CN43-1258/TP
ISSN 1007-130X

计算机工程与科学
COMPUTER ENGINEERING & SCIENCE

2012 年第 34 卷第 5 期
Vol. 34, No. 5, 2012

文章编号:1007-130X(2012)05-0178-06

基于遗传算法的自动组卷系统的设计与实现*

Design and Implementation of an Automatic Test Paper Generation System Based on Genetic Algorithms

张　琨,杨会菊,宋继红,赵学龙
ZHANG Kun, YANG Hui-ju, SONG Ji-hong, ZHAO Xue-long
(南京理工大学计算机科学与技术学院,江苏 南京 210094)
(School of Computer Science and Technology,
Nanjing University of Science and Technology, Nanjing 210094, China)

摘　要:组卷问题是一个在一定约束条件下的多目标参数优化问题,采用传统的数学方法求解十分困难,自动组卷的效率和质量完全取决于试题库设计以及抽题算法的设计。本文以省级《数据结构》精品课程建设为背景,在分析传统组卷算法的优缺点和组卷策略参数的基础上,选用遗传算法,设计并实现了一个自动组卷系统。该算法按照试题类型、数量、难度、区分度、分值和时间等约束条件进行快速搜索并寻找最优解,其中采用分组自然数编码,减少了染色体长度空间;运用自适应理论改进交叉概率及变异概率,使得算法总能找到合适的交叉和变异概率。系统采用 C#.NET 编程实现,目前已应用于实际教学,取得了良好的教学效果。

Abstract: The generation of test papers is an optimized problem with multi-objective parameters under a certain restrictive condition. The optimization is implemented very difficultly by traditional methods. The quality and efficiency of auto-generation is determined by the design of test question databases and algorithms to extract questions. Based on the construction of the provincial excellent course Data Structure, an analysis of the features of traditional test paper algorithms and the parameters of strategy, this paper presents the design and realization of an automatic test paper generation system based on genetic algorithms. This algorithm searches for the best answer according to such restrictive conditions as test question types, quantity, difficulty level, difference level, score and answering time. In addition, the natural code is used in this algorithm in order to decrease the space of chromosomes. The crossover probability and the mutation probability are improved with the self-adaptation theory, so that the proper numbers of crossover probability and mutation probability can be found. After the accomplishment based on C#.NET, the system has been applied in practice and achieved good effect.

关键词:自动组卷;遗传算法;试题库;数据结构

Key words: automatic test paper generation; genetic algorithms; item bank; data structure

doi: 10.3969/j.issn.1007-130X.2012.05.035

中图分类号:TP301　　**文献标识码**:A

* 收稿日期:2011-03-15;修订日期:2011-06-20
基金项目:江苏省《数据结构》精品课程建设项目(2010104);南京理工大学《数据结构》精品课程建设项目(201008)
通讯地址:210094 江苏省南京市孝陵卫 200 号南京理工大学计算机科学与技术学院
Address: School of Computer Science and Technology, Nanjing University of Science and Technology, 200 Xiaolingwei, Nanjing, Jiangsu 210094, P. R. China

1　引言

《数据结构》是计算机课程体系中的核心专业基础课程,是计算机程序设计的重要理论和实践基础,提高该课程的教学质量对提高学生的实践和创新能力具有重要的意义。其中,自动组卷系统的实现即是提高教学质量和促进教学改革的一项重要内容。自动组卷是指根据用户输入的组卷要求和考核目的,按一定算法从试题库中选择一组题目生成一份试卷的过程。它不仅能实现教考分离,而且通过组卷算法的设计还能提高出题的科学性,更客观地体现教与学的实际水平[1]。本文实现的自动组卷系统由两部分组成:试题库和应用软件。应用软件提供用户访问试题库的接口,其核心和关键是自动组卷算法。

2　组卷系统的设计

2.1　组卷系统的功能设计

本系统的功能设计主要针对学生和教师两类用户,其功能结构如图 1 所示。

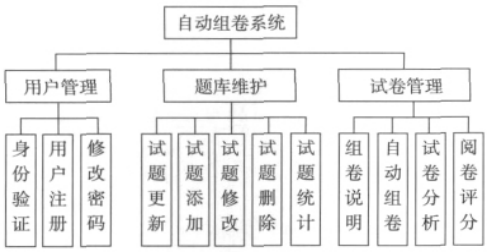

图 1　系统功能结构图

2.2　系统设计的相关指标和要求

2.2.1　试卷的评价指标

难易适当和鉴别力强是一份好试卷的要素,通常用试卷的难度和区分度来衡量。由于试题的难度和区分度在随后小节介绍的系统组卷算法中需用到,因此在此处一并介绍。表 1 是本系统所采用的评价指标的计算公式[2]。

