

基于用户分层的电子商务智能信息推荐策略

何 波

(重庆理工大学 计算机科学与工程学院, 重庆 400050)

摘 要: 针对全体用户采用同一推荐方法的传统推荐策略已不能满足当前电子商务企业和用户的要求, 取而代之的是以用户为中心, 在充分挖掘用户价值的基础上进行有针对性的信息推荐. 将电子商务用户分层为新用户、高端用户和普通用户, 提出了基于用户分层的电子商务智能信息推荐策略. 针对用户的不同, 设计有针对性的推荐方法, 以实现企业与用户的双赢. 实验结果表明, 提出的推荐策略和推荐方法是有效的.

关键词: 用户分层; 协作过滤; 马尔可夫决策过程; 智能信息推荐

中图分类号: TP18

文献标识码: A

文章编号: 1000-7180(2011)12-0021-04

E-commerce Intelligent Information Recommendation Strategy Based on User Segmentation

HE Bo

(School of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400050, China)

Abstract: The traditional recommendation strategy that aimed at all users adopting the same recommendation method could not satisfy current E-commerce enterprises and users' needs. The advanced one was that centered in users and implemented information recommendation on the base of the fully mined users' value. This paper divided E-commerce users into new user, advanced user and ordinary user. It proposed the E-commerce intelligent recommendation strategy based on user segmentation. According to the difference of users, the paper designed different recommendation methods to guarantee enterprises and users' profit. The experimental results indicate that the proposed recommendation strategy and the designed recommendation methods are efficient.

Key words: user segmentation; collaborative filtering; MDP; intelligent information recommendation

1 引言

随着电子商务^[1-2]的迅猛发展, 如何有效地为电子商务用户提供信息推荐服务^[3]成为一个重要的研究课题. 传统的推荐策略是针对全体用户采用同一推荐方法, 没有考虑不同用户对企业利益的影响, 不能满足当前企业和用户的要求. 改进的措施是以用户为中心, 充分挖掘用户价值的不同, 进行有针对性的信息推荐, 即对用户进行分层的基础上进行信息推荐, 实现电子商务企业与用户的双赢. 论文将电子商务用户分层为新用户、高端用户和普通用户, 提出了基于用户分层的推荐策略. 针对新用户, 设计基

于用户事务模式聚类的信息推荐方法(IRUMC). 针对高端用户, 设计基于协作过滤的 Web 智能信息推荐方法(WIIRM). 针对普通用户, 采用基于马尔可夫决策过程的信息推荐方法.

2 Web 日志预处理

Web 日志的数据预处理是进行信息推荐的基础, 其主要步骤包括: 数据净化、用户识别、用户访问操作识别、路径完善和用户事务模式识别等. 数据预处理步骤的详细描述见文献[3]. 经过数据预处理后, 可获得用户事务模式, 并且生成每个用户的事务模式文件, 每一个用户事务模式文件包含了若干

收稿日期: 2011-03-01; **修回日期:** 2011-05-08

基金项目: 教育部人文社会科学研究项目基金(09yjc870032)

个用户事务模式. 由于要考虑用户访问页面的时间特性, 所以得到的是带有时长的用户事务模式. 带有时长的用户事务模式如定义 1 所示.

设 L 为用户访问操作集合, 每一条访问记录 $l \in L$ 包括: 用户主机地址 $l.ip$, 用户 ID 号 $l.uid$, 被访问 Web 页的 URL 地址 $l.url$, 访问时间戳 $l.time$ 和访问 Web 页的时长 $l.timelength$.

定义 1 带有时长的用户事务模式定义为三元组, 形式化表示如下:

$$t = \langle ip_i, uid_i, URL_i \rangle;$$

$$URL_i = \{ (l'_1.url, l'_1.time, l'_1.timelength), \dots, (l'_m.url, l'_m.time, l'_m.timelength) \}$$

其中 $l'_k \in L, ip_i = l'_k.ip, uid_i = l'_k.uid, 1 \leq k \leq m; l'_k.timelength = l'_{k+1}.time - l'_k.time, 1 \leq k \leq m-1$.

