第 38 卷 第 11 期
Vol.38 No.11

计 算 机 工 程
Computer Engineering

2012 年 6 月
June 2012

·软件技术与数据库· 文章编号：1000—3428(2012)11—0034—04 文献标识码：A 中图分类号：TP301.6

基于多特征的个性化图书推荐算法

李克潮[1]，梁正友[2]

(1. 广西民族师范学院图书馆，广西 崇左 532200；
2. 广西大学计算机与电子信息学院，南宁 530004)

摘 要：现有推荐算法计算读者之间或图书之间的相似性不准确、推荐精确度不高。为此，提出一种基于多特征的个性化图书推荐算法。根据中图分类法及图书的特征向量计算图书的相似性，依据读者的特征向量及借阅记录计算读者的相似性。在此基础上产生 2 种预测结果并对其进行加权，产生最终推荐。实验结果表明，该算法具有较高的图书推荐精确度。

关键词：中图分类法；图书特征向量；读者特征向量；相似性；推荐算法；中图分类树；专业分类树

Personalized Book Recommendation Algorithm Based on Multi-feature

LI Ke-chao[1], LIANG Zheng-you[2]

(1. Library, Guangxi Normal University for Nationalities, Chongzuo 532200, China;
2. School of Computer and Electronics Information, Guangxi University, Nanning 530004, China)

【Abstract】To address the problem of recommendation algorithm computing readers similarity or books similarity with low accuracy and recommendation quality, Personalized recommendation algorithm based on multi-feature is proposed. It computes books similarity based on Chinese library classification method and books feature vector, computes readers similarity based on readers feature vector and borrow records. Based on this, two prediction results are produced. The last recommendation is produced to readers by weight of this two prediction results. Experimental result shows that the proposed algorithm achieves more recommendation accuracy on books.

【Key words】Chinese library classification method; book feature vector; reader feature vector; similarity; recommendation algorithm; Chinese library classification tree; professional classification tree

DOI: 10.3969/j.issn.1000-3428.2012.11.011

1 概述

随着网络的发展，数字图书馆资源越来越丰富。用户(下文统称为读者)不知道如何在众多的图书资源中快速寻找到自己真正需要的资源(下文统称为图书)[1-5]。

个性化推荐系统因此诞生，它通过分析读者的行为，获取读者的兴趣偏好，实现个性化推荐。其中，协同过滤推荐是最成功的推荐技术之一[6-8]，其通过计算读者或图书的相似性，产生读者或图书的最近邻居，再由最近邻居向目标读者推荐。

然而，现有的协同过滤推荐，一方面计算图书相似性时，需要计算目标图书与其他全部图书的相似性，算法复杂度大。另一方面，不考虑读者不同时间段访问图书、访问图书的时间长短对推荐的影响，也不考虑受大众欢迎的图书对推荐的影响。

针对以上问题，本文提出一种基于多特征的个性化图书推荐算法。首先，根据中图分类号，构建中图分类树，再根据中图分类树计算基于分类号的图书相似性。从图书的页数、被外借与归还的时间间隔、平均被外借的次数、出版日期计算图书受大众读者欢迎的程度。根据本科专业目录，构建专业分类树，再根据专业分类树计算基于专业的读者相似性。综合专业、年级、性别相似性计算读者之间的综合特征相似性。最后，综合读者和图书的借阅预测，推荐图书给读者。

2 基于多特征的图书推荐算法

2.1 图书特征

图书馆的任意 2 本图书都有相似性和相异性的特征。这些特征对向读者推荐图书至关重要。例如：与计算机相关的图书，其封面都贴有以“TP”开头的索书号。通过索书号，可知道图书属于中图分类号中的哪一类。属于同一中图分类号的图书，具有相似的主题。

每本书都有总页数，总页数比较多，读者一般需要比较长的时间才能看完。写得比较好的图书会受到较多读者的欢迎，在相同的时间内被借阅的次数较多。对于新书，即出版日期较晚的图书，也比较容易得到读者的青睐，一般被借阅的可能性比旧书的大。因此，分类号、总页数、被借阅次数及出版日期是本文考虑的图书特征。

2.1.1 中图分类树

中国图书馆图书分类法[9]，简称中图分类法，具有从总到分、从一般到具体的特点。其采用汉语拼音 22 个字母和阿拉伯数字相结合的混合编码，每个字母表示一个大类。字母

