基于遗传算法的智能商品搜索策略的研究

康晓凤,邵晓根

(徐州工程学院信电学院,江苏 徐州 221008)

[**摘要**] 电子商务购物系统为我们的日常生活带来了极大的便利,但是,随着现有电子商务购物系统中商务信息的急剧增加,导致用户搜索耗时太长,影响了交易的顺利进行. 为解决这种问题,提出了基于遗传算法的智能搜索策略. 首先根据用户输入的初始搜索字段,利用实数编码构造进化个体. 然后提出了基于用户行为的个体适应函数值的评价模型,辅助用户尽快搜索到满意商品. 最后基于 Java 平台开发了智能搜索引擎,通过与传统搜索引擎在搜索耗时和成功率方面的比较验证了该方法的有效性.

[**关键词**] 搜索引擎,智能,遗传算法,自然编码

[**中图分类号**]TP20 [**文献标志码**]A [**文章编号**]1001-4616(2014)04-0126-05

Study of Intelligent Goods Search Strategies Based on Genetic Algorithm

Kang Xiaofeng, Shao Xiaogen

(Department of Information and Electrical Engineering, Xuzhou Institute of Technology, Xuzhou 221008, China)

Abstract: The e-commerce shopping systems have greatly offered benefits to our daily life, however, the systems developed now often make the user spend a lot of time finding the satisfactory goods in e-commerce due to the increasing of commerce information. An Intelligent search strategies based on Genetic Algorithm is proposed to overcome the above shortages. Firstly, a natural coding was designed to encode the goods according to their key words and the ones entered by the user. Then, For effectively comparing the satisfied degree of the user on all displayed goods, a fitness function was built based on the goods the user had evaluated. Lastly, a personalized search strategy with Java was developed and compared with the traditional Search Strategy, and the results show that our algorithm is obviously advanced in saving user time and improving trade success.

Key words: search engine, intelligence, genetic algorithm, natural coding

当前电子商务以其方便、快捷、高效等特点,迅速得到广泛应用,并逐步演化成一种全球范围内分布的动态的商品信息资源库,也成为广大消费者购买或者浏览商品的主要形式之一. 目前的电子商务系统的运行过程可以分为 3 个部分:一是基于用户需求,如输入的关键词信息,进行相关信息的匹配;二是按照某种准则对相关的商品进行排序;最后是将所有相关的商品信息按照上述排序的值提交给用户,由其自由选择. 当用户对待购商品如商品品牌名称以及相应的商品编码足够了解时,系统可以快捷地为其提供相关信息. 但是,更多情况下,用户无法提供详细商品信息,只能提供非常简单的关键词,此时,用户要搜索到其满意的商品往往需要耗费很多时间,有时甚至在耗费了大量精力后,仍无法找到其满意的商品信息.

若能在用户提供简单关键词的情况下,将用户浏览信息所隐含的知识反馈给电子商务系统,然后基于该反馈信息,利用智能搜索算法,不断地进行新的匹配和排序,逐渐缩小用户的搜索范围,引导系统尽可能地为用户呈现其感兴趣的商品,则可大大改进现有电子商务系统在信息搜索方面存在的不足. 交互式遗传算法正是一种基于用户认知和偏好评价方案的智能搜索算法,即通过人机交互以及进化搜索过程,帮助用户找到满意解. 因此,可将该算法应用于电子商务系统中,而目前没有见到相关的研究.

收稿日期:2014-08-16.

基金项目:科技部技术创新项目(13C262132018660)、江苏省科技计划项目(BC2010058)、徐州工程学院 2012 科研项目(XKY2012308).

通讯联系人:康晓凤,讲师,研究方向:智能计算和信息安全. E-mail:kxfeng07@163.com

1 基于用户兴趣建模的商品智能搜索

目前用户搜索个人满意商品的过程实质是一个优化过程,现有的系统中,仅根据用户提供的关键词进行匹配和排序,没有进一步基于用户后续选择的信息进行辅助搜索[1]. 心理学研究表明,用户的行为反应了用户的认知、偏好和目的,可以通过分析用户的行为,隐式地感知用户的兴趣[2-4]. 根据用户的浏览行为,如对某类网页或者产品的点击率、对某网页的浏览时间、对某类商品的关注程度以及相关操作等,获取用户的兴趣模型,进而用于个性化推荐系统设计或者个性化检索等,从而在电子商务中拥有更多的用户群,成为许多网络企业的主要任务. 因此,基于用户兴趣发现的个性化推荐,近年来在电子商务中得到了广泛研究,而其核心即是用户兴趣发现和建模.

