

基于双向关联规则的网络消费者偏好挖掘研究

刘枚莲, 刘同存, 肖吉军

(桂林电子科技大学 商学院, 广西 桂林 541004)

摘 要: 通过分析电子商务网站的 Web 服务器日志提取网络消费者的商品浏览行为信息, 利用消费者的浏览频率、浏览时间、链接路径数及路径深度估计商品对网络消费者当前浏览期间偏好的影响权重. 结合双向关联规则理论和 Apriori 算法挖掘具有相互依赖关系的商品, 找出网络消费者的商品偏好浏览路径, 根据消费者当前的浏览行为发现其可能感兴趣的物品, 并进一步计算消费者对商品的偏好程度. 最后利用自主开发的旅游电子商务网站的 Web 日志数据进行仿真实验, 挖掘网络消费者的旅游偏好. 实验结果表明, 在相同的实验条件下, 与基于关联规则的偏好挖掘方法相比, 基于双向关联规则的偏好挖掘方法的推荐精度增加, 推荐覆盖率扩大.

关键词: 电子商务; 网络消费者; 偏好挖掘; 双向关联规则

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2013)03-0020-07

Web Consumer Preference Mining Based on Bidirectional Association Rules

LIU Mei-lian, LIU Tong-cun, XIAO Ji-jun

(Business School, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: By analysis of web server logs in e-commerce to extract the information of web consumer's products browsing behavior, the weight of a product of web consumer's current preference has been estimated by means of consumer's browsing frequency, browsing time, the number of paths and links depth. The theory of bidirectional association rules was combined with the idea of Apriori to find out interdependent products with to discover consumer's product preference path, candidate products of consumer preferred have been found based on consumer's current browsing behavior, and then each preference degree of consumer to those products has been calculated respectively. Finally, the web server logs of a self-developed e-commerce web site has been used in the simulation to find out web consumers tour preference. Experimental results shows that the recommendation accuracy of preference mining method based on bidirectional association rules has improved greatly and the coverage has been expanded compared with preference mining method based on association rules when the experimental condition is equal.

Key words: E-commerce; web consumer; preference mining; bidirectional association rules

1 引言

电子商务应用的普及使得互联网商品信息成指数增长^[1], 这拓展了消费者的选择空间, 同时也增加了消费者的选择成本. 为了降低过量商品信息给消费者带来的困扰, 国内外学者提出了多种信息过滤

技术. 推荐系统是网络营销领域广泛应用的信息过滤技术, 通过分析大量历史购买数据挖掘消费者的潜在需求, 推荐消费者可能感兴趣的物品. 消费者偏好是一种消费心理, 反映消费者对某种产品或服务的兴趣或爱好程度, 受消费者个性特征以及外界环境等主客观因素的共同影响, 把握消费者偏好的变

收稿日期: 2012-06-13; 修回日期: 2012-07-24

基金项目: 国家自然科学基金项目(70862001)

化规律比较困难,但是消费者的偏好往往通过购买行为体现^[2],因此国内外众多学者着眼于消费者的购买历史,利用不同方法研究消费者偏好。

利用消费者的历史购买行为挖掘消费者偏好存在一些不足,最典型的是忽略消费者购买决策过程中的信息,而仅仅只考虑决策结果。在网络环境下,Web服务器日志可以记录网络消费者购买过程中的相关信息,如点击流、访问时间、停留时间等^[3]。因此,本文通过分析日志文件,利用双向关联规则挖掘消费者的商品访问模式,进而了解消费者偏好的变化特征,为网络营销和商品推荐决策提供依据。

2 研究综述

目前,通过挖掘网络消费者偏好来进行商品推荐与服务的方法主要分为基于协同过滤的推荐算法、基于关联规则的推荐算法和基于内容匹配的推荐算法^[4]。其中,协同过滤算法将消费者对商品的评分视为其对该商品的偏好程度,根据偏好相似的最近邻居用户对商品的偏好挖掘目标用户的偏好^[5-6]。该方法已应用于各类电子商务系统中。然而,随着电子商务系统规模的扩展,协同过滤算法面临评分数据稀疏性导致的偏好提取精确度降低的困境^[7]。为了缓解数据稀疏性问题,可以给缺省值赋予固定值或平均值,增加共同评分的项目数,但是该方法忽略了用户对项目偏好的差异性^[8]。其次是结合奇异值分解和用户特征降低项目评分数据的空间维度以提高评分密度^[9],或者利用主成分分析法处理评分数据集中的缺省值,不过这些方法极易导致数据丢失。为此,可以通过计算项目间的相似性评估缺省值^[10],但是该方法仍采用传统的度量方法计算项目相似性,评估不够精确。鉴于此,基于影响集^[11]和云模型的推荐方法应运而生^[12],前者利用当前对象的影响集来提高项目的评分密度,后者在知识层面比较用户偏好相似性以解决稀疏性问题。