表 1　试卷的评价指标计算公式

试卷的难度 P	试卷的区分度 D	试题 i 的难度 P_i	试题 i 的区分度 D_i
$P=1-\overline{X}/W$	$D=(\overline{X_h}-\overline{X_l})/W$	$P_i=1-\overline{X_i}/W_i$	$D_i=(\overline{X_{ih}}-\overline{X_{il}})/W_i$

i 为试题编号($i=1,2,3,\cdots,NR$),NR 为试卷要求的试题个数,$\overline{X}$ 为试卷平均分,W 为试卷满分,$\overline{X_i}$ 为所有考生第 i 题的平均分,W_i 为第 i 题的满分。按照试卷考分排序后,$\overline{X_h}$ 是指所有考生中排分在前 27%的学生的试卷平均分,$\overline{X_l}$ 是指后 27%的学生的试卷平均分。按照第 i 题的考分排序后,$\overline{X_{ih}}$ 是指所有考生中排分在前 27%的考生第 i 题得分的平均值,$\overline{X_{il}}$ 是指后 27%的考生第 i 题得分的平均值。由于 P_i 和 D_i 需要依据学生做完试卷后的成绩确定,因此本系统中 P_i 和 D_i 的初值是根据经验值预先设定的。

2.2.2　试题库的设计要求

试题库是大量具有一定参数的试题的有机组合。试题库的设计要求如下:

(1)试题库中的题目按题型分类,以便于采用遗传算法。为实现此项要求,本系统在对试题库中的试题编码时采用 3.3.1 节和 3.3.2 节的编码和分类方式;

(2)试题库的题量足够大,比例合理,题型难度接近正态分布,知识点分布均匀;

(3)试题库中的每一道试题由题目属性指标和题目本身组成,其中题目属性指标能反映题目的唯一性;

(4)试题库能够动态维护,包括添加、删除、修改和统计等功能。

为满足上述设计要求,实现的数据库中主要包括:试题难度表、试题区分度表、试题类型表、知识点表、各类型试题的信息表(5 个表)、试卷信息表、学生答案信息表、学生成绩信息表、学生信息表和管理员信息表等。其中某些关键表的结构如下:

(1)试题难度表:保存试题难度信息。2 个字段:难度级别编码(7 个级别,int 类型)和难度级别名称(10 个字符)。本系统中试题的难度分级规则如表 2 所示。

(2)试题区分度表:保存试题区分度信息。2 个字段:区分度级别编码(4 个级别,int 类型)和区分度级别名称(10 个字符)。本系统中试题区分度分级规则如表 3 所示。

表 3　区分度值与区分度级别对应表

试题区分度级别	1(好)	2(较好)	3(一般)	4(差)
P_i	(0.4,1]	[0.3,0.4]	[0.2,0.3)	[0,0.2)

表 2　难度值与难度级别对应表

试题难度级别	7(难)	6(较难)	5(中偏难)	4(中等)	3(偏易)	2(较易)	1(易)
P_i	(0.9,1]	(0.8,0.9]	(0.6,0.8]	(0.4,0.6]	(0.2,0.4]	(0.1,0.2]	[0,0.1]

(3)试题类型表:保存试题类型信息。2 个字段:试题类型编码(5 个类型,int 类型)和类型名称(20 个字符)。结合数据结构考试需求,本系统试题类型为 5 种:选择题、填空题、判断题、简答题和程序题。

(4)试题信息表:保存试题信息。除选择题有 14 个字段外,其他四类题型仅有 10 个字段。optionA、optionB、optionC 和 optionD 只在选择题类型的试题信息表中存在,如表 4 所示。

(5)试卷信息表:存储试卷总体描述信息。13 个字段,如表 5 所示。

3 自动组卷算法的设计与实现

自动组卷问题实质上是在一定约束条件下的多目标参数优化问题,因此应用具有收敛速度快和全局寻优特性的遗传算法进行求解是目前较为实用和有效的方法。以下就该算法所涉及的三个主要方面:组卷策略、组卷目标和遗传算法寻优求解进行介绍。其中,算法通过个体的适应度函数值(3.3.3 节)来区分群体中个体的优劣,各个体的产生经历"生成初始种群-选择-交叉和变异"过程最终确定,各个体是否被选择作为最终试卷中的试题的评价指标是个体的适应度值。