3 基于用户分层的电子商务智能信息推荐策略

3.1 用户分层^[4]思想

随着网络经济时代的到来, 针对全体用户采用同一推荐方法的传统推荐策略已不能满足当前企业和用户的要求. 正如 Pareto^[3] 原则所表示的内容, 企业大部分利润是由小部分用户所创造. 这就是著名的 80/20 法则, 即 80% 的企业利润是由 20% 的高端用户所创造的. 而且, 对电子商务企业来说, 存在着比较特殊的用户, 即新用户. 新用户被电子商务网站收集的信息非常少, 采用一般的推荐方法并不适合. 针对新用户、高端用户和普通用户的不同, 电子商务网站需要采用不同的推荐方法, 以实现电子商务企业与用户的双赢.

3.2 用户分层的设计

针对电子商务用户 $User$ 的特点, 将其分层为新用户 $NewUser$ 、高端用户 $AdvancedUser$ 和普通用户 $OrdinaryUser$.

新用户是根据当前日期与用户第一次访问电子商务站点日期的差值来确定. 设定时间阈值 σ , 设某用户第一次开始访问商务站点日期为 $User_start$, 当前日期为 $User_current$. 如果 $User_current - User_start \leq \sigma$, 则该用户为新用户, 否则为其他用户.

对其他用户再次分层为高端用户和普通用户. 高端用户与普通用户的确定需要考虑多个因素, 包括访问站点新度 $Recency$ 、访问站点时长 $Time$ 、访问站点频度 $Frequency$ 、消费总金额 $Money$ 和消费数

量 $Amount$.

访问站点新度 $Recency$ 是一个相对值, 表征一个用户对站点的访问时间的鲜程度, 描述为:

$$Recency = \frac{Customer_end - Customer_start}{Customer_current - Customer_start},$$

其中 $User_end$ 为用户最近一次访问站点的结束日期.

访问站点时长 $Time$ 为用户每次访问站点时长之和, 描述为: $Time = \sum_{i=1}^{Count} (Time_i)$, 其中 $count$ 为一个用户访问站点的次数, $Time_i$ 为用户第 i 次访问站点所花费的时间.

访问站点频度 $Frequency$ 为用户访问站点的频繁程度, 描述为

$$Frequency = \frac{Customer_current - Customer_start}{count}.$$

消费总金额 $Money$ 为用户在电子商务站点消费总金额, 描述为: $Money = \sum_{i=1}^{Count} (Money_i)$, 其中 $Money_i$ 为用户第 i 次访问站点的消费金额.

消费数量 $Amount$ 为用户消费的总数, 描述为: $Amount = \sum_{i=1}^{Count} (Amount_i)$, 其中 $Amount_i$ 为用户第 i 次访问站点的消费数量.

对于各个因素对应的值, 利用模糊函数将它们转换到 $[0, 1]$ 范围, 这五个因素是其值越大, 转换后的值也越大. 各个因素所对应的权重是不同的, $Money$ 的权重最大, 设定为 α , $Recency$ 和 $Frequency$ 的权重较大, 都设为 β , $Time$ 和 $Amount$ 的权重最小, 都设为 γ , 即 $1 > \alpha > \beta > \gamma > 0$, 并且 $\alpha + 2\beta + 2\gamma = 1$. 设定用户阈值 θ , 如果 $\beta * Recency + \gamma * Time + \beta * Frequency + \alpha * Money + \gamma * Amount > \theta$, 则该用户为高端用户, 否则为普通用户.

电子商务用户分层可整体描述为三元组, 形式化表示为

$$User =$$

$$\langle NewUser, AdvancedUser, OrdinaryUser \rangle$$

$$NewUser =$$

$$\{ User \mid User_current - User_start \leq \sigma \}$$

$$AdvancedUser = \{ User \mid User_current - User_start > \sigma \text{ and } \beta * Recency + \gamma * Time + \beta * Frequency + \alpha * Money + \gamma * Amount > \theta \}$$

$$OrdinaryUser = \{ User \mid User_current - User_start > \sigma \text{ and } \beta * Recency + \gamma * Time + \beta * Frequency + \alpha * Money + \gamma * Amount \leq \theta \}$$

其中, σ 为时间阈值, θ 为用户阈值, $1 > \alpha > \beta > \gamma > 0$, 并且 $\alpha + 2\beta + 2\gamma = 1$.