随着关联规则数据挖掘技术的日益成熟,基于关联规则的商品推荐技术打破了协同过滤过度依赖评分数据的局限^[7],拓展了消费者偏好挖掘的研究领域。通过分析用户交易数据库中商品之间的关联性,挖掘用户的偏好变化轨迹,构建用户偏好文档,然后将用户的当前购买行为与偏好文档比较,达到偏好预测的目的。但是该方法在规则挖掘过程中将所有的商品同等对待,忽略了用户对商品偏好的差异性。为此,研究者提出使用加权的关联规则挖掘用户偏好的方法^[13],为每件商品赋予不同的权重以反

映其对用户偏好的影响程度。基于用户的历史交易记录仅仅能够反映消费者的历史偏好,为了挖掘用户当前购物过程的真实偏好,可以通过分析Web日志挖掘用户的偏好浏览路径^[14]。传统的关联规则技术是假设消费者的偏好是确定不变的,即消费者的偏好域是确定的,事实上消费者对商品的偏好是模糊的、不确定的,因此可以利用模糊集合理论解决消费者偏好程度边界问题^[15-16]。

基于内容匹配的商品推荐只能推荐与用户历史偏好相似的商品,时效性较差,不能发现用户的新兴趣^[17]。为了弥补上述三种模型存在的缺陷,把现有模型进行有机结合,提出混合的商品偏好挖掘方法^[5]。这些方法可以挖掘出用户对商品的偏好倾向,然而研究的主体是商品,忽视了消费者的主体地位。消费者作为市场活动的主体,其偏好受各种因素的影响而不断变化,而且其偏好最终通过浏览行为或者购买行为来体现,因此有必要建立以消费者为主体的偏好挖掘方法。

3 基于双向关联规则的网络消费者偏好挖掘

3.1 双向关联规则的定义

关联规则是用于挖掘大量数据集中项目之间空间关联性的数据挖掘技术,广泛应用于零售业,挖掘具有较高可能性被消费者同时购买的某些商品,可为商品陈列和促销提供决策依据。关联规则的形式化定义为^[18]:对于项目集 I ,事务数据库 T ,每个事务 T_i 均是项目集 I 的子集,即 $T_i \subset I$,如果有集合 $A \subseteq T, B \subseteq T$,并且 $A \subseteq I, B \subseteq I, A \cap B = \emptyset$,则将形如 $A \rightarrow B$ 的蕴含式称为关联规则, A 为规则前件, B 为规则后件。关联规则反映项集 A 对 B 的依赖关系,依赖程度用支持度 s 和置信度 c 衡量。 s 是事务 T 中同时包含 A 和 B 的百分比, c 是事务 T 中包含 A 的同时也包含 B 的百分比。

基于支持度—置信度框架的关联规则得到广泛的应用,但是只能发觉具有单项依赖关系的关联规则,而现实生活中却存在大量相互依赖关系的事务,因此,双向关联规则思想诞生了^[19-21]。将形如 $A \leftrightarrow B$ 的规则称为双向关联规则,并对其合理性进行论证。项集 A 和 B 互为规则前后件,相互依赖程度同样用支持度和置信度衡量,将同时满足最小支持度和置信度阈值的双向规则称为强双向关联规则^[21]。

命题1 如果规则 $A \rightarrow B$ 和 $B \rightarrow A$ 的支持度和置信度均满足最小支持度和置信度阈值,那么规则 $A \leftrightarrow B$ 也满足最小支持度和置信度阈值。

