改进的遗传算法在 Web 使用挖掘中的应用

雷 亮[1],李善君[2],彭 军[1]
LEI Liang[1],LI Shan-jun[2],PENG Jun[1]

1.重庆科技学院 电子信息工程学院,重庆 401331
2.重庆市科协技术信息中心,重庆 401147
1.School of Electronic Information Engineering,Chongqing University of Science and Technology,Chongqing 400050,China
2.Information Center of Chongqing Association for Science and Technology,Chongqing 401174,China
E-mail:cqlei.l@163.com

LEI Liang,LI Shan-jun,PENG Jun.Application of Web usage mining based on improved genetic algorithm.Computer Engineering and Applications,2009,45(8):135-137.

Abstract: Web usage mining is a hot research direction of Web data mining.In this paper,an improved genetic algorithm is proposed to overcome the shortage of early convergence and stagnation in the traditional genetic algorithm,which is based on unconvertible rate of crossover operator and mutation operator.Moreover,the user page interest measure threshold is introduced into the association rules mining.Lastly,the improved genetic algorithm is successfully applied to a commerce Web server log mining and the experiment results indicate that the proposed algorithm is an effective method to avoid early convergence.

Key words: Web data mining;Web usage mining;genetic algorithm;interest measure

摘 要:Web 使用挖掘是近年来 Web 数据挖掘中的研究热点。针对传统遗传算法在提取关联规则问题时常采用固定染色体交叉概率和染色体变异概率,容易出现早熟、收敛速度较慢的问题,提出了改进的遗传算法,并在关联规则的提取中增加了用户页面兴趣度这一阈值,成功地运用到某商业网站服务器日志挖掘。实验证明,这种改进的遗传算法能够有效避免早熟收敛现象,是一种有效的方法。

关键词:Web 数据挖掘;Web 使用挖掘;遗传算法;兴趣度
DOI:10.3778/j.issn.1002-8331.2009.08.041 **文章编号**:1002-8331(2009)08-0135-03 **文献标识码**:A **中图分类号**:TP181

1 引言

随着 Internet 的快速发展,电子商务已经成为企业从事商业活动的一种重要方式。相对于传统商务模式,电子商务领域的竞争更加激烈,客户只需点击鼠标就可以在不同的商家之间转换、比较。如何了解到顾客尽可能多的爱好和价值取向,为顾客提供更优质的服务成为电子商务发展迫切要解决的问题。

电子商务网站的顾客在 Web 上的行为都会产生大量数据信息,不仅包括本次交易信息而且还有利用搜索引擎,以及在站点内进行浏览的相关数据。如何充分了解网络客户的个性化需求以及网络客户的浏览行为, 成为电子商务网站结构设计、内容设计和服务设计的前提条件。网络交易可以生成大量的交易记录,同时,客户的浏览行为也有意或无意地在网上留下了许多重要信息。通过数据挖掘技术,对所获得的各种 Web 信息(包括交易记录、Web Log、注册信息、cookies 信息、所访问站点的结构与页面信息等)进行 Web 使用挖掘,发现客户的访问模式,从而可以对客户进行分类、聚类、发现潜在的客户、改进站点的设计,方便客户的浏览和交易,并为客户提供个性化的服务,这使得 Web 使用挖掘成为目前 Web 数据挖掘的研究热点。

