文章编号：0253-374X(2013)06-0953-08 DOI:10.3969/j.issn.0253-374x.2013.06.025

考虑长期与短期兴趣因素的用户偏好建模

王洪伟，邹 莉
（同济大学 经济与管理学院，上海 200092）

摘要：鉴于电子商务网站推荐系统的需要，将用户兴趣分为长期兴趣和短暂兴趣，并提出一种基于长期兴趣和短暂兴趣的用户偏好表示法.利用web服务器数据库的数据，采用无监督学习方法，对用户注册信息进行挖掘，提取出用户长期兴趣.基于向量映射，对web服务器日志上的用户使用记录数据和内容数据进行分析，提取用户短暂兴趣.通过用户反馈信息修正“粗糙”用户偏好文档，使得用户偏好文档更新得以实现.最后，应用了实证案例验证了该方法的合理性和有效性.

关键词：web数据挖掘；长期兴趣；短暂兴趣；用户偏好
中图分类号：C931.6 **文献标志码**：A

Modeling Users' Preference Based on Long- and Short-term Interests

WANG Hongwei, *ZOU Li*
(College of Economics and Management, Tongji University, Shanghai 200092, China)

Abstract: In view of the needs of E-commerce website for recommendation system, user interests are divided into the long-term interest and the short-term interest, furthermore, based on the long-term interest and the short-term interest, a way to describe users' preference is proposed. On the basis of the data from the web server database, users' registration information can be fully mined to abstract users' long-term interest by using unsupervised learning. Both the records data and content data on the server log are analyzed to abstract users' short-term interest by vector mapping. Moreover, the rough profile presenting users' preference can be modified by dealing with users' feedback, as a rescut, updating users' preference profile becomes possible. Case analysis illustrates to a certain extent this method is reasonable and feasible.

Key words: web data mining; long-term interest; short-term interest; users' preference

电子商务环境下，构建用户偏好模型是实现个性化服务的关键环节.及时了解用户需求，准确把握用户偏好，对用户而言，能帮助他们在海量信息空间中高效率地筛选信息，对商家而言，能够针对性地为用户设计并推荐产品，提高服务效率.然而，出于对个人私密信息安全性的顾虑，多数用户不愿意在线提交过多的个人信息，这给用户偏好的识别带来困难.因此，如何根据用户在线行为动态地识别用户兴趣与偏好，是电子商务系统实现自动推荐和个性化服务急需解决的问题.

1 相关研究综述

关于用户偏好建模的研究，主要从用户兴趣的提取和用户偏好的建模两方面展开.

1.1 用户兴趣的提取

在电子商务的信息检索与过滤中，建立用户偏好文档使跟踪用户的行为和兴趣成为可能，有助于商家为用户提供个性化信息及产品服务[1-2].在心理学中，兴趣被认为是人们探究某种事物或者从事某种活动的心理倾向.人对感兴趣的东西会表现出很大的积极性，并且产生某种肯定的情绪体验.可以认为，用户兴趣的大小和用户对这个兴趣相关的信息的需求量是相关的.特别是在电子商务中，用户兴趣的大小对兴趣相关产品的接受程度是相关的，如图1所示.因此，识别并提取用户的兴趣成为了亟待解决的问题.

人的兴趣可以分为长期兴趣和短暂兴趣，长期兴趣是由个体的倾向性引起，相对稳定，与个人的成长背景、学历、人生观、价值观等因素相关联.而短暂兴趣通常是由于当前环境下的某些条件和刺激而产

收稿日期：2012-06-19
基金项目：国家自然科学基金(70971099)；中央高校基本科研业务费专项资金(1200219198)；上海市科技发展基金软科学研究博士生学位论文资助(12692193000)
第一作者：王洪伟(1973—)，男，副教授，博士生导师，管理学博士，主要研究方向为商务智能与情感计算.
E-mail：hwwang@tongji.edu.cn

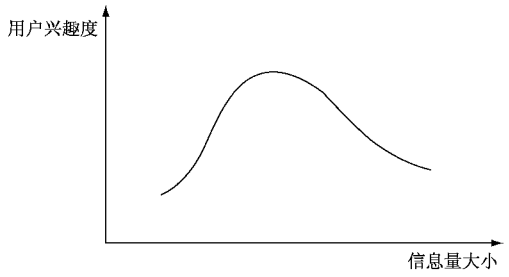

图1 兴趣与信息量关系示意图

Fig.1 Correspondence of interests and information amount

生,相对不稳定,容易消逝,但却对用户的当前偏好起着重要的实时影响作用,成为了商家最为关注的部分.