证明: 设 $L_k = \{I_1, I_2, \dots, I_k\}$ 为事务数据库 D 中的

一个频繁项集,非空真子集 $L_A = \{I_i\}, L_B = \{I_1, I_2, \dots, I_{i-1}, I_{i+1}, \dots, I_k\}$, 最小支数 $s_{\min}$, 最小置信度 $c_{\min}$. 由支持度和置信度的定义可知:

$$S_{L_A \rightarrow L_B} = \frac{P(I_1, I_2, \dots, I_i, \dots, I_k)}{|D|} \geq s_{\min}$$

$$C_{L_A \rightarrow L_B} = \frac{P(P(I_1, I_2, \dots, I_i, \dots, I_k))}{P(i)} \geq c_{\min}$$

$$S_{L_B \rightarrow L_A} = \frac{P(I_1, I_2, \dots, I_i, \dots, I_k)}{|D|} \geq s_{\min}$$

$$C_{L_B \rightarrow L_A} = \frac{P(I_1, I_2, \dots, I_i, \dots, I_k)}{P(I_1, I_2, \dots, I_{i-1}, I_{i+1}, \dots, I_k)} \geq c_{\min}$$

由此可知, $L_A \cap L_B = \emptyset$ 且 $L_A \cup L_B = L_k$, 由于 $S_{L_A \rightarrow L_B} = S_{L_B \rightarrow L_A} \geq s_{\min}$, 所以 $S_{L_A \leftrightarrow L_B} \geq s_{\min}$ 成立. 由 $C_{L_A \rightarrow L_B} > C_{L_B \rightarrow L_A}$, 且 $C_{L_B \rightarrow L_A} \geq c_{\min}$, 因此, $C_{L_A \leftrightarrow L_B} \geq c_{\min}$ 也成立.

综上所述, 命题成立.

3.2 双向关联规则的挖掘

3.2.1 数据预处理

本文通过数据过滤、用户识别、会话识别和事务识别技术对收集到的 Web 日志文件进行预处理, 删除噪声数据, 转换为便于进行数据挖掘的规范化数字存储格式. 考虑到消费者通常会在不同时间点对电子商务网站及其商品多次访问, 将消费者的每次访问会话期均作为一条事务, 构造成关系型的事务数据库. 若电子商务网站的所有商品构成的集合为 $I = \{i_1, \dots, i_2, \dots, i_n\}$, 则事务数据库可表示为 $T = \{t_1, t_2, \dots, t_i, \dots, t_n\}$, 每一条事务均表示消费者在特定会话期对商品的浏览记录, 为了区分每个商品对消费者的重要性, 用 k 维的 2 元组向量表示每条事务, 即, $t_i = \langle \text{sessionid}, \{(i_1, w_1), \dots, (i_i, w_i), \dots, (i_k, w_k)\} \rangle$, 且 $i_i \in I$. sessionid 为会话 ID, 唯一标识一条交易事务, (i_i, w_i) 表示该消费者在会话 sessionid 中浏览了商品 i_i , w_i 为该商品对消费者的重要性权重, k 为消费者在该会话期所浏览的商品总数.

3.2.2 双向关联规则提取

规则提取过程就是找到事务数据库中同时满足最小支持度和置信度阈值的规则, 可分为频繁项集挖掘和规则提取两步. 具体的规则提取过程包括商品权重设定、支持度计算和规则提取.

(1) 商品权重设定. 权重 w_i 设置的合理性直接影响到规则提取的精确性^[13]. 研究发现, 消费者对其感兴趣的商品通常会长时间、高频率的浏览^[13]. 因此本文将权重表示成消费者对商品 i_i 的归一化单字节浏览时间 $D(i_i)$ 和单路径浏览频率 $F(i_i)$ 的函数:

$$w(i_i) = \frac{D(i_i)}{F(i_i)} \quad (1)$$

其中 $D(i_i)$ 和 $F(i_i)$ 分别为:

$$D(i_i) = \frac{t_i/s_i}{\sum_{j=1}^n t_j/s_j} \quad (2)$$

$$F(i_i) = \frac{(f_i/n_i) * (h_i/\sum_{k=1}^{n_i} h_k)}{\sum_{j=1}^n [(f_j/n_j) * (h_j/\sum_{k=1}^{n_j} h_k)]} \quad (3)$$

s_i 为商品 i_i 的描述字节数, n_i 为能够到达该商品的路径数, h_k 为每一路径的深度, t_i 为消费者对该商品的总浏览时间, f_i 为总浏览频率.

(2) 支持度计算. 传统关联规则挖掘算法中的支持度计算没有考虑到项集在事务及事务数据库中的重要性. 为此, 本文综合考虑这两者因素对传统的计算方法加以改进, 以便精确挖掘有效的双向关联规则. 为了便于描述, 首先给出如下定义.