本文在介绍 Web 使用挖掘的基本知识后,提出了使用改进的遗传算法,从大量的 Web 服务器日志中提取关联规则。

2 Web 使用挖掘

Web 数据挖掘是将 Web 技术与数据挖掘技术结合起来,获取 Web 知识的过程。Web 数据挖掘一般的定义为:从与 www 相关的资源和行为中抽取感兴趣的、有用的模式和隐含信息。Web 数据挖掘可分为三类[1-3]:Web 内容挖掘(Web content mining)、Web 结构挖掘(Web structure mining)和 Web 使用挖掘(Web usage mining)。Web 内容挖掘是从文档内容或其描述中抽取知识的过程。Web 结构挖掘是从 Web 页的组织结构和超链接关系以及 Web 文档自身的结构信息(如 Title、Heading、Anchor 标记等)推导出 Web 内容以外的知识。Web 使用挖掘主要是从 Web 的访问记录(服务器日志)中抽取感兴趣的模式,其数据原有服务器的日志、用户注册数据、跟踪文件的数据记录、用户访问期间的事务、用户查询、书签数据和鼠标移动点击的信息。其主要特点是对用户信息数据进行抽取、转换、分析和其他模型化处理,从中提取辅助企业决策的关键性数据。

基金项目:重庆市科技攻关计划(the Key Technologies R&D Program of Chongqing,China under Grant No.CSTC2007AB2047)。
作者简介:雷亮(1973-),男,博士研究生,讲师,主要研究领域为图像处理、模式识别和数据库挖掘;彭军(1970-),男,博士,教授,主要研究领域为图像处理、混沌密码学研究等;李善君(1971-),男,工程师,主要研究领域为数据库处理、Web 应用程序开发。
收稿日期:2008-11-03 **修回日期**:2009-01-14

Web 使用挖掘的一般过程如图 1 所示,包括:(1)数据的预处理。这是用户访问信息最关键的阶段。数据预处理包括数据净化、用户识别、会话识别、页面过滤、事务识别、路径补充等过程。(2)模式发现。对数据预处理所形成的文件,利用数据挖掘的一些有效算法(本课题主要用关联规则)来发现隐藏的模式和规则。(3)模式分析。主要是对挖掘出来的模式、规则进行分析,找出用户感兴趣的模式,提供可视化的结果输出。

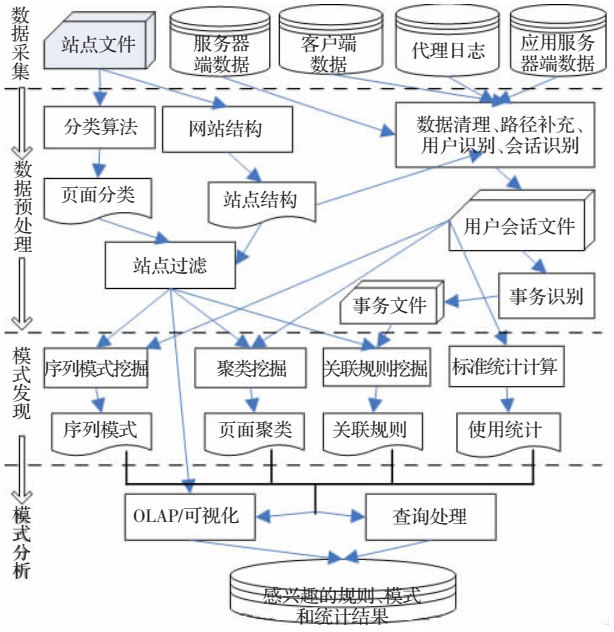

图 1 Web **使用挖掘的一般过程**

Web 使用挖掘应用的技术主要有路径分析、关联规则分析、序列模式分析、聚类分析、统计分析等。Web 使用记录挖掘可以挽留老客户发现潜在的新用户、改进电子商务网站的建设、制定产品营销策略、降低运营成本、提高站点点击率、增加个性化服务等。

3 关联规则的相关概念

挖掘关联规则是指在数据库中挖掘出具有如下形式的规则:由于某些事件的发生而引起另外一些事件的发生。它在决策支持系统、专家系统和智能信息系统等各个方面起着重要的作用。并且,随着数据库应用的普及,数据挖掘的应用越来越广,在近几年内备受人们的关注。在 Web 使用挖掘中,关联规则主要用于发现用户之间、页面之间以及用户浏览页面和网上行为之间存在的潜在关系[4-6]。例如,在超市的购买记录中,可以发现"购买尿布湿的男性顾客又买了啤酒的可能性是 80%"。在 Web 服务器日志中,可以发现"访问页面 P_1 和 P_3 的用户中有 35%的用户也访问了页面 P_5"。这样,可以调整尿布湿的价格来促进啤酒的销售;调整页面路径,为用户预取一些 Web 页面,加快用户获取页面的速度。