基金项目：广西自然科学基金资助项目(桂科自 0832059)；广西民族师范学院 2011 年度基金资助项目(XYYB2011030)

作者简介：李克潮(1982－)，男，硕士，主研方向：个性化推荐系统；梁正友，教授、博士

收稿日期：2011-09-21 **E-mail：**my472360924@sina.com

后面跟随数字，用以表示大类下的子类，数字使用小数点制进行编号，其分类可描述为如图 1 所示的中图分类数。

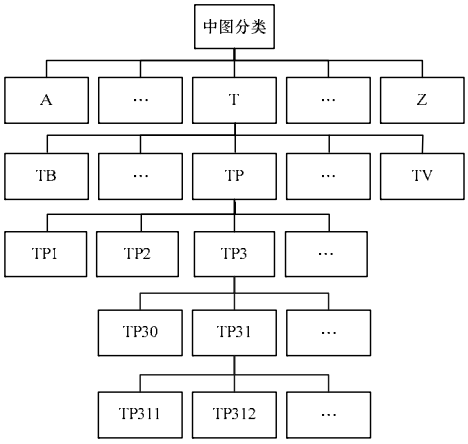

图 1　中图分类树

2.1.2　基于分类号的图书相似性

传统推荐算法计算资源之间相似性，需要遍历整个资源表，算法复杂性较大。本文根据中图分类法计算资源相似性，减少了算法复杂性。

中图分类法将属于相同的学科、具有相同主题的图书归为一大类。将属于更细相同学科、具有更多相同主题的图书归为同一大类下的子类。如此依次归类下去，相似性最强的图书的分类号处于中图分类树的最下层。

根据图 1 可知，比较 2 本图书的相似性，需要先比较索书号中分类号最左边的字母。字母相同时，按照数字大小比较字母后的第 1 位数字，第 1 位数字相同时，比较第 2 位数字的大小，如此类推。可构建目标读者 r(reader，记为 r)未外借图书 j 与已外借图书 i 基于分类号特征向量的相似性 $simb(j,i)$ ：

$$simb(j,i)=\begin{cases}\dfrac{Layerb(j,i)-1}{Layerb(all)} & c(j)\neq c(i)\\ 1 & c(j)=c(i)\end{cases} \tag{1}$$

其中，$Layerb(j,i)$ 为图书 j 的分类号 $c(j)$ 与 i 的分类号 $c(i)$ 在中图分类树中最近的共同父结点所在的层；$Layerb(all)$ 为图书分类树的总层数。对 n 位读者 m 本图书的数据，现有协同过滤推荐计算图书相似性，如采用 Pearson 相关相似性方法，需要的算法复杂度为 $O(n\times m)$，而根据式(1)所需的复杂度为 $O(1)$。

2.1.3　图书的受欢迎程度

不同的读者对图书的喜好，除了具有特殊性外，还具有普遍性。通常，受大众欢迎的图书，目标读者 r 喜欢的可能性比较大。这些书具有这样的特征：平均每次被外借的时间长，被外借次数多，是新书。

假设外借过页数为 $Page(j)$ 的图书 j 的所有读者集合为 $S(allbj)$，读者 $r'\in S(allbj)$ (其中，$r\notin S(allbj)$)外借和归还 j 的日期分别为 $Bo_{date}(r',j)$ 和 $Re_{date}(r',j)$，图书 j 基于页数特征向量的外借与归还时间间隔 $Re_{Interal}(r',j)$ 定义为：

$$Re_{Interal}(r',j)=\frac{Re_{date}(r',j)-Bo_{date}(r',j)}{Page(j)} \tag{2}$$

所有外借过图书 j 的读者，平均每次外借图书 j 到归还图书 j 的时间间隔 $Av_{Interval}(j)$ 为：

$$Av_{Interval}(j)=\frac{\sum_{r'\in S(allbj)}Re_{Interval(r',j)}}{records(allbj)} \tag{3}$$