3.1 组卷策略

组卷策略是自动组卷算法的基础,是指试卷控制参数的组织和表现形式。本系统选择的试卷控制参数是:试卷难度(P)、试卷区分度(D)、试卷的总分值(W)、试卷的估计用时(T)、试卷的试题类型 j($j=1,2,3,4,5$,数字 1~5 代表 5 种不同的试题类型)、试卷的试题个数(N)、各题型的试题个数(TN_j)、各题型的分数(TW_j)。依据实际组卷经验,往往会对组成试卷的各题型进行个数和分数的限定,因此加入了各题型的试题个数和分数这两个控制参数。此外,由于 2.2.1 节中介绍的试卷难度和区分度的计算需要考试结束后才能统计,因此在应用遗传算法进行寻优求解时,试卷难度 P 和区分度 D 的计算采用如下公式:

$$P=\sum_{j=1}^{5}(TP_j\times TW_j)/W$$

$$D=\sum_{i=1}^{N}(D_i\times W_i)/W$$

其中,TP_j 表示各题型的难度。依据以往阅卷经验,各题型的难度对试卷质量的影响也较大,如选择题和填空题等批阅严格的题型的难度不易过高,否则试卷均分往往较低。因此,不仅在组卷目标中加入对 TP_j 的限定,在 P 的计算中也引入了 TP_j 参数,它的计算公式如下:

$$TP_j=\sum_{i=1}^{N}(P_i\times W_i)/TW_j,i\in j\text{ 类型的试题}$$

3.2 组卷目标

组卷目标是指对所生成试卷的期望目标。在实际运算中,用对 3.1 节中的试卷控制参数的定量值或约束限定来表示。本系统中的组卷目标如下:

(1)试卷的实际试题个数等于试卷要求的试题个数(NR),即 $NR=N=\sum_{j=1}^{5}TN_j$;

(2)试卷中所有题型的实际总分和等于试卷要求的总分(WR),即 $WR=W=\sum_{j=1}^{5}TW_j$;

(3)试卷中所有试题的实际考试时间总和等于试卷要求的考试时间(TR),即 $TR=T=\sum_{i=1}^{N}T_i$,其中 T_i 表示第 i 题的答题时间;

(4)试卷的实际难度(P)和要求的难度(PR)相差不大,即 $(|PR-P|/PR)\leqslant\omega_P$,其中 ω_P 表示试卷的难度误差限定比率;

(5)试卷的实际区分度(D)和要求的区分度(DR)相差不大,即 $(|DR-D|/DR)\leqslant\omega_D$,其中 ω_D 表示试卷的区分度误差限定比率;

(6)各题型的实际难度(TP_j)和要求的难度

表 4 试题信息表结构

ID	que_type	que_difficulty	que_differentation	que_score	que_time	que_subject	que_point	optionA	optionB	optionC	optionD	que_answer	Note
试题编号	试题类型	试题难度	试题区分度	题满分值	答题时间	试题题干	题知识点	选项 A	选项 B	选项 C	选项 D	试题答案	备注

表 5 试卷信息表结构

title	score	subject	num1	num2	num3	num4	num5	type1	type2	type3	type4	type5
试卷标题	试卷总分	试题总数	选择题个数	填空题个数	判断题个数	简答题个数	程序题个数	选择题型的分值	填空题型的分值	判断题型的分值	简答题型的分值	程序题型的分值

(TPR_j)相差不大,即 $(|TPR_j - TP_j|/TPR_j) \leqslant \omega_{P_j}$,$\omega_{P_j}$ 表示题型 j 的难度误差限定比率。

3.3　基于遗传算法的组卷算法

3.3.1　编码方案

为避免简单遗传算法(Simple Genetic Algorithm,简称 SGA)中种群染色体编码过长的问题,本系统采用自然数分区编码方法[3]。首先根据题型将编码分区(本系统中分为 5 个区),随后依次为每种题型根据组卷要求从试题库中选题,每个选中试题代表一个基因,基因编号就是该试题的编号,一组选择结束时即由这些基因组成一个染色体,染色体的基因个数等于 N。