设 $I=\{i_1,i_2,\cdots,i_m\}$是 m 个不同项目的集合,即项目集,D 为事务数据库(或称事务集),其中每个事务 T 是一个项目子集($T\subseteq I$)。事务 T 包含项目集 X,表示为 $X\subseteq T$。

定义 1 关联规则是形如 $X\Rightarrow Y$ 的逻辑蕴含式,其中,$X\subset T$,$Y\subset T$,并且 $X\cap Y=\Phi$。X 称作前提,Y 称作结果。有两个因子与这条规则有关:可信度(Confidence)和支持度(Support)。

定义 2 可信度(Confidence)。如果事务数据库 D 里包含 X 的事务中有 c%的事务同时也包含 Y,那么称关联规则 $X\Rightarrow Y$的置信度为 c。简单地说,可信度就是指在出现了项目集 X 的事务 T 中,项目集 B 也同时出现的概率有多大。

定义 3 支持度(Support)。如果事务数据库 D 中有 s%的事务包含 $X\cup Y$,那么就称关联规则 $X\Rightarrow Y$ 的支持度为 s。支持度描述了 X 和 Y 这两个项目集的并集 C 在所有的事务中出现的概率有多大。

最著名的关联规则挖掘方法是 R.Agrawal 提出的 Apriori 算法。关联规则的发现都遵循两个步骤:(1)迭代识别所有的频繁项目集,要求频繁项目集的支持率不低于用户设定的最小支持度;(2)从频繁项目集中构造可信度不低于用户设定的最小置信度。

事实上,在 Web 使用挖掘中,仅考虑规则的支持度和可信度是不够的。在本文中,对关联规则的提取,将增加用户兴趣度这一阈值。用户兴趣度的计算办法:设 S_i 是一个页面上的一个会话,$S_i=\{[c_1,t_1,sb_1,f_1],[c_2,t_2,sb_2,f_2],[\cdots],[c_n,t_n,sb_n,f_n]\}$, 其中,$c_n$ 表示第 n 个页面,t_n 是第 n 个页面所花时间信息,sb_n 是第 n 个页面的发送字节数信息,f_n 是第 n 个页面被浏览的次数。页面 j 的兴趣度 p_j 可以按照公式(1)计算得到:

$$p_j=\frac{t_j*f_j}{sb_j}\quad (j=1,2,\cdots,n) \tag{1}$$

式中,n 表示页面数。

4 改进的遗传算法

提取关联规则的核心问题是发现最大项目集。发现最大项目集的过程其实就是全局的搜索过程,遗传算法是一种全局优化算法,因此它避免了搜索过程中的局部最优。将遗传算法用在规则的发现和提取方面能够发现真正有用的规则。

标准遗传算法是针对一个宏观的种群进行选择、交叉、变异三种操作,类似于人类进化过程,一群人随着时间的推移而不断地进化,具备越来越多的优良品质。然而,由于他们的生长、演化、环境和原始祖先的局限性,经过相当长时间后,将逐渐进化到某些特征相对优势的状态,当一个种群进化到这种状态,称之为平衡态,这个种群的特性就不会再有很大的变化。双种群遗传算法是一种并行遗传算法, 它使用多种群同时进化,并交换种群之间优秀个体所携带的遗传信息,以打破种群内的平衡态达到更高的平衡态,跳出局部最优。