其中，$records(allbj)$ 为外借图书 j 的总记录数。

设图书 j 入库(指图书放入书库，读者才可从书库外借)的日期为 $Indate(j)$，当前日期为 $Nowdate$，$Bo(j)$ 为图书 j 自入库以来 $Nowdate-Indate(j)$ 时间段内被外借的总次数。图书 j 基于时间段的平均被外借次数 $BoT(j)$ 为：

$$BoT(j)=\frac{Bo(j)}{Nowdate-Indate(j)} \tag{4}$$

设图书 j 的出版日期为 $Pdate(j)$。若当前日期与图书出版日期之差比较小，则说明该书比较新。本文定义当前日期与出版日期之差的倒数 $Newb(j)$ 衡量图书为新书的程度：

$$Newb(j)=\frac{1}{Nowdate-Pdate(j)} \tag{5}$$

$Newb(j)$ 的值越大，说明该书越新。

图书 j 受大众读者欢迎($Welcome$，记为Wel)的程度，可通过式(6)表示：

$$Wel(j)=\alpha\times Av_{Interval}(j)+\beta\times BoT(j)+(1-\alpha-\beta)\times Newb(j) \tag{6}$$

其中，α 和 β 分别为调整的系数，$\alpha+\beta=1$。

2.2　读者特征

不同特征的读者，借阅的图书通常不一样。从读者的历史借阅记录，可以看出读者借阅哪类图书较多。未来一段时间内，读者很可能再借阅与历史记录相似的图书。因专业学习的需要，读者借阅与专业相关的图书可能性较大。不同年级的读者，借阅的图书也有差别。对于与专业相关不大的图书，女生比男生更倾向于借阅情感方面的图书，而男生倾向于球类、游戏方面的图书。

2.2.1　专业分类树

与中图分类树类似，根据全国普通高等学校本科专业目录[10]，可构建如图 2 所示的专业分类树。

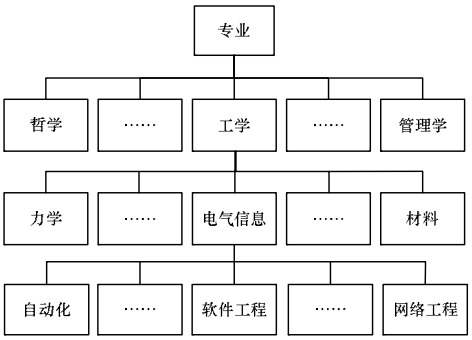

图 2　专业分类树

2.2.2　读者特征向量的相似性

在专业分类树中，具有相似专业的归为一大类(如：将与工学相关的专业归到工学这一大类)。将具有较多相似的专业归为同一大类下的子类(如：将自动化、软件工程、网络工程归到工学大类下的电气信息子类)。相似性最强的专业，其处于专业分类树的最下层。

同理，由图 2 的专业分类树，可得到目标读者 r 与 r' 基于专业特征向量的相似性 $simp(r,r')$ ：

$$simp(r,r')=\begin{cases}\dfrac{Layerp(r,r')-1}{Layerp(all)} & p(r)\neq p(r')\\ 1 & p(r)=p(r')\end{cases} \tag{7}$$

其中，$Layerp(r,r')$ 为目标读者 r 的专业 $p(r)$ 和读者 r' 的专业 $p(r')$ 在专业分类树中最近的共同父结点所在的层；$Layerp(all)$ 为专业分类树的总层数。

设目标读者 r 和 r' 的年级分别为 $g(r)$ 和 $g(r')$，性别分别为 $s(r)$ 和 $s(r')$，则目标读者 r 和 r' 基于年级特征向量的相似性 $simg(r,r')$ 为：

$$simg(r,r')=\begin{cases}1 & g(r)=g(r')\\ 0 & g(r)\neq g(r')\end{cases} \tag{8}$$

则目标读者 r 和 r' 基于性别特征向量的相似性 $sims(r,r')$ 为：

$$sims(r,r')=\begin{cases}1 & s(r)=s(r')\\ 0 & s(r)\neq s(r')\end{cases} \tag{9}$$

则目标读者 r 和 r' 基于专业、年级、性别综合特征向量的相似性 $simpgs(r,r')$ 为：

$$simpgs(r,r')=\chi\times simp(r,r')+\delta\times simg(r,r')+(1-\chi-\delta)\times sims(r,r') \tag{10}$$