3.3.2　生成初始种群

为避免 SGA 中初始种群解空间分布不均匀的问题,本系统在生成初始种群时利用了分类策略。首先将试题库中的试题根据题型分类为集合 A_j,随后对集合 A_j 再依据知识点 $k(k=1,2,3,\cdots,m)$ 分类为集合 B_{jk},最后在试卷知识点覆盖率的限定下从 B_{jk} 中按照试题编号递增的顺序选择 TN_j 个题目,最终构成个体样本。该方法不仅能满足知识点范围、题型和题量的要求,且各题型的试题题号均按升序排列,可有效降低交叉操作产生非法染色体的概率。

3.3.3　适应度函数

适应度函数由目标函数变换而得,在遗传算法中以适应度函数值来区分群体中个体的优劣,其值越大越好。在计算中往往采用个体对应的试题属性与组卷要求之间的误差作为目标函数,进而转换得到适应度函数。本系统的种群中个体 $b(b=1,2,3,\cdots,B,B$ 为种群个体数量)的目标函数定义为:

$$F_b = F_{1b} + F_{2b} + F_{3b} + F_{4b} + F_{5b}$$

其中,$F_{1b} = |PR - P|$,$F_{2b} = |DR - D|$,$F_{3b} = |TR - T|$,$F_{4b} = |WR - W|$,$F_{5b} = \sum_{j=1}^{5} |TPR_j - TP_j|$。

F_b 值越小,意味着个体的相应属性越接近试卷的要求。采用指数比例变换后的适应度函数定义如下:

$$Y_b = Y_{1b} + Y_{2b} + Y_{3b} + Y_{4b} + Y_{5b}$$

$$Y_\phi = e^{-\beta F_\phi}, q = 1,2,3,4,5$$

其中系数 β(本系统中取 0.02)决定了复制的强制性。

3.3.4　选择

本系统采用最佳个体保持法和轮盘赌相结合的方法实现选择操作[4],其步骤如下:

(1)依次计算种群中每个个体的适应度值 Y_b;

(2)按适应度值排序,将值最高的前 10%的染色体直接遗传到下一代;

(3)计算每个个体的选择概率 $G_b = Y_b / \sum_{b=1}^{B} Y_b$ 和累计选择概率 $C_b = \sum_{l=1}^{b} G_l$;

(4)随机产生一个概率 $r \in [0,1]$;

(5)依次用 C_b 与 r 进行比较,若 $r \leqslant C_1$,则种群中编号为 1 的染色体被选中,若 $C_{b-1} < r \leqslant C_b (2 \leqslant b \leqslant B)$,则编号为 b 的染色体被选中;

(6)重复第(4)和第(5)步,直至产生满足所需数量的个体。

3.3.5　交叉和变异

为加快搜索效率并防止算法陷入局部最优,本系统采用 Srinvivas[5] 提出的自适应交叉概率 Z_c 和变异概率 Z_m,其计算公式如下:

$$Z_c = \begin{cases} k_1 \times (Y_{\max} - Y') / (Y_{\max} - Y_{avg}), Y' \geqslant Y_{avg} \\ k_2, Y' < Y_{avg} \end{cases}$$

$$Z_m = \begin{cases} k_3 \times (Y_{\max} - Y_b) / (Y_{\max} - Y_{avg}), Y_b \geqslant Y_{avg} \\ k_4, Y_b < Y_{avg} \end{cases}$$

其中,$Y_{\max}$ 为群体中最大的适应度值,Y_{avg} 为当前群体的平均适应度值,Y' 为要交叉的两个个体中的较大的适应度值,Y_b 为要变异个体的适应度值,k_1,k_2,k_3,k_4 为(0,1)之间的值,本系统中取 $k_1 = k_2 = 0.8$,$k_3 = 0.1$,$k_4 = 0.2$。

3.3.6　停机条件

当达到规定的最大遗传代数 $H_{\max}$ 或者连续 $H_{\max}/5$ 代内得到的最优个体适应度值无变化则停机。$H_{\max}$ 一般取 500～1 000 代。