传统的双种群遗传算法(AGA)在提取关联规则时,首先建立两个遗传算法群体,即种群 A 和种群 B,分别独立地进行自然选择、染色体交叉、染色体变异操作,且交叉概率、变异概率固定。当每一代运行结束以后,产生一个随机数 num,分别从 A、B 中选出最优染色体和 num 个染色体进行杂交,以打破平衡态。

但是,在实际应用传统的双种群遗传算法[7]的过程中,由于交叉概率、变异概率固定不变,容易出现过早收敛而仅得到局部最优解的现象,需要对算法进行改进,让交叉概率和变异概率能够自适应调节,使算法寻优速度加快而且不易陷入局部最优解。

文献[8-11]提出了一种改进的自适应遗传算法(IAGA),其交叉概率 P_c 和变异概率 P_m 分别如公式(2)和(3)所示。

$$P_c=\begin{cases}P_{c1}-\dfrac{(P_{c1}-P_{c2})(f'-f_{avg})}{f_{max}-f_{avg}}, & f'\geqslant f_{avg}\\ P_{c1}, & f'\leqslant f_{avg}\end{cases} \tag{2}$$

$$P_m=\begin{cases}P_{m1}-\dfrac{(P_{m1}-P_{m2})(f_{\max}-f)}{f_{\max}-f_{avg}}, f\geqslant f_{avg}\\ P_{m1}, f\leqslant f_{avg}\end{cases} \quad (3)$$

其中,$f_{\max}$ 代表群体中最大适应度;f_{avg} 代表每代群体的平均适应度;f'代表要交叉的两个个体中较大适应度;f 代表要变异个体的适应度。

但是,运用本算法,较差个体的变异能力较低,容易产生停滞现象。而精英保留策略虽然起到了保护和推广优秀个体的作用,但是其个体数目不宜过大,否则会使种群进化陷入停滞不前,造成局部收敛。

为此,本文对交叉概率 P_c 和变异概率 P_m 进行了改进,提出了一种新的自适应遗传算法(NAGA)。为了更好地描述这种算法,引入 N_1 和 N_2。

$$N_1=\frac{f_{\max}-\bar{f}}{f_{\max}-f_{\min}+\varepsilon} \quad (4)$$

$$N_2=\frac{\bar{f}-f_{\min}}{f_{\max}-f_{\min}+\varepsilon} \quad (5)$$

$$p_c=\begin{cases}0.9 & N_1(z)\leqslant N_1(z-1), z\in[1,gen]\\ 0.5 & N_1(z)>N_1(z-1), z\in[1,gen]\end{cases} \quad (6)$$

$$p_m=\begin{cases}0.3 & N_2(z)\leqslant N_2(z-1), z\in[1,gen]\\ 0.01 & N_2(z)>N_2(z-1), z\in[1,gen]\end{cases} \quad (7)$$

在式(4)和式(5)中 $f_{\max}$ 代表群体中最大适应度;$f_{\min}$ 代表群体中最小适应度;$\bar{f}$表示本代群体的平均适应度;ε 是一个无穷小正数,主要是为了防止分母为 0。

在式(6)和式(7)中,z 表示遗传的当前代,$z-1$ 表示上一代;gen 表示进化的总代数。

改进后的交叉概率和变异概率不但能够随适应度自动改变,而且使种群中最大适应度值的个体的交叉概率和变异概率不为零,这就相应地提高了种群中表现优良的个体的交叉概率和变异概率,使得它们不会处于一种近似停滞不前的状态,从而使算法跳出局部最优解。