定义 1 若项集为 A , 事务为 t_i , 则将 A 中项目的权重之和对 t_i 中所有项目的权重之和的比值定义为 A 在事务 t_i 中的重要性权重. 表示为:

$$w(A, t_i) = \begin{cases} \frac{\sum_{i \in A} w(i_i)}{\sum_{k \in t_i} w(i_k)} & A \subseteq t_i \\ 0 & A \not\subseteq t_i \end{cases} \quad (4)$$

$|A|$ 和 $|t_i|$ 分别为项集 A 和事务 t_i 所包含的项目总数, $w(i_i)$ 为项目 i_i 的权重.

定义 2 对于事务 t_k , 将 t_k 中项目的权重之和对所有事务所包含项目的权重之和的比值定义为该事务在事务数据库 T 中的重要性权重. 表示为:

$$w(t_k, T) = \frac{\sum_{i=1}^{|t_k|} w(i_i)}{\sum_{k=1}^{|T|} w(i_k)} \quad (5)$$

$|t_k|$ 和 $|T|$ 分别为事务和事务数据库中所包含的项目总数.

依据上述定义, 对于任意项集 X 在事务数据库 T 中的重要性权重可表示为^[13]:

$$w(X, T) = \frac{\sum_{t_i \in TX} w(t_i, T) \times w(X, t_i)}{\bar{w} \times \sum_{k=1}^{|T|} w(t_k, T)} \quad (6)$$

$\bar{w}$ 为所有项目在事务数据库中的平均权重, TX 为事务数据库中包含项集 X 的事务集. 因此, 事务数据库中任意项集 X 的支持度可以用 $ws_p(X) = w(X, T) \times \text{support}(X)$ 表示, $\text{support}(X)$ 为事务数据库中包含项集 X 的百分比, 从而双向关联规则 $X \leftrightarrow Y$ 的支持度计算

方法为:

$$\begin{aligned} \text{usp}(X \leftrightarrow Y) &= \min\{w(X, T), w(Y, T)\} \times \\ &\quad \text{support}(X \leftrightarrow Y) \\ &= \min\{w(X, T), w(Y, T)\} \times \frac{\text{count}(X \cup Y)}{|T|} \end{aligned} \quad (7)$$

$\text{count}(X \cup Y)$ 为事务数据库中同时包含项集 X 和 Y 的数量, $|T|$ 为事务数据库规模。

(3) 双向关联规则生成。关联规则挖掘的核心就是找到事务数据库中的频繁项集, Apriori 算法是最具影响力的频繁项集挖掘算法^[18], 算法使用层次顺序搜索的循环方法产生频繁项目集, 频繁 k 项集用于找到频繁 $(k+1)$ 项集。由于本文考虑了项目权重, 传统的 Aprior 算法不能挖掘出有效的加权频繁项集, 因此本文使用加权的关联规则频繁项集 Aprior 挖掘算法挖掘频繁项集。然后, 按照给定的最小置信度阈值生成双向关联规则, 并以 $R = \langle (I_1, I_2, \dots, I_k), (I_{k+1}, I_{k+2}, \dots, I_{k+l}), (w_1, w_2, \dots, w_{k+l}), s, c \rangle$ 的格式存储在双向关联规则数据库 R_D 中, 其中 $(I_1, I_2, \dots, I_k)$ 和 $(I_{k+1}, I_{k+2}, \dots, I_{k+l})$ 分别为规则前件和后件, $(w_1, w_2, \dots, w_{k+l})$ 为各个项目的权重, s 和 c 分别为该规则的支持度和置信度。

3.3 网络消费者的偏好挖掘

大规模电子商务网站存储大量用户对商品的浏览日志数据, 通过分析这些数据可以挖掘商品之间的内在关联性, 从而达到挖掘消费者潜在偏好的目的。

为了能够发现目标用户的偏好, 首先根据用户的当前浏览行为和双向关联规则数据库 R 提取其可能感兴趣的候选商品列表, 然后找到每个商品的关联集, 计算用户对该商品的偏好程度。为了便于描述, 首先给出双向关联集的概念。

定义 3 设有目标项目 I_i , 关联规则数据库 R_D , 将 R_D 中包含 I_i 的规则前件构成的集合用 $H = \{H_{i1}, H_{i2}, \dots, H_{im}\}$ 表示, 其中 $H_{ij} = \{I_1, \dots, I_i, \dots, I_k\}$, 将包含 I_i 的规则后件构成的集合用 $B = \{B_{i1}, B_{i2}, \dots, B_{in}\}$ 表示, 其中 $B_{ij} = \{I'_1, \dots, I'_i, \dots, I'_k\}$ 。以 H 中的元素 H_{ij} ($H_{ij} \in H$) 为规则前件的规则后件项目集定义为 B_{ij}' , 以 B 中的元素 B_{ij} ($B_{ij} \in B$) 为规则后件的规则前件构成的项目集定义为 H_{ij}' , 则将 $R_i = \bigcap_{j=1}^m B_{ij}' \cap \bigcap_{j=1}^n H_{ij}'$ 称为项目 I_i 的双向关联集。

