-level integration of chinese semantics and system design of spring festival couplets [Ph.D. Thesis]. Beijing: Institute of Automation Chinese Academy of Science, 1999 (in Chinese with English abstract).

[11] Zhou M. Microsoft's generation system of Chinese couplets. Microsoft Research Asia natural language processing group. Beijing. 2006(in Chinese with English abstract). http://duilian.msra.cn/

[12] Zhou CL. An Introduction to Computation of Mind and Brain. Beijing: Tsinghua University Press, 2003 (in Chinese).

[13] Bailey RW. Computer-Assisted poetry: The writing machine is for everybody. In: Mitchell JL, ed. Computers in the Humanities. Edinburgh: Edinburgh University Press, 1974. 283−295.

[14] Van Mechelen MV. Computer poetry. 1992. http://www.trinp.org/Poet/ComP/ComPoe.HTM

[15] Gervás P. Exploring quantitative evaluations of the creativity of automatic poets. In: Proc. of the 2nd Workshop on Creative Systems, Approaches to Creativity in Artificial Intelligence and Cognitive Science, the 15th European Conf. on Artificial Intelligence (ECAI 2002). 2002. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.6.3026&rep=rep1&type=pdf

[16] Hartman CO. Virtual Muse: Experiments in Computer Poetry. Middletown: Wesleyan University Press, 1996.

[17] Boden MA. The Creative Mind: Myths and Mechanisms. London: Weidenfeld and Nicolson, 1990.

[18] Kurzweil R. Kurzweil's cybernetic poet. 2001. http://www.kurzweilcyberart.com/poetry

[19] Rubaud J, Lussonnal P, Braffort P. ALAMO: Atelier de Littérature Assistée par la Mathématique et les Ordinateurs. 2000. http://alamo.mshparisnord.org/rialt/pagaccalam.html

[20] Luger GF, Tubblefield WA. Artificial Intelligence: Structures and Strategies for Complex Problem Solving. 3rd ed., Reading: Addison Wesley Longman, Inc., 1998.

[21] Aamodt A, Plaza E. Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications, 1994,7(1):39−59.

[22] Gerv´as P. An expert system for the composition of formal spanish poetry. Journal of Knowledge-Based Systems, 2001,14(3-4): 181−188.

[23] Diaz-Agudo B, Gerv´as P, Gonz´alez-Calero PA. Poetry generation in COLIBRI. In: Susan C, Alun P, eds. Proc. of the 6th European Conf. on Case Based Reasoning (ECCBR 2002). Aberdeen: Springer-Verlag, 2002. 157−159.

[24] Kempe V, Levy R, Graci C. Neural networks as fitness evaluators in genetic algorithms: Simulating human creativity. In: Moore JD, Stenning K, eds. Proc. of the 23rd Annual Conf. of the Cognitive Science Society. Edinburgh: Lawrence Erlbaum Associates, 2001. Poster Session 1. http://citeseerx.ist.psu.edu/viewdoc/download? doi=10.1.1.126.1464&rep=rep1&type=pdf

[25] Gruber H, Davis S. Inching our way up Mount Olympus: The evolving systems approach to creative thinking. In: Sternberg RJ, ed. The Nature of Creativity. New York: Cambridge University Press, 1988. 243−269.

[26] Sims K. Artificial evolution for computer graphics. Computer Graphics, 1991,25(4):319−328.

[27] Manurung HM. An evolutionary algorithm approach to poetry generation [Ph.D. Thesis]. Edinburgh: University of Edinburgh, 2003.

[28] Su JS, Zhou CL, Li YH. The establishment of the annotated corpus of Song dynasty poetry based on the statistical word extraction and rules and forms. Journal of Chinese Information Processing, 2007,21(2):52−57 (in Chinese with English abstract).

[29] Su JS. Research on the establishment of Song dynasty poetry corpus and the computational methods of style identification and emotion analysis [MS. Thesis]. Xiamen: Xiamen University, 2007 (in Chinese with English abstract).

[30] Ying Y, Zhou F, Zhou CL. A research on emotion tagging of chinese understanding by designing an experiment system. Journal of Chinese Information Processing, 2002,16(2):27−33 (in Chinese with English abstract).

[31] Wang ZP, Liu ZM. Dictionary of Song Ci. Nanjing: Fenghuang Publishing House, 2003 (in Chinese).

[32] Jin QH. Dictionary of Allusion to Complete Song Ci. Jilin: Jilin Literature and History Press, 1991 (in Chinese).

[33] Pan S. Dictionary of Rhythm of Song Ci. Shanxi: Shanxi People's Publishing House, 1982 (in Chinese).

[34] The Authorized Collection of Ci Poem. Beijing: Chinese Book Store, 1983 (in Chinese).