1.2 用户偏好的建模

用户偏好模型的表示主要有三种方式:空间布尔模型、向量空间模型和潜在语义索引模型.布尔模型就是给定一系列具有二值逻辑的特征变量.这些变量从文档中抽取出来,用来描述文档的特征,如关键字和索引词等.通过布尔操作符把表示文档信息的特征变量构成布尔表达式[3].为了改进传统的布尔模型,后来又提出了扩展的布尔模型信息检索系统.这个布尔模型信息检索系统是介于布尔查询处理和向量处理模型中间,在以布尔模型为基础的查询构架上增加了关键词相对于查询或文档的重要程度,即权重.基于 Salton 等提出的向量空间模型,用户偏好文档用具有最高权值的文档关键字组成的向量表示,这里的权值用 TF-IDF(termfrequency-inverse document frequency)方法计算[4].Chen 等构建了一种帮助用户进行网络浏览和检索的代理 WebMate,它通过监视用户的浏览行为和利用用户的主动反馈来建立用户偏好文档[5].潜在语义索引模型利用字项与文档对象之间的内在关系形成信息的语义结构[6].这种语义结构反映了数据间最主要的联系模式,忽略了个体文档对词的不同使用风格.文档多以多维向量来表示,关键字向量中的值表示字在文档中出现的频率.

总体来看,用户偏好建模已有相关的理论研究,并得到一定的应用.但相关理论尚不成熟,尤其是从心理学角度看,用户偏好受长期个人兴趣和短期个人兴趣两方面影响,并且存在一定的相互联系.而目前的用户偏好建模忽略了这一点,导致对用户兴趣的提取和转换难以客观地描述.本文将阐述利用用户兴趣进行用户偏好的提取过程.

2 用户偏好提取的理论模型

用户偏好受长期个人兴趣和短期个人兴趣两方面影响,所以用户的兴趣偏好文档可以描述为

$$\boldsymbol{P}=\{\boldsymbol{L},\boldsymbol{S}\} \tag{1}$$

式中:$\boldsymbol{L}$ 表示用户的长期兴趣,$\boldsymbol{S}$ 表示用户的短期兴趣.由于兴趣的多样性,所以 $\boldsymbol{L}$ 和 $\boldsymbol{S}$ 可以分别表示为

$$\boldsymbol{L}=\{l_1,l_2,\cdots,l_m\},\boldsymbol{S}=\{s_1,s_2,\cdots,s_n\}$$

此时 $\boldsymbol{P}=\{l_1,l_2,\cdots,l_m,s_1,s_2,\cdots,s_n\}$ 表述为用户兴趣偏好.

在用户的各种长短期兴趣中,为了清楚地区分其类别和用户感兴趣的程度,兴趣向量应包含更多的信息.因此,对于每个 $l_i,s_j(i=1,2,\cdots,m;j=1,2,\cdots,n)$,引入表示类别属性的变量 c_i,c_j 和表示权重属性的变量 ω_i,ω_j,所以 l_i,s_j 可以扩展成

$$\begin{aligned} l_i &=<l_i,\omega_i,c_i>,\ i=1,2,\cdots,m \\ s_j &=<s_j,\omega_j,c_j>,\ j=1,2,\cdots,n \end{aligned} \tag{2}$$

综合式(1)和(2)可以得到用户偏好文档可表示成一个二维表

$$\boldsymbol{P}=\begin{Bmatrix} l_1 & l_2 & \cdots & l_m & s_1 & s_2 & \cdots & s_n \\ \omega_1 & \omega_2 & \cdots & \omega_m & \omega_{m+1} & \omega_{m+2} & \cdots & \omega_{m+n} \\ c_1 & c_2 & \cdots & c_m & c_{m+1} & c_{m+2} & \cdots & c_{m+n} \end{Bmatrix} \tag{3}$$

为了方便表示,缩写成

$$\begin{aligned} \boldsymbol{P}=\{&<l_1,\omega_1,c_1>,<l_2,\omega_2,c_2>,\cdots,<l_m,\omega_m,c_m>,<s_1,\omega_{m+1},c_{m+1}>,<s_2,\omega_{m+2},c_{m+2}>,\\ &\cdots,<s_n,\omega_{m+n},c_{m+n}>\} \end{aligned}$$

式中:l_m,s_n 分别为代表长期兴趣和短暂兴趣的某个属性值;c_{m+n} 表示用户兴趣对应的产品所属产品类别;ω_{m+n} 是以属性值词汇为代表的兴趣权重,表示用户对某个类别产品的感兴趣程度,它是一个随着用户反馈信息不断变动的值,并且有 $\omega_1+\omega_2+\cdots+\omega_{m+n}=1$.为了帮助理解利用长短期兴趣表示偏好文档,下面举例了淘宝网的产品分类标准,用户在淘宝网站上搜索产品时也是利用该标准进行产品选择.

例如,淘宝网中的商品类别可细分为服装、电子产品、护肤品、化妆品、电器、图书、健身器材等;产品属性包括颜色、尺寸、价格、风格、产地、质地等;描述产品属性的属性值词汇集合就是描述产品的具体的属性值词汇(例如:红色、休闲、中国、1 000～1 200 RMB、棉等).为此,可以构建"类别-属性-属性值"的树形结构,如图2所示.