其中，χ 和 δ 分别为调整的系数，$\chi+\delta=1$。

2.2.3 读者兴趣度

对同一位读者，若某本图书是他比较喜欢的，通常他看了第 1 轮还想看第 2 轮，即从外借到归还这本图书所需的时间间隔比较长。

假设图书 i 的页数为 $Page(i)$，目标读者 r 外借 i 的日期 $Bo_{date}(r,i)$ 与归还的日期 $Re_{date}(r,i)$ 基于页数的平均时间间隔 $Re_{Interal}(r,i)$ 为：

$$Re_{Interal}(r,i)=\frac{Re_{date}(r,i)-Bo_{date}(r,i)}{Page(i)} \tag{11}$$

设 $records(rb)$ 为目标读者 r 外借图书的总记录数。在目标读者 r 借阅的图书集合 $S(rb)$ 中，若很多图书与目标读者 r 未借图书 j 的相似性比较高，并且图书 j 受大众读者欢迎的程度比较高，意味读者更倾向于借阅图书 j。读者兴趣度可用图书 j 与目标读者 r 已外借的图书集合 $S(rb)$ 的综合相似性程度 $simb(j,S(rb))$ 表示为：

$$simb(j,S(rb))=\frac{\sum_{i\in S(rb)} simb(j,i)\times Re_{Interal}(r,i)\times Wel(j)}{records(rb)} \tag{12}$$

其中，$simb(j,i)$ 由式(1)得到；$Re_{Interal}(r,i)$ 由式(11)得到；$Wel(j)$ 由式(6)得到。

2.3 借阅预测

2.3.1 基于图书的借阅预测

设由式(1)产生图书 j 的邻居集合为 N_j，$\overline{Re_{Interal}(j)}$ 为图书 j 平均每次被外借与归还的时间间隔，$Re_{Interal}(r,i)$ 为目标读者 r 外借与归还图书 $i\in N_j$ 的时间间隔，$\overline{Re_{Interal}(i)}$ 为图书 $i\in N_j$ 平均每次被外借与归还的时间间隔，基于图书的预测目标读者 r 对未借阅图书 j 的外借与归还时间间隔 $P_{r,j}$ 为：

$$P_{book}(r,j)=\overline{Re_{Interal}(j)}+\frac{\sum_{i\in NN_j} sim(j,i)\times(Re_{Interal}(r,i)-\overline{Re_{Interal}(i)})\times simb(j,S(rb))}{\sum_{i\in NN_j}(|sim(j,i)|)} \tag{13}$$

其中，$sim(j,i)$ 由式(1)得到；$simpgs(r,r')$ 由式(10)得到；$simb(j,S(rb))$ 由式(12)得到。

2.3.2 基于读者的借阅预测

设由式(10)产生目标读者 r 的邻居集合为 N_r，$\overline{Re_{Interal}(r)}$ 为目标读者 r 外借与归还每本图书的平均时间间隔，$Re_{Interal}(r',j)$ 为 $r'\in N_r$ 外借与归还图书 j 的时间间隔，$\overline{Re_{Interal}(r')}$ 为 $r'\in N_r$ 外借与归还每本图书的平均时间间隔，基于读者预测目标读者 r 对未借阅图书 j 的外借与归还时间间隔 $P_{reader}(r,j)$ 为：

$$P_{reader}(r,j)=\overline{Re_{Interal}(r)}+\frac{\sum_{r'\in N_r} simpgs(r,r')\times(Re_{Interal}(r',j)-\overline{Re_{Interal}(r')})\times simb(j,S(rb))}{\sum_{r'\in N} simpgs(r,r')} \tag{14}$$

其中，$simpgs(r,r')$ 由式(10)得到。

2.3.3 基于图书和读者的借阅预测

根据基于图书的借阅预测和基于读者的借阅预测，基于综合的预测目标读者 r 对未借阅图书 j 的外借与归还时间间隔 $P(r,j)$：

$$P(r,j)=\lambda\times P_{\text{read}}(r,j)+(1-\lambda)\times P_{\text{book}}(r,j) \tag{15}$$

其中，$\lambda\in[0,1]$ 为权重因子。

2.4 推荐算法

输入 每本图书基于分类号 C、页数 $Page$、出版日期 $Pdate$、入库日期 $Indate$ 的特征向量及外借日期 Bo_{date}、归还日期 Re_{date}，每位读者专业 P、年级 $Grade$、性别 Sex 的特征向量，每位读者已经外借的图书记录集合 S