3.3 基于用户分层的推荐策略

在进行用户分层的基础上提出了基于用户分层的推荐策略。根据用户的不同特点,采用不同的推荐方法。

新用户被收集的个人信息较少,为了吸引新用户,设计基于用户事务模式聚类信息推荐方法^[5](IRUMC)。它有三个关键步骤:第一步,根据第2节的预处理步骤,获取带有时长的用户事务模式;第二步,将相似的用户事务模式聚类,生成用户事务聚类模式;第三步,将用户当前访问路径与用户事务聚类模式进行匹配,根据设定的最小匹配门限 τ ,形成推荐集。IRUMC将相似兴趣的用户事务模式聚类到一起,利用相似兴趣的其他用户知识获得推荐集,扩大了推荐范围。它比较适合新用户和有新颖性信息需求的用户。

高端用户是电子商务企业利润的主要贡献者,为了留住高端用户,设计基于协作过滤的Web智能信息推荐方法^[6](WIIRM)。它有三个关键步骤:第一步,根据第2节的预处理步骤,获取带有时长的用户事务模式。第二步,设计改进的协作过滤方法,首先,获取用户对项目的评价,建立评价矩阵;其次,根据评价矩阵计算用户之间的相似度;最后,采用阈值可调的阈值法^[6]或 k 值可调的最优 k 个近邻法^[6]来选择近邻用户。第三步,进行基于协作过滤的信息推荐。WIIRM利用协作过滤的方法,考虑了用户访问页面的时间特性,同时实现了离线处理与在线推荐的有效结合,能够实现较为准确的信息推荐,这对于留住高端用户是很有意义的。

普通用户对电子商务企业利润贡献较小,对其的推荐服务考虑的是为企业获取更大的利益,采用基于马尔可夫决策过程的信息推荐方法^[7]。该方法利用马尔可夫决策过程,将推荐过程看作连续最优化过程,考虑了每次推荐的长期效益和期望值,目的是发现最优策略以使平均奖赏或折扣奖赏和最大。由于利用了马尔可夫决策过程,该方法考虑了一些特殊的推荐。例如,推荐一些被购买的可能性小,但利润极高的商品;推荐近期回报较小,但其远期回报较高的商品。它能较好地让企业从普通用户中获取更大的利益。

基于用户分层的推荐策略简单描述如下:

if (当前用户为新用户)

采用基于用户事务模式聚类信息推荐方法(IRUMC)进行推荐;

else

if (当前用户为为高端用户)

采用基于协作过滤的Web智能信息推荐方法(WIIRM)进行推荐;

else

采用基于马尔可夫决策过程的信息推荐方法进行推荐;

4 相关实验

测试环境为一台Dell服务器,配置为P4 2.4 GHz,内存1 GB内存,Windows 2003 Server操作系统,SQL Server 2005数据库管理系统。实验数据取自建立站点的访问log文件,总的数据量是一个月的访问记录。经过预处理,获得大约800个用户事务模式。选择20位用户对推荐页面的准确度进行评价,每个用户进行3次测试,第1次是对基于用户事务模式聚类信息推荐方法(IRUMC)进行测试,第2次对基于协作过滤的Web智能信息推荐方法(WIIRM)进行测试,第3次对基于马尔可夫决策过程的信息推荐方法进行测试。20个用户通过3次测试对推荐页面准确性进行评价,得到的平均准确度值如图1所示。