Choi,Huang,Su,邓春晖[5-8]等通过对浏览网站或者商品的特征信息,赋予一定的权值,通过记录用户浏览网页的行为,采用简单相加或者取大形式,确定相关的权重,获取用户兴趣,而没有将所建立的用户兴趣模型应用于进化算法中,通过与用户交互的融合,动态调整用户兴趣模型的同时,对搜索过程进行优化,从而尽快辅助用户找到满意的信息.

所谓基于遗传算法的智能搜索即将遗传算法应用于搜索中,在此过程中,动态调整适应度函数,重复进化操作,直到找到用户需要的商品信息. 网络购物中书籍因为知识含量高,购买前不需试用等特性,更适合于网络销售. 据 ACNielsen 的数据显示,网上购物者选择最受欢迎的网上商品中,书籍达到了 34% 的比例,所以本文用书作为案例来验证算法的有效性.

2 智能商品搜索约束模型

智能搜索的约束条件有很多可供选择的属性,约束条件可由属性值计算而得[9-10]. 为了提高搜索结果的用户满意度和搜索的效率,用户在搜索之前可以指定搜索的条件,例如名称、价格、出版社、使用对象、学科专业、销售成功率和用户的偏好等,使用对象主要是指这本书主要的使用人群,例如高职高专、大学本科、研究生等. 学科专业就是这本书的专业属性,例如计算机科学与技术类、电子信息类、自动化控制类、土木工程类等. 销售排行主要反映了这本书使用人群的多少,比如有些人认为买得人多了这本书肯定就好. 假设我们把这本书的浏览量记为 F,购买成功率记为 H,则这本书的销售成功率 Q 就可以表示为 H/F. 我们在建库的时候每本书都有销售量和浏览数量这个属性,每销售或浏览了一本书则这个属性的计数就加 1. 用户的喜好是指用户比较偏爱某个人写的书或者某个出版社的书,这些信息一般在申请会员时进行填写. 定义的目标图书的指标体系 $I=\{$名称、价格、出版社、使用对象、学科专业、销售成功率和用户的偏好$\}$. 假设目标书籍有 m 个,每本书有 n 个属性,则最后的搜索结果可以用 $m\times n$ 的目标矩阵 $\boldsymbol{R}$ 来表示,行表示目标书籍,列表示书籍的属性. $\boldsymbol{R}$ 中每一列向量表示一个分目标,其中 r_{ij} 为第 i 本书的第 j 个属性值,则

$$\boldsymbol{R}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix}.$$

3 智能商品搜索算法描述

3.1 算法模型

(1)染色体编码及初始种群

目前染色体主要的编码方式为二进制编码和实数编码. 二进制编码必须要进行解码,这样会增加计算量,十进制编码不需要解码操作,计算量小,效率高,故本文采用实数编码. 染色体基因直接采用某本书的属性表示,染色体代表某本书,书的染色体编码为$(G_1,G_2,\cdots,G_m)$,其中 $G_i(i=1,2,\cdots,m,m$ 为某本书的属性). 基因按学科专业有序排列,因此同学科的书总是放在同一区间内. 最初随机生成的 n 本书作为初始群体,种群规模根据需要给出. 群体中的个体由从图书中按学科专业随机选择 m 本书产生.

(2)适应度函数

由于初始化种群时,染色体已满足了学科专业的要求,故目标函数中只考虑图书的名称、价格和出版社即可,目标函数可以如下构建.

设数据库中的某条图书的记录为 r_x,用户输入特征所对应的目标图书 r_y,根据上面的矩阵可知图书所对应的属性为 r_{ij},则一本图书的属性评价函数可用下式表示:

$$f_i=\frac{r_{xj}-r_{yj}}{r_{yj}}, \tag{1}$$

式中:f_i 越小代表越能满足用户的要求. 对于一本书的其他属性的评价函数设用 $f_1,f_2,\cdots,f_n$ 来表示. 假设找到的目标图书与用户要求之间的总误差为 f. 由于上述各项约束的重要程度不同,因此一本书总误差 f 设计为各项误差 f_i 的加权和,$f=\sum_{i=1}^{n} w_i f_i$,其中 w_i 表示第 i 个指标的权值,反映约束条件的强弱,$\sum_{i=1}^{n} w_i=1$. 搜索算法的目标就是要使搜索的总误差 f 尽可能小.

(3)选择操作

选择操作指从群体中选择染色体遗传到新一代种群的过程,根据式(2)计算个体选择概率. 然后据 s 值进行从大到小的顺序进行排序.

$$s=\frac{1}{w_1\left(\frac{r_{x1}-a}{a}\right)^2+w_2\left(\frac{r_{x2}-b}{b}\right)^2}, \tag{2}$$

其中:w_1、w_2 为权值系数,且满足 $w_1>0$、$w_2>0$,x 为变量,对应着书的编号是整数,r_{x1} 为书名,a 为关键词的值,r_{x2} 为书的价格,b 为用户输入的价格. w_1、w_2 为权值,如果用户比较看重这本书的某种属性,就可以把这种属性的权值提高[11-13].