设消费者 u 当前的商品浏览序列为 $\langle I, W \rangle$, I 为项目集合, W 为每个项目对应的权重。扫描关联规则数据库, 包含商品序列中一个或多个商品的规则前件的规则后件序列为 $B_{u,i}$, 则该消费者在当前购买过程中可

能感兴趣的商品序列集为 $C_u = \bigcup_{i=1}^k B_{u,i}$, k 为找到的规则后件总数。为了能够发现消费者对 C_u 中商品 I_k 的偏好程度, 根据定义 4 找到目标项目的双向关联集 R_i , 则消费者对商品 I_k 的偏好程度为:

$$I_k = \frac{2 \text{fresh}(I_k) * (w(I_k) / \sum_{j \in R_i} w(I_j))}{\text{fresh}(I_k) + w(I_k) / \sum_{j \in R_i} w(I_j)} \quad (8)$$

其中 $\text{fresh}(I_k)$ 为用户对商品 I_k 的访问新鲜度^[13], 离当前时间越近访问的商品, 新鲜度越高, 用 $\text{fresh}(I_k) = i/n$ 表示, i 为访问次序, n 为双向关联集 R_i 的商品数。

最后, 对候选集中的商品按照偏邓程度降序排列, 前个偏好程度较高的商品可推荐给目标用户, 为其购买决策提供依据。

4 实验仿真

实验平台是 PC(Pentium T4400, CPU 2.20GHz, 内存 2GB), Windows XP 操作系统, 使用 C# 语言, 在 Microsoft Visual Studio 2008 开发环境下编写算法程序。

4.1 实验数据

本文以自主开发的旅游电子商务网站 (<http://www.youdaike.com>) 的 Web 访问日志为例进行仿真实验。该网站主要提供旅游产品服务, 包括景区门票预订、旅游线路预订和车辆出租等。实验选取 2011 年 10 月 6 日—2011 年 11 月 6 日, 共一个月的 Web 日志进行分析。先使用数据清洗技术, 删除与商品无关的日志记录, 同时删除用户最后访问的商品, 后者无法计算其对该商品的浏览时间。另外, 为了排除个别用户的偶然浏览行为, 仅保留用户在会话期对商品访问数大于 10 的 Web 日志记录, 最后得到 1331 条实验数据, 共包含 95 个 IP 用户和 67 件商品。另外, 为了简化后面实验的工作量, 利用公式(1)、(2)和(3)计算每个商品对用户偏好的影响权重, 部分数据如表 1 所示。

为了便于实验比较, 将实验数据进一步划分为训练集和测试集, 对此, 本文引入变量 x , 表示训练集占整个数据集的百分比。在本实验中, 将 x 设为 0.8。

4.2 评价指标

本文采用精确度和覆盖率两个指标进行衡量^[13]。精确度是指系统推荐的商品中, 满足用户偏好的商品占所推荐商品的比例。覆盖率是指系统推荐的商品占用户所有偏好商品的比率。设推荐的项目集为 R , 测试集中的项目为 T , 则推荐精确度和覆盖率数学表达式分别为:

表 1 实验数据示例

访问时间	商品	总字节	浏览 时间	浏览 频率	路径 数	平均 深度	项目 权重
220. 181. 94. 226	06/10/2011:01:07 42	xianlu_34. html	306241046	3	3	2	0. 977
67. 195. 115. 113	06/10/2011:01:51:14	jingqu_14. html	32523657	10	8	1. 4	0. 322
67. 195. 115. 113	06/10/2011:06:00:59	xianlu_30. html	32954358	3	5	1. 8	0. 578
203. 208. 60. 73	06/10/2011:07:37:02	jingqu_43. html	2931776	6	11	2. 45	0. 100
203. 208. 60. 73	06/10/2011:07:40:48	jingqu_24. html	3136777	6	1	3	0. 095
203. 208. 60. 73	06/10/2011:07:42:05	nkjingqu_48. html	26836283	4	1	2	0. 612
220. 181. 94. 226	06/10/2011:08:29:58	tuangou_6. html	256111192	4	1	3	0. 998
203. 208. 60. 210	06/10/2011:08:38:39	jingqu_48. html	29637953	1	10	3	5. 311
220. 181. 94. 226	06/10/2011:08:48:21	nkjingqu_47. html	278361111	3	1	2	1. 141