4　组卷系统的实现

为便于进行在线考试,本系统的实现基于 B/S 分布式体系架构。具体开发平台是微软的.NET 平台,开发语言是 C#[6],数据库系统采用 SQL Server2005。实现的组卷算法的伪代码如下所示:

```
/*基于遗传算法的组卷算法 TestPaper_GeneticAlgorithm*/
/*ds:数据库,tbname:表名,result:最优染色体结果数组,num:染色体个数,Parameters:组卷参数*/
```

```
public void TestPaper_GeneticAlgorithm(ds,tbname,
result,num,Parameters){
    /*设置组卷约束条件,最大遗传代数 agemax,种群规
模 pmax*/
      int agemax=800;int pmax=30;int[,] pp=new int
[pmax, num];double[] av=new double[pmax];
      int[] rold; int[] rnew; /*数组定义:用于记录适
应值排序前 10%的染色体编号*/
      createPopulation(ds,num,tbname,pp,pmax); /*
生成试卷初始种群,pp:存放初始种群*/
      getPAdaptiveValue(ds,tbname,pp,av,num,pmax,
Parameters);//各染色体适应度值计算存于 av
      getMaxF(av,pmax,oldav,rold); /*选择适应度值
排序的前 10%的染色体存入 rold*/
      for (i = 0; i < agemax; i++) {
        doSelect(ds,tbname,pp,pmax,num,av,Parame-
ters);//选择过程,生成的子代个体存于 pp
        if (num > 2) {
          doCrossover(ds,tbname,pp,pmax,num,Pa-
rameters,av);//交叉过程,pp:存放交叉后的子代种群
          getPAdaptiveValue(ds,tbname,pp,av,num,
pmax,Parameters);//计算各染色体适应度存于 av
        }
        doMutation(ds,tbname,pp,av,num,pmax);
//变异过程,pp:存放变异后的子代种群
        getPAdaptiveValue(ds,tbname,pp,av,num,
pmax,Parameters); //计算各染色体适应度值存于 av
        getMaxF(av,pmax,newav,rnew); /*选择适应
度值排序的前 10%的染色体存入 rnew*/
        if (valueSame(rold,rnew))/*判断父、子代适应
度值是否相同,用 count 记录相同次数*/{
          count++;
          if (count >= agemax / 10) break;/*达到最
大适应度值相同次数 agemax / 10,算法结束*/ }
          else { rold = rnew; count = 0; }
      }//达到最大迭代次数时,算法结束
      for (i = 0; i < num; i++) { result[i] = pp[r,
i]; /*将结果存入 result 数组*/ }
    }
```

在调用该算法时,需传入存有系统数据试题信息的数据库名和相应要访问的表名。该算法是系统的关键核心,当用户点击界面左侧的“自动组卷”按钮后由系统自动调用。

教师用户登录系统后,在图 2 所示界面中根据组卷要求,设置试题个数、难度、区分度、答题时间以及试题分值等参数后,点击开始组卷,系统即会自动生成试卷,并以网页形式显示。图 3 显示的就是一份自动生成的试卷界面。

图 2 组卷要求参数设置界面

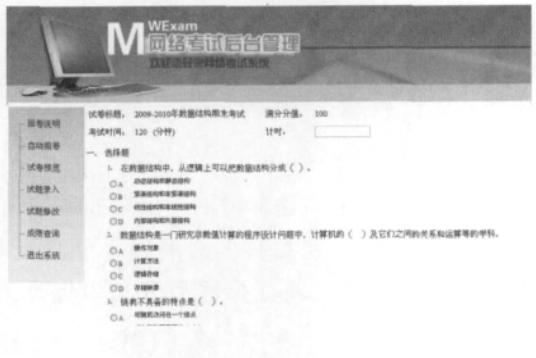

图 3 生成的一套试卷

本组卷系统自 2010 年实现以来,已在本学院《数据结构》考试组卷过程中应用。组卷的默认要求是:试卷总分 100,用时 120 分钟。从实际使用效果看,本系统生成的试卷知识点覆盖合理,组卷效率和成功率均较高。

5 结束语

组卷问题是一个在一定约束条件下的多目标求解问题,在应用遗传算法解决智能组卷问题时,通过根据实际问题设计的组件编码方案和有效的选择、交叉、变异算子,直接在问题解上进行遗传操作,提高了遗传算法的搜索能力,取得了良好的组卷效果。本组卷系统目前已应用于实际教学中,作为本门课程省级精品教学建设的一部分,取得了良好的教学效果。

参考文献:

[1] 黄国政. 基于遗传算法的自动组卷系统的设计与实现:[硕士学位论文][D]. 南京:南京理工大学, 2008.

[2] 余红朝. 基于遗传算法的组卷研究及题库系统实现:[硕士学位论文][D]. 重庆:重庆大学,2008.