[35] Long YS. Rhythm of Ci Poem of Tang and Song Dynasty. Shanghai: Shanghai Guji Publishing House, 1978 (in Chinese).

[36] Tang GZ. Complete Song Ci. Shanghai: Chinese Publishing House, 1997 (in Chinese).

[37] Zhang L, Zhang B. Research on the mechanism of genetic algorithms. Journal of Software, 2000,11(7):945−952 (in Chinese with English abstract). http://www.jos.org.cn/ch/reader/view_abstract.aspx?flag=1&file_no=20000712&journal_id=jos

[38] Church KW, Hanks P. Word Association norms, Mutual Information, and Lexicography. Computational Linguistics, 1990,16(1): 22−29.

[39] Hu JF, Yu SW. Word meaning similarity analysis in Chinese ancient poetry and its applications. Journal of Chinese Information Processing, 2002,16(4):39−44 (in Chinese with English abstract).

[40] Liu Q, Li SJ. Similarity computation of vocabulary semantic based on HowNet. In: Proc. of the 3rd Symp. on Semantics of Chinese Words(in Chinese with English abstract). http://www.keenage.com/html/paper.html.

[41] Dong ZD. Research on internal semantic of vocabulary and Chinese knowledge dictionary. Applied Linguistics, 2000,(1):29−31 (in Chinese with English abstract).

附中文参考文献:

[1] 刘岩斌,俞士汶,孙钦善.古诗研究的计算机支持环境的实现.中文信息学报,1996,11(1):27−35.

[2] 穗志方,俞士汶,罗凤珠.宋代名家诗自动注音研究及系统实现.中文信息学报,1998,12(2):44−53.

[3] 罗凤珠,李元萍,曹伟政.中国古代诗词格律自动检索与教学系统.中文信息学报,1999,13(1):35−42.

[4] 罗凤珠.诗词语言切分与语意分类标记之系统设计及应用.In:第 4 届数位典藏技术研讨会议论文集.2005.

[5] 俞士汶,胡俊峰.唐宋诗之词汇自动分析及应用.语言暨语言学,2000,4(3):631−647.

[6] 胡俊峰.基于词汇语义分析的唐宋诗计算机辅助深层研究[博士学位论文].北京:北京大学,2001.

[7] 李良炎.基于词联接的自然语言处理技术及其应用研究[博士学位论文].重庆:重庆大学,2004.

[8] 李良炎,何中市,易勇.基于词联接的诗词风格评价技术.中文信息学报,2005,19(6):98−104.

[9] 易勇.计算机辅助诗词创作中的风格辨析及联语应对研究[博士学位论文].重庆:重庆大学,2005.

[10] 费越.汉语语义的多层次集成研究--及春联艺术系统设计[博士学位论文].北京:中国科学院自动化研究所,1999.

[11] 周明.微软对联生成系统.微软亚洲研究院自然语言组.2006.http://duilian.msra.cn/

[12] 周昌乐.心脑计算举要.北京:清华大学出版社,2003.

[28] 苏劲松,周昌乐,李翼鸿.基于统计抽词和格律的全宋词切分语料库建立.中文信息学报,2007,21(2):52−57.

[29] 苏劲松.全宋词语料库建设及其宋词风格与情感分析的计算方法研究[硕士学位论文].厦门:厦门大学,2007.

[30] 应英,周峰,周昌乐.汉语情感意义的机器标注研究初探.中文信息学报,2002,16(2):27−33.

[31] 王兆鹏,刘尊明.宋词大辞典.南京:凤凰出版社,2003.

[32] 金启华.全宋词典故考释辞典.吉林:吉林文史出版社,1991.

[33] 潘慎.词律辞典.山西:山西人民出版社,1982.

[34] 钦定词谱.北京:中国书店,1983.

[35] 龙榆生.唐宋词格律.上海:上海古籍出版社,1978.

[36] 唐圭璋.全宋词.上海:中华书局,1997.

[37] 张铃,张钹.遗传算法机理的研究.软件学报,2000,11(7):945−952. http://www.jos.org.cn/ch/reader/view_abstract.aspx?flag=1&file_no=20000712&journal_id=jos

[39] 胡俊峰,俞士汶.唐宋诗中词汇语义相似度的统计分析及应用.中文信息学报,2002,16(4):39−44.

[40] 刘群,李素建.基于《知网》的词汇语义相似度计算.见:第 3 届中文词汇语义学研讨会论文集.http://www.keenage.com/html/paper.html

[41] 董振东.汉语知识词典及词汇内部语义描述研究.语言文字应用,2000,(1):29−31.

周昌乐(1959－),男,江苏太仓人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为人工智能及其应用技术.

丁晓君(1975－),女,讲师,主要研究领域为审美认知.

游维(1982－),女,博士生,主要研究领域为自然语言处理.