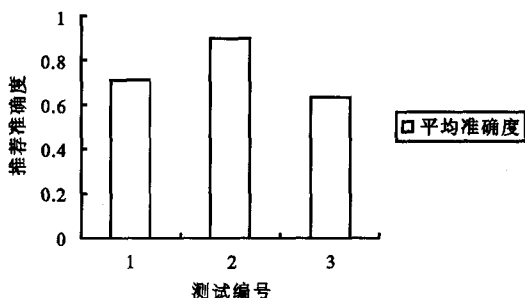


图1 推荐页面平均准确度图

从测试结果看,第1次推荐页面平均准确度值71%,第2次推荐页面平均准确度值90%,第3次推荐页面平均准确度值63%。准确度最高的是基于协作过滤的Web智能信息推荐方法,该方法针对高端用户,需要比较准确的推荐。准确度最低的是基于马尔可夫决策过程的信息推荐方法,该方法针对一般用户,在推荐的时候更多的考虑了企业的利益。准确度处于中间的是基于用户事务模式聚类信息推荐方法,该方法主要是针对新用户。实验结果表明,这三种推荐方法比较适合针对的用户群体;基于用户分层的推荐策略充分考虑了三种用户的不同特点,能实现企业与用户的双赢。

5 结束语

为了实现电子商务企业与用户的双赢^[8],文中

将电子商务用户分层为新用户、高端用户和普通用户,提出了基于用户分层的电子商务智能信息推荐策略。初步的实验结果表明,提出的推荐策略和推荐方法是有效的。下一步工作是对三种推荐方法进行进一步对比实验,并将该推荐策略应用到实际的智能信息推荐系统。

参考文献:

- [1] 许海玲,吴潇,李晓东,等. 互联网推荐系统比较研究[J]. 软件学报,2009,20(2):350-362.
- [2] 郭悦红,齐莉丽. 基于 WS-Security 的电子商务安全支付系统[J]. 微电子学与计算机,2010,27(3):151-153.
- [3] 张梅,文静华,张焕国. 基于 ATL 的电子商务协议建模与形式化分析[J]. 微电子学与计算机,2009,26(8):193-195.
- [4] Xu M, Qiu Y H, Qiu J, et al. Segmenting your users in recommender system[C]// Proceeding of the IASTED International Conference: Modeling and Simulation' 2003. Palm Springs, CA, USA. ACTA Press, 2003:584-588.
- [5] 何波,杨武,张建勋,等. 基于用户模式聚类的智能信息推荐算法[J]. 计算机工程与设计,2006,27(13):2360-2361.
- [6] 何波. 基于协作过滤的 Web 智能信息推荐方法[J]. 图书情报工作,2010,54(19):115-118.
- [7] Guy Shani, David Heckerman, Ronen I Brafman. An MDP-based recommendation system[J]. Journal of Machine Learning Research, 2005(6):1265-1295.
- [8] 田晓珍,尚东娟. Web 的个性化服务[J]. 重庆理工大学学报:自然科学版,2008,22(7):76-80.

作者简介:

何波 男,(1978-),硕士,副教授.研究方向为智能信息推荐、数据挖掘.

(上接第 20 页)

参考文献:

- [1] 郑宏超,范隆,岳素格,等. 基于库替换技术的单粒子效应故障注入仿真[J]. 微电子学与计算机,2009,26(7):60-64.
- [2] 侯明东,刘杰,马峰,等. 宇航器件中的单粒子效应及其加速器地面模拟[J]. 核物理动态,1996,13(1):32-35.
- [3] Woods J P, Nichols D K, Price W E. Investigation for single-event-upset in MSI devices IEEE trans[J]. Nucl. Sci.,1981,NS-28(6):4022.
- [4] 董代洁,郭怀里. 基于 FPGA 的可编程 SOC 设计[M]. 北京:北京航空航天大学出版社,2006.
- [5] 王磊. 基于 MicroBlaze 软核的 FPGA 片上系统设计[J]. 单片机与嵌入式系统应用,2004(7):52-54.

作者简介:

董攀 男,(1986-),硕士研究生.研究方向为抗辐照加固验证技术、嵌入式测试技术.

范隆 男,研究员,硕士生导师.研究方向为抗辐照加固验证技术.

岳素格 女,研究员,硕士生导师.研究方向为抗辐照加固设计和验证技术.

杜守刚 男,(1983-),工程师.研究方向为抗辐照加固验证技术、嵌入式测试技术.

郑宏超 男,(1983-),工程师.研究方向为抗辐照加固验证技术、嵌入式测试技术.