Precision = $\frac{T \cap R}{R}$ (11)

Coverage = $\frac{T \cap R}{T}$ (12)

4.3 实验过程与结果分析

4.3.1 阈值选取实验

最小支持度和置信度阈值是影响关联规则质量的主要因素,因此首先通过实验说明在不同的置信度和支持度水平下,双向关联规则与关联规则数量的变化规律. 首先将置信度设为 50%,观察支持度以 1%的间隔在 9%~16%之间变动时所找到的规则数量的变化规律;然后将支持度设为 9%,观察置信度以 10%的间隔在 50%~90%之间变动时所找到的规则数量的变化规律. 实验结果如图 1~图 2 所示.

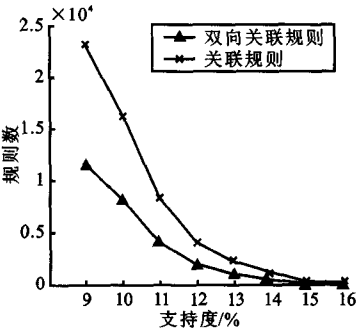


图 1 支持度阈值与规则数量之间的关系

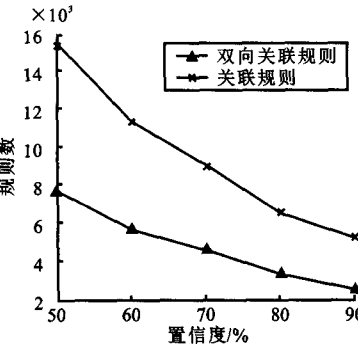


图 2 置信度阈值与规则数量之间的关系

实验结果表明,当置信度或支持度水平一定情况下,所挖掘到的关联规则和双向关联规则数量均随着支持度或置信度水平的降低而逐渐增多,当降低到一定程度时,规则数量呈剧增状态. 在置信度和支持度较低情况下,所挖掘到的规则质量较低,商品之间的关联性不强. 另外,图中可以发现,在支持度和置信度相同情况下,所挖掘到的双向关联规则数量明显少于普通关联规则数量,这是由于在规则提取时同时考虑规则前件和后件的置信度,仅保留商品之间具有较强相互依赖关系的规则.

为了进一步比较不同的置信度和支持度水平对推荐精度的影响,为后续实验选取恰当的支持度和置信度阈值,本文将置信度水平以 10%的间隔在 50%~90%之间滑动,支持度以 2%为间隔在 10%~16%之间滑动进行实验,每次为用户推荐 5 件商品,实验结果如图 3~图 4 所示.

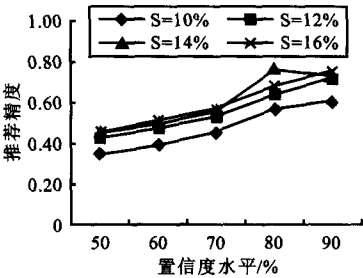


图 3 置信度阈值对推荐精度的影响

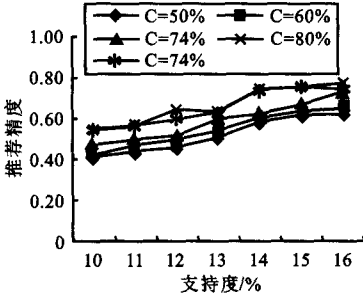


图 4 支持度阈值对推荐精度的影响

由图 3 可知,在推荐数量及支持度不变情况下,推荐精度随着置信度水平的提高而逐渐提高,当置信度水平达到一定水平后,推荐精度的提高幅度逐渐减小.在同一置信度水平下,支持度越高则推荐精度越高,支持度达到一定水平后,推荐精度随支持度增加而提高的幅度逐渐减小.由图 4 可知,在推荐数量及置信度水平不变情况下,推荐精度随着支持度水平的提高而逐渐提高,支持度达到一定水平后,推荐精度的增长幅度逐渐减小.总体来说,较高的置信度和支持度水平可以提供较高的推荐精度,这是

由于在高置信度和支持度下,能够较为精确挖掘具有相互依赖关系的商品,精确地挖掘出消费者的偏爱路径,达到商品精确推荐的目的.但是,如果设置较高的阈值则会导致原本在关联集中的项目被遗漏掉,因此,具体的阈值设置需要根据实际情况动态调整.