输出 推荐度排在最前面的 N 本图书作为目标读者 r 的推荐集 $top-N$

Step1 根据式(1)，计算目标读者 r 未外借任一图书 j 与已外借图书 i 的相似性 $simb(j,i)$。

Step2 根据式(6)或根据式(2)~式(5)，计算图书 j 受大众读者欢迎的程度 $Wel(j)$。

Step3 根据式(10)或根据式(7)~式(9)，计算目标读者 r 和 r' 基于专业、年级、性别综合特征向量的相似性 $simpgs(r,r')$。

Step4 根据式(12)，计算图书 j 与目标读者 r 已外借的图书集合 $S(rb)$ 的综合相似性 $simb(j,S(rb))$。

Step5 根据式(15)，得到目标读者 r 对未外借图书 j 的综合借阅预测 $P(r,j)$。

Step6 根据 $P(r,j)$ 的大小进行排序，获得目标读者 r 的推荐集 $top-N$。

3 实验结果与分析

3.1 实验数据集及度量

本文采用的实验数据来自广西民族师范学院图书馆的借阅记录。以一个学期的借阅记录作为训练集，后一个学期的借阅记录作为测试集。

每条借阅记录含有读者借阅证号(ID)、图书索书号、出版日期、外借日期、归还日期字段。并且从 OPAC 集成系统中获取每本图书的页数、入库日期。

对每位读者 r，按照他在训练集中的借阅记录，计算他的 N 本图书推荐集 $top-N$。若读者 r 的 N 本图书推荐集 $top-N$ 中的某本图书，出现在该读者的外借记录测试集里面，说明提出的算法生成了一本正确的图书推荐。采用 $Precision$ 作为图书推荐的精确度：

$$Precision=\frac{Hits}{N} \tag{16}$$

其中，$Hits$ 为推荐算法正确生成的推荐数目；N 为推荐算法总的推荐数目；$Precision$ 越大，说明算法的精确度越高。

3.2 结果及分析

实验取图书和读者的推荐数都为 20，在式(6)中 α 和 β 分别取 0.2、0.3，在式(10)中 χ 和 δ 分别取 0.2、0.3。

实验 1 图书和读者的最近邻数都分别取 10、20、30、40、50，在式(15)中 λ 分别取 0.2、0.3、0.4，推荐精确度实验结果对比如图 3 所示。

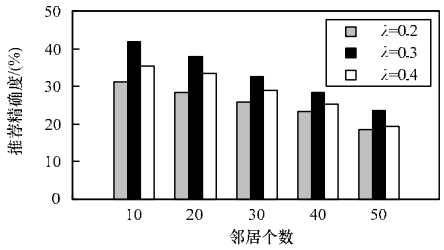

图 3　不同 λ 对推荐精确度的影响

从图 3 可以看出：(1)随着邻居数的增加，推荐精确度会减小。(2) λ 取不同的值，即基于读者的借阅预测和基于图书的借阅预测权重不一样时，推荐效果不一样。(3)当 λ =0.3 时，推荐精确度达到最大，即算法推荐质量最好。

实验 2 取 λ =0.3，专业数分别为 5、10、15、20、25，观察推荐精确度的变化如图 4 所示。

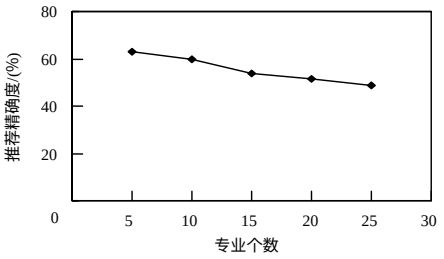

图 4　不同专业个数对推荐精确度的影响

从图 4 可以看出，随着专业个数的增长，推荐精确度略有减少，但总体的推荐质量比较好。

实验 3 本文算法(当 λ =0.3 时)与 PBRS[2]、CBDR[4]算法的推荐精确度比较如图 5 所示。

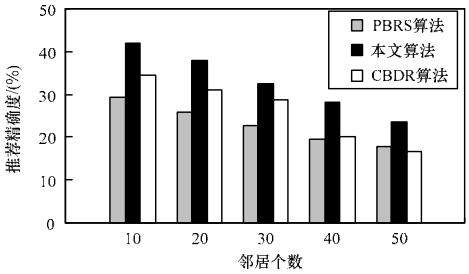

图 5　本文算法与其他算法的推荐精确度比较

常见的图书推荐算法只是从读者的某些兴趣或图书的个别特征来向读者做推荐，没有充分考虑读者及图书的多个综合特征，更没有对多个特征进行综合加权后再向读者推荐，因此，推荐质量不高。从图 5 可以看出，本文引入图书及读者的多个特征并做综合加权后提出的算法与其他算法相比，在不同的邻居数下，推荐精确度有了较大的提高，而且优势很明显。