(4)交叉操作

在交叉过程中选定两个染色体的基因座依照交叉概率 P_c 进行交换,使算法探索新的领域. 算法采用分组单点交叉策略. 随机产生交叉点和1个(0,1)内的随机数 R_c,如果 $R_c<P_c$,交换两个个体的基因座,否则不交换.

(5)变异操作

变异操作模拟生物由于各种偶然因素引起的基因突变,这种基因突变依照变异概率 P_m 进行. 对每个个体产生(0,1)间的一个随机数 R_m,若 $R_m<P_m$,则变异染色体中的基因座,否则不变异. 在基因变异时,也要满足使用对象、学科专业、书名和价格等关键因素的约束. 若不满足,需要重新产生变异位置和 R_m.

3.2 算法描述

遗传算法智能搜索的主要步骤描述如下:

(1)参数初始化,例如初始种群规模、最大迭代次数、交差参数、变异概率等,生成初始种群.

(2)计算种群个体适应函数值.

(3)是否达到遗传最大迭代次数,如果是转向(5),否则进行遗传操作产生中间种群.

(4)判断中间种群中是否有适应函数值相同的个体,如果有就根据销售成功率、喜欢的出版社和出版日期等信息进行二次筛选.

(5)算法结束,输出搜索的结果.

4 智能商品搜索算法的应用

4.1 参数设置

系统中的图书信息包括图书名称、类别、价格、适用对象、出版社、关键词和销售率等属性. 目标图书属性应该满足的约束条件如下:

约束1:使用对象:高职高专、大学本科和研究生.

约束2:学科专业:计算机、电子信心、数学、经济管理、人文和外语.

约束3:书名:根据用户的输入的内容,计算机关键词的值. 对于数据库中每一条记录,我们都给出其

相应的关键词,并给这个关键词赋值. 比如网络原理,这条记录的关键词就可以是网络. C++计算机网络编程技术,这条记录的关键词就是 3 个分别是 C++、计算机和网络. 对于符合多个关键词的情况我们采用,最后的关键词=1/n. 关键词 1 的值+2/n. 关键词 2 的值+……的方法来求每条记录的关键词. 我们在建库的时候专门创建一张表存储关键词和关键词的值.

约束 4:价格:用户输入的价格.

约束 5:根据书名和价格来计算数据库中的每条记录的适应值.

我们进行选择的时候给数据库中的每一条记录一个编号,根据用户输入的信息计算适应值,然后根据适应函数值 按照从大到小的顺序输出满足条件的记录,让用户在这些记录中找出比较满意的.

例如查找网络原理这本书,输入的价格为 30 元,假设关键词网络的值为 25. 如数据库中有一条记录为计算机网络原理,29 元. 适应函数值公式应该为:

$$s=\frac{1}{w_1\left(\frac{x(1)-25}{25}\right)^2+w_2\left(\frac{29-30}{30}\right)^2}. \tag{3}$$

这里 x_1 表示关键词计算机和网络的平均值. 如果经过计算适应函数值相同的记录,则根据销售成功率、喜欢的出版社和出版日期再进行第二次筛选. 销量就是销量高排在前面,因为销量在一定程度上翻译了用户对这本书的认可度. 喜欢的出版社是用户申请时,用户填入的信息,因为有的用户对某个出版社比较认同,或者是因为学科的关系,某个出版社的书籍特别适合于某个学科. 出版日期一般就是选择比较新的,因为有的高校在选择教材时规定是近 3 年出版的,特别是对于一些技术更新特别快的技术例如计算机编程技、网络技术等,更要选择新出版的,否则书籍中讲述的技术与新技术之间就产生了断代.

4.2 交互界面

交互界面采用 Java 的 JGAP 软件开发包设计,用户首先进行登录,然后进行信息查询,查询页面如图 1 所示,主要包括类别、使用对象、书名和市场价格 4 个信息输入窗口;3 个按钮分别是提交、重置和刷新. 当用户输入主要的信息之后,点击提交就把这些信息提交给相应的程序去处理,如果用户对新生成的信息不满意,可以点击刷新按钮,这时就把上次处理的结果重新反馈给用户,重复适应函数值计算和遗传操作,直至用户找到满意信息;如果用户想清空输入的信息,就可以点击重置按钮.