4.3.2 双向关联规则生成与偏好挖掘分析

本文在上述实验基础上选择 14% 的最小支持度阈值和 70% 的最小置信度阈值生成双向关联规则,生成的双向关联规则如表 2 所示.

表 2 双向关联规则示例

双向关联规则	支持度 (%)	A→B 置信度 (%)	B→A 置信度 (%)
jingqu_10. html ↔ jingqu_14. html	29.3	92.03	70.90
jingqu_10. html ↔ tuangou_20. html, jingqu_14. html	27.9	87.87	83.65
jingqu_17. html ↔ jingqu_30. html, jingqu_24. html, jingqu_45. html	15.7	97.14	89.03
jingqu_15. html, jingqu_48. html ↔ xianlu_12. html, tuangou_11. html	17.5	71.20	96.05
xianlu_20. html, tuangou_11. html ↔ jingqu_15. html, jingqu_48. html	18.2	91.48	74.17
xianlu_20. html, tuangou_20. html ↔ jingqu_15. html, xianlu_12. html	16.7	79.07	84.40
tuangou_11. html, xianlu_12. html ↔ jingqu_48. html, jingqu_47. html	17.5	96.46	71.78
jingqu_48. html, jingqu_13. html ↔ xianlu_20. html, xianlu_12. html	19.3	78.53	97.70
jingqu_48. html, jingqu_13. html, jingqu_15. html ↔ xianlu_20. html, tuangou_20. html	15.2	90.80	71.82
xianlu_20. html, jingqu_48. html, jingqu_13. html ↔ xianlu_12. html, jingqu_15. html	19.6	91.85	99.05

假设网络消费者在当前购物初始阶段依次所浏览的商品为 jingqu_10. html、jingqu_15. html、jingqu_13. html、xianlu_20. html,则根据生成的双向关联规则挖掘到的网络消费者可能感兴趣的商品集合为:{jingqu_49. html, tuangou_11. html, jingqu_46. html, xianlu_21. html, xianlu_5. html, tuangou_20. html, xianlu_30. html, xianlu_17. html, jingqu_11. html, jingqu_14. html, xianlu_24. html}.对于每件商品,找到其对应的双向关联集,并计算消费者对商品的喜好程度为{0.33,0.52,0.37,0.47,0.51,0.093,0.35,0.69,0.14,0.23,0.27}.根据计算结果,可知该消费者最可能偏好的商品序列为{xianlu_17. html, tuangou_11. html, xianlu_5. html, xianlu_21. html, jingqu_46. html},因此应该将这些商品推荐给当前网络消费者,为其购买决策提供依据.

4.3.3 与其他方法的比较

本文进一步验证基于双向关联规则挖掘偏好的有效性,实验以传统的基于关联规则偏好挖掘方法进行对比.实验以 2 为间隔将商品推荐数由 2 逐渐增加到 12,通过多次实验计算推荐精确度和覆盖率,最后取平均值观察变化趋势,实验结果如图 5~图 6 所示.

由图可知,在商品推荐数量不变情况下,基于双向关联规则的偏好挖掘推荐精度和覆盖率明显优于

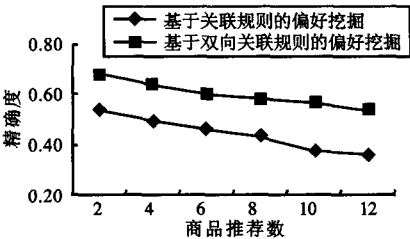


图 5 推荐精度比较

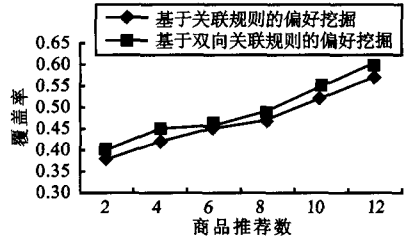


图 6 覆盖率比较

基于关联规则的偏好挖掘方法,并且随着商品推荐数量的增加,两种方法的推荐精度均逐渐降低,而推荐覆盖率却逐渐增加,但是双向关联规则的方法始终优于关联规则方法.