4　结束语

本文提出一种综合图书和读者多特征的个性化图书推荐算法。综合考虑了图书分类号、页数、被外借与归还的时间间隔、平均被外借的次数、出版日期特征，及读者专业、年级、性别、借阅记录特征。基于图书和读者的加权预测产生推荐。实验结果表明，提出的算法预测推荐精确度优于常见的算法。下一步将研究受复本限制下的纸质图书推荐、新读者及新图书的推荐等问题。

参考文献

[1] 曾庆辉，邱玉辉. 一种基于协作过滤的电子图书推荐系统[J]. 计算机科学, 2005, 32(6): 147-150.

[2] Kuroiwa T, Bhalla S. Dynamic Personalization for Book Recommendation System Using Web Services and Virtual Library Enhancements[C]//Proc. of the 7th IEEE International Conference on Computer and Information Technology. Washington D. C., USA: IEEE Press, 2007: 212-217.

[3] 马　炎. 一种自适应的协作过滤图书推荐系统研究[J]. 情报杂志, 2008, 27(5): 105-106, 109.

[4] 武建伟，俞晓红，陈文清. 基于密度的动态协同过滤图书推荐算法[J]. 计算机应用研究, 2010, 27(8): 3014-3015.

[5] 丁　雪. 基于数据挖掘的图书智能推荐系统研究[J]. 信息系统, 2010, 33(5): 107-110.

[6] 王　茜，王均波. 一种改进的协同过滤推荐算法[J]. 计算机科学, 2010, 37(6): 226-243.

[7] Blattner M. B-rank: A Top N Recommendation Algorithm[EB/OL]. (2009-11-23). http://dblp.uni-trier.de/journals/corr/corr0908.html#abs-0908-2741.

[8] Ghazanfar M A, Prugel-Bennett A. Building Switching Hybrid Recommender System Using Machine Learning Classifiers and Collaborative Filtering[J]. International Journal of Computer Science, 2010, 37(3): 272-287.

[9] 国家图书馆《中国图书馆分类书》编辑委员会. 中国图书馆分类法[M]. 5 版. 北京: 北京图书馆出版社, 2010.

[10] 中国教育在线. 全国普通高等学校本科专业目录[EB/OL]. (2011-07-01). http://www.eol.cn/html/g/benkezy.shtml#0806/2011-7-1.

编辑　陆燕菲

(上接第 29 页)

[5] Nurmi D, Wolski R, Grzegorczyk C, et a1. The Eucalyptus Open-source Cloud-computing System[C]//Proc. of the 9th IEEE/ACM International Symposium on Cluster Computing and the Grid. [S. l.]: IEEE Computer Society, 2009.

[6] Armbrust M, Fox A, Griffith R, et al. Abore the Clouds: Aberkeley View of Cloud Computing[R]. Berkeley, USA: University of California, Berkeley: Technical Report: UCB/EECS-2009-28, 2009.

[7] Yan Baoqiang, Rhodes P J. Toward Automatic Parallelization of Spatial Computation for Computing Clusters[C]//Proc. of the 17th International Symposium on High Performance Distributed Computing. New York, USA: ACM Press, 2008.

[8] Nagarajan A B, Mueller F, Engelmann C, et al. Proactive Fault Tolerance for HPC with Xen Virtualization[C]//Proc. of the 21st Annual International Conference on Supercomputing. New York, USA: ACM Press, 2007.

[9] Pinar A, Hendrickson B. Exploiting Flexibly Assignable Work to Improve Load Balance[C]//Proc. of the 14th Annual ACM Symposium on Parallel Algorithms. New York, USA: ACM Press, 2002.

[10] 刘　怡，张　勘. 基于负载均衡和经验值的工作流任务分配策略[J]. 计算机工程, 2009, 35(21): 57-59.

[11] 陈国良. 遗传算法及其应用[M]. 北京: 人民邮电出版社, 1996.

[12] Liang Yanchun, Ge Hongwei, Zhou Chunguang. Solving Traveling Salesman Problem by Genetic Algorithms[J]. Progress in Natural Science, 2003, 13(2): 135-142.

编辑　顾姣健