图 1 智能搜索界面

Fig. 1 Intelligent search interface

4.3 性能比较

为了说明智能搜索的有效性,首先考虑在输入同样关键字和固定搜索时间的条件下,分别利用传统方法(关键字匹配技术)和本文方法,统计用户找到满意的解的次数和成功率. 本系统中设定搜索关键字为:计算机、大学本科、计算机网络和 30. 测试时系统独立运行 80 次,限时 1 min,比较这两种方法在 80 次搜索中,找到满意的解的次数,结果如表 1 所示. 可以看出在有限的时间内,采用本文算法找到满意的解的成功率大于传统方法.

表 1 搜索成功率

Table 1 The successful rate of the search

比较参数	传统方法	本文方法
用户满意的解次数	52	68
成功率	65%	85%

然后不固定搜索时间,统计用户分别利用传统方法和本文方法找到满意的解所花费的时间. 本系统中设定同样的搜索关键字,独立运行 80 次,用户需要的时间如图 2 所示. 从图 2 可以看出,智能搜索引擎在查找耗时方面明显优于传统搜索引擎.

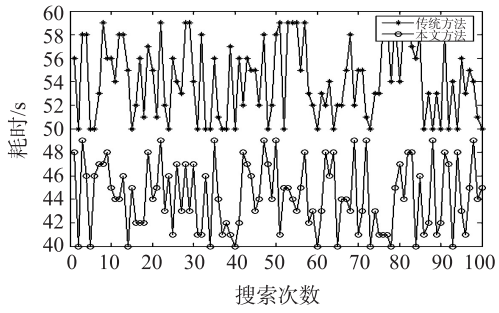

图 2 搜索耗时比较图

Fig. 2 A comparison of search time

5 结论

为了提高电子商务中用户商品搜索的效率,针对其本质为优化问题的特性,提出基于遗传算法的智能搜索策略.根据用户输入的初始搜索字段,采用自然编码表示数据库中相关的关键词,作为进化个体,然后基于用户选择的潜在感兴趣的商品信息,估计进化个体适应值,并利用遗传算法辅助用户尽快搜索到满意商品.开发了一个智能搜索引擎,与传统的搜索引擎相比减小了用户耗时、提高了搜索成功率.将该算法框架应用于更多实际的电子商务平台中将是需进一步研究的内容.

[参考文献]

[1] Tokui N, Iba H. Music composition with interactive evolutionary computation[C]//Proceedings of the 3rd International Conference on Generative Art, Milan, 2000:215-226.

[2] Nia Xingliang, Lub Yao, Quanc Xiaojun, et al. User interest modeling and its application for question recommendation in user-interactive question answering systems[J]. Information Processing & Management, 2012, 3:218-233.

[3] Issa, Taroub. How Web applications complement search engines[C]//2013 Palestinian International Conference on Information and Communication Technology, Singapore, 2013:99-106.

[4] Shen Wei, Wang Jianyong, Luo Ping, et al. Linking named entities in Tweets with knowledge base via user interest modeling [C]//KDD'13 Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, 2013:68-76.

[5] Choi D H, Ahn B S. Eliciting customer preferences for products from navigation behavior on the web:a multicriteria decision approach with implicit feedback[J]. IEEE Transactions on System, Man, and Cybernatics—Part A: Systems and Humans, 2009, 39(4):880-889.

[6] Huang C Y, Yang Y L, Tzeng G H, et al. 4G mobile phone consumer preference predictions by using the rough set theory and flow graphs[J]. Proceedings of Technology Management for Global Economic Growth, 2010, 1:10.

[7] Su J W, Wang B W, Hsiao C Y, et al. Personalized rough-set-based recommendation by integrating multiple contents and collaborative information[J]. Information Sciences, 2010, 180:113-131.

[8] 伊春晖,邓伟.基于用户浏览行为分析的用户兴趣获取[J].计算机技术与发展,2008,18(5):37-39.

[9] 蒋在帆,王斌.基于用户行为分析的个人信息检索研究[J].中文信息学报,2011(1):9-14.

[10] 周艳聪,刘艳柳.遗传模拟退火智能组卷策略研究[J].计算机工程与设计,2011(3):1 066-1 069.

[11] 肖理庆,徐晓菊.改进遗传算法智能组卷研究[J].计算机工程与设计,2012,10(5):3 970-3 974.

[12] 巩敦卫,郝国生,严玉若.交互式遗传算法基于用户认知不确定性的定向变异[J].控制与决策,2010(1):74-78.

[13] Sun X Y, Gong D W, Ma X P. Directed fuzzy graph based surrogate model assisted interactive genetic algorithms with uncertain individual's fitness[J]. Proceedings of IEEE Congress on Evolutionary Computation, Trondheim, 2009:2 395-2 402.

[责任编辑:陆炳新]