5 结束语

本文利用双向关联规则挖掘技术分析 Web 日志中消费者的商品浏览行为信息,找出网络消费者的浏览偏爱路径,然后根据消费者的当前浏览行为

找到其可能感性兴趣的商品以及偏好程度,最后选择偏好程度较高的商品推荐给消费者.实验结果表明,基于双向关联规则的偏好挖掘方法能够较好的挖掘消费者的潜在偏好,推荐精度和覆盖率显著优于传统的基于关联规则的挖掘方法.

参考文献:

- [1] Cao YK, Li YF. An intelligent fuzzy-based recommendation system for consumer electronic product[J]. Expert Systems with Applications, 2007, 23(1): 230-240.
- [2] 张志宏,寇纪淞,陈富赞,等. 基于遗传算法的顾客购买行为特征提取[J]. 模式识别与人工智能, 2010, 23(2): 256-266.
- [3] 过蓓蓓,方兆本. 基于 SVM 的 web 日志挖掘及潜在客户发现[J]. 管理工程学报, 2010, 24(1): 129-133.
- [4] Zhang Y Y, Jiao J X. An associative classification-based recommendation system for personalization in B2C e-commerce applications[J]. Expert Systems with Applications, 2007, 33(2): 357-367.
- [5] Kim D, Yum B J. Collaborative filtering based on iterative principal component analysis[J]. Expert Systems with Applications, 2005, 28(4): 823-830.
- [6] Kim B M, Li Q, Park C S, et al. A new approach for combining content-based and collaborative filters[J]. Journal of Intelligent Information Systems, 2006, 27(1): 79-91.
- [7] Lee TQ, Park Y, Park YT. A time-based approach to effective recommender systems using implicit feedback[J]. Expert Systems with Applications, 2008, 34(4): 3055-3062.
- [8] Garcia I, Sebastia L, Onaindia E. On the design of individual and group recommender systems for tourism[J]. Expert Systems with Applications, 2011, 38(6): 7683-7692.
- [9] Vozalis M G, Margaritis G M. Using SVD and demographic data for the enhancement of generalized Collaborative Filtering[J]. Information Sciences, 2007, 117(15): 3017-3037.
- [10] 邓爱林,朱扬勇,施伯乐. 基于项目评分预测的协同过滤推荐算法[J]. 软件学报, 2003, 14(9): 1621-1628.
- [11] 陈健,印鉴. 基于影响集的协作过滤推荐算法[J]. 软件学报, 2007, 18(7): 1685-1694.
- [12] 张光卫,李德毅,李鹏,等. 基于云模型的协同过滤推荐算法[J]. 软件学报, 2007, 18(10): 2403-2411.
- [13] Forsati R, Meybodi M R. Effective pages recommendation algorithms based on distributed learning automata and weighted association rules[J]. Expert Systems with Applications, 2010(37): 1316-1330.
- [14] Chou P H, Li P H, Chen K K, et al. Integrating web mining and neural network for personalized e-commerce automatic service[J]. Expert Systems with Applications, 2010, 37(4): 2898-2910.
- [15] Leung C W K, Chan S C F, Chung F L. A collaborative filtering framework based on fuzzy association rules and multiple-level similarity[J]. Knowledge and Information Systems, 2006, 10(3): 357-381.
- [16] Zenebe A, Zhou L, Anthony FN. User preferences discovery using fuzzy models[J]. Fuzzy set and systems, 2010(161): 3044-3063.
- [17] Kim H N, Ha I, Lee K S, et al. collaborative user modeling for enhanced content filtering in recommender systems[J]. Decision Support Systems, 2011, 51(4): 722-781.
- [18] Chen G Q, Wei Q, Liu D, et al. Simple association rules (SAR) and the SAR-based rule discovery[J]. Computers & Industrial Engineering, 2002, 43(4): 721-733.
- [19] Campbell J I, Nicole N D Robert. Bidirectional associations in multiplication memory: conditions of negative and positive transfer[J]. Journal of Experimental Psychology: Learning, Memory, and Cognition, 2008, 34(3): 546-555.
- [20] Du Y J, Li H M. Strategy for mining association rules for web pages based on formal concept analysis[J]. Applied Soft Computing, 2010, 10(3): 772-783.
- [21] 曾令明,唐常杰,阴小雄,等. 双向关联规则挖掘及其相关性分析[J]. 计算机工程与设计, 2005, 26(10): 2585-2633.

作者简介:



刘枚莲 女, (1972-), 博士, 教授, 硕士生导师. 研究方向为电子商务、物流工程.

刘同存 男, (1984-), 硕士研究生. 研究方向为电子商务.
肖吉军 男, (1975-), 博士研究生. 研究方向为电子商务.