[3] 王玉芬, 贾燕茹, 郭晓娟. 遗传算法的改进及在自动组卷系统中的应用[J]. 信阳师范学院学报:自然科学版, 2009,22(1):130-133.

[4] 刘英, 王琰, 杜炎, 等. 基于自适应算子的遗传算法多目标

智能组卷[J]. 计算机应用, 2008,28(6):22-24.
[5] 吴飞. 自适应遗传算法解决组卷问题的探讨[J]. 重庆科技学院学报(自然科学版), 2007, 9(2):132-135.
[6] 靳华. ASP. NET 3. 5 宝典[M]. 北京:电子工业出版社, 2009.

张琨(1977-),女,河北昌黎人,博士,副教授,研究方向为计算机应用和网络安全。**E-mail**:zhangkun@mail. njust. edu. cn

ZHANG Kun, born in 1977, PhD, associate professor, CCF member (E200006375M), her research interests include computer application, and network security.

杨会菊(1988-),女,河北辛集人,研究方向为计算机应用和网络安全。**E-mail**:yanghuiju2006@sina. com

YANG Hui-ju, born in 1988, her research interests include computer application, and network security.

宋继红(1958-),女,黑龙江虎林人,工程师,研究方向为计算机应用。**E-mail**:jsjx@mail. njust. edu. cn

SONG Ji-hong, born in 1958, engineer, her research interest includes computer application.

2012 年第 4 届计算与信息科学国际学术会议征稿通知

会议网站: www. iccis. net/

第 4 届信息与计算科学国际学术会议(The fourth International Conference on Computational and Information Sciences (ICCIS2012)) **将于 2012 年 8 月 10 日至 12 日在重庆市召开,主办单位为美国纽约州立大学,会议得到 IEEE 成都分会的技术支持,承办单位为西南石油大学,会议论文集由 IEEE CS 出版并送交 EI 检索。大会将主要围绕高性能计算、并行与分布式计算/算法、生物信息与生物计算、数据挖掘与隐私保护、文本、视频及多媒体技术、智能信息处理、图形、图像和视频处理、电子商务与 web 挖掘等 8 个主题展开讨论。会议将邀请国内外知名学者参会,共同就"信息与计算科学"前沿、热点问题展开深入的交流与讨论。希望广大科研工作者、研究生和产业界人士积极投稿。**

所属学科:计算机科学与技术、信息科学与系统科学、工程与技术科学的计算、计算生物信息学、计算统计学计算经济管理学、数据处理和分析石油工程

会议论文集检索:EI,ISTP

开始时间:2012-08-10　　**结束时间:**2012-08-12

征稿范围

大会议题将主要包含以下内容(但不仅限于此):

(1)计算科学与工程:高性能计算与算法、并行与分布式计算/算法、数值模拟和数值算法、计算数学、计算物理、计算化学、计算生物学、计算解剖学、计算财经学、大型科学和工程计算、应用数值分析、生物建模与仿真、基于 Web 和网格的计算和模拟。

(2)信息科学计算:数据挖掘与算法分析、信息检索、医疗信息、生物信息学,基因组学和生物统计学、计算图形学、图像处理、电子商务与 Web 数据挖掘、文本,视频,多媒体挖掘、智能分析和评价、安全和信息的隐私、科学可视化、高性能的信息处理和算法、计算信息学中的并行,分布式和可扩展算法。

(3)先进的计算理论及应用:神经网络、进化计算和遗传算法、模糊系统与软计算、蚁群优化、粒子群优化、人工鱼群算法、人工生命和人工免疫系统、系统生物学和神经生物学、支持向量机、粗糙和模糊粗糙集、知识发现与数据挖掘、内核方法、监督和半监督学习、云计算、进化学习系统、混合系统。

主办单位: State University of New York at Brockport, USA

承办单位:西南石油大学

会议主席: Wensheng Shen, State University of New York at Brockport, USA.

组织委员会主席: Jie Chen, University of Missouri-Kansas City, USA

程序委员会主席: Yin Wang, Lawrance Technological University, USA.

重要日期:

论文截稿日期:2012-05-10　　**论文录用通知日期:**2012-05-20　　**交修订版截止日期:**2012-05-30

通讯地址:四川省成都市新都区西南石油大学计科院

会务组联系方式:

联系人:耿老师 杨老师

联系电话:028－83037528(O);13980668364(M);13540478330(M)

传真: 028－83037503