5　实验结果与分析

Web 服务器日志一般包含[12]:用户请求页面的日期(date)、用户请求页面的具体时间(time)、客户端主机的 IP 地址或 DNS(cs-ip)、客户端的用户名(cs-username)、用户代理(User-Agent)、服务器 IP 地址(s-ip)、服务器端口(s-port)、用户的请求方法(cs-method)、用户的请求页面(cs-Uri-stem)、用户欲进行的查询(cs-Uri-query)、协议状态(cs-status)、服务器名(s-computername)、服务器发送的字节数(sc-byte)、客户收到的字节数(cs-byte)、完成浏览所花费的时间(time-taken)、服务器的操作系统(host)、协议版本(cs-version)、Cookie(Cookie)、用户浏览的上一页(Reference)等信息。经过数据的预处理后,得到 Web 使用挖掘任务有关的用户 IP 地址、用户 ID、请求访问的 URL 页面、页面字节数(sc-byte)和用户请求页面的具体时间等信息。

以某商业网站服务器日志为实验平台,对取得的一个月的 Web 服务器日志进行预处理,得到如表 1 所示的数据(为保护用户的个人隐私,隐去了用户 IP 地址,用用户 ID 代替)。

为便于关联规则的提取,采用上述公式(1),采用文献[13]的方法进行页面兴趣度的计算,同时过滤掉页面兴趣度小于或等于 0.1 的页面记录。同时,变换数据存储格式,使得一个用户的一次会话用一条记录表示,如表 2 所示。

表 1　经过预处理后的 Web 日志

用户 ID	请求访问的 URL 页面	页面字节数	请求页面的具体时间
User1	product/214/214390/	1 286.5	20.2
User1	product/87/87485/	1 998.5	71.3
User1	product/265/265045/	1 876.3	45.6
User1	product/49/49236/	1 598.6	56.2
User1	product/338/338458/	2 109.8	19.3
User1	product/297/297384/	2 104.2	120.3
User1	product/162/162485/	1 987.3	45.3
User2	product/297/297397/	2 203.4	54.4
User2	product/162/162500/	1 890.3	56.3
User2	product/297/297441/	1 847.6	34.2
User2	product/297/297384/	2 104.2	81.3
User2	product/265/265045/	1 876.3	76.5
User2	product/87/87485/	1 998.5	90.2
User2	product/338/338458/	2 109.8	34.2
User2	product/297/297396/	2 289.3	10.4
…	…	…	…

表 2　根据页面兴趣度过滤后的 Web 日志

用户 ID	请求访问的 URL 页面
User1	214390,87485,265045,49236,297384,162485
User2	297397,162500,297441,297384,265045,87485, 338458
…	…

运用上面介绍的改进的遗传算法,进行关联规则提取,得到这样一些强规则:40%的用户访问页面 214390 时也访问了 265045;35%的用户访问页面 214390 时也访问了 49236;29%的用户访问页面 162500 时也访问了页面 162485;36%的用户在访问页面 162485 时访问了页面 162500…。利用这些强关联规则,可以更好地组织站点的 Web 空间,减少用户过滤信息的负担,实行有效的推销策略,增加交叉销售量。

另外,为了比较新算法(NAGA 算法)的收敛性,选取一个简单的单峰值函数(De Jong 球函数)进行实验,此函数是一个简单的平方和函数,用遗传算法易于求解。将 NAGA 算法与传统双种群遗传算法(AGA)、改进的自适应遗传算法(IAGA)进行独立实验结果比较。为求得函数值 99.999 9,用这三种方法分别独立运行 500 次、1 000 次和 2 000 次,将收敛次数分别记录,如表 3 所示。

表 3　两种遗传算法执行结果对比

运行次数	500	1 000	2 000
AGA	482	968	1 946
IAGA	491	983	1 974
NAGA	499	998	1 998

从表 3 可以看出,NAGA 在收敛性上改善了现有的一些自适应遗传算法的性能,其计算效果比较理想,是一种有效的、稳定的、十分实用的算法。

6　结论

随着网络资源的日益丰富,Web 使用挖掘已经成为热点,通过挖掘可以更好地管理服务器资源、改进电子商务网站的建设、制定产品营销策略、降低运营成本、提高站点点击率、增加个性化服务。本文简略介绍了 Web 使用挖掘的一般过程和常

(下转 171 页)