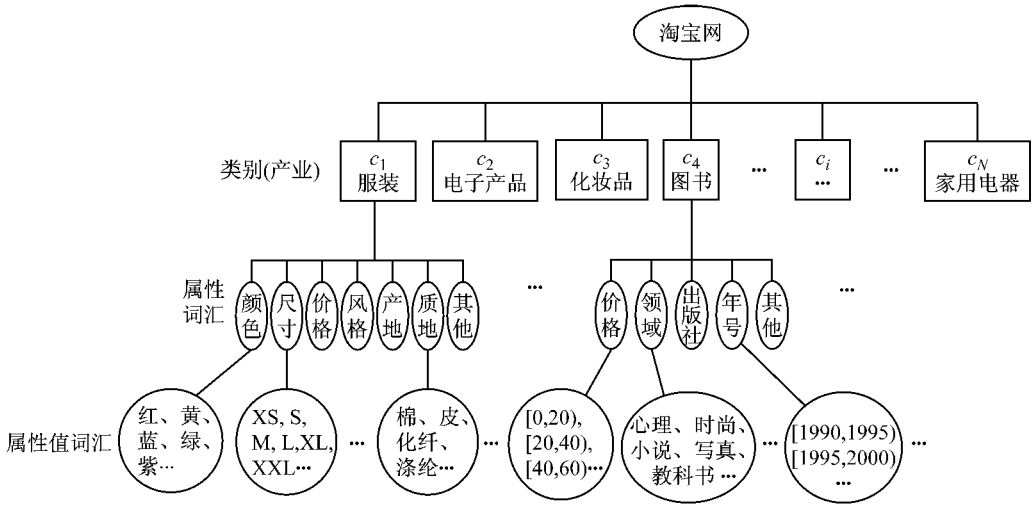

图 2　类别–属性–属性值树形结构图

Fig. 2　The tree structure of products-attributes-features

3　用户偏好提取过程

第 2 节已经提出了基于用户长短期兴趣建立用户偏好文档的理论模型，对于整个模型应用的关键在于如何识别用户兴趣二维表 $\boldsymbol{P}$.

用户兴趣的获取有两种方式：① 显式获取，即通过用户主动提供自己的兴趣来获得用户偏好[1-2]. 这要求用户非常清楚自己当前的信息需求，适用于专业性用户. ② 隐式获取，即通过用户访问 web 的相关反馈信息来获得用户偏好[7-8]. 电子商务环境下，尤其在 B–C 和 C–C 模式下，客户专业性不强，为了准确地提取用户偏好，采用隐式获取方式比较合理.

根据用户的浏览内容和浏览行为来构建用户的兴趣模型，分为三步：① 通过用户注册信息来判断用户类别，获得用户长期兴趣向量. ② 通过观察和跟踪用户的浏览行为(来源于服务器网络日志、收藏夹以及 cookies 记录)来建立和更新用户兴趣向量. ③ 通过用户对推荐信息和产品的反馈来修正用户偏好. 建立用户兴趣模型的框架，见图 3.

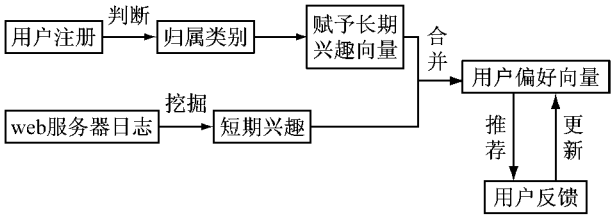

图 3　建立用户偏好模型框架

Fig. 3　Modeling frame of users preference

3.1　基于聚类分析的用户长期兴趣偏好构建

在电子商务网站上，无论用户要浏览网站信息还是要购买产品，都需要用户首先注册账号，目的在于备份用户个人基本信息，进一步了解用户需求，为用户提供满意度高的个性化服务. 其中用户注册时所需提供的信息一般有：性别、年龄、学历、职业、收入、兴趣爱好. 其中几个重要个人因素决定着用户的长期兴趣及偏好. 下面具体说明基于聚类分析构建用户长期兴趣.

3.1.1　注册信息向量的建立

注册信息一般包括：性别、年龄、学历、所在城市、职业、收入、兴趣爱好. 此时，可建立用户基本信息向量 $\boldsymbol{X}$ =(性别，年龄，学历，所在城市，职业，收入，兴趣爱好)，即 $\boldsymbol{X}=(x_1,x_2,\cdots,x_7)$，其中 $x_j(j=1,2,\cdots,7)$为向量分量的数值表示. 例如，对于性别，1 为男，0 为女.

3.1.2　基于 K-MEANS 算法的用户聚类

在用户基本信息向量的基础上，采用无监督学习 K-MEANS 聚类算法，可把用户聚类为 K 类稳定客户集合. 具体过程如下：设数据点的集合 $D=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n)$，这里代表稳定客户集合，其中 $\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{i7}),i=1,2,\cdots,n$. 将其分为 K 个组 $C_1,C_2,\cdots,C_K$，则有下面性质：① $C_i\neq\varnothing,i=1,2,\cdots,K$；② $C_i\cap C_j=\varnothing$且$\bigcup_{i=1}^{K}C_i=D,i=1,2,\cdots,K,j=1,2,\cdots,K,i\neq j$. 在众多聚类方法中，K-MEANS 算法的简洁和效率使其成为最为广泛使用的算法. 具体算法如下：

步骤 1：令 $N=1$，从 n 个点集合$(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n)$中随机选取 K 个点$(\boldsymbol{Z}_1(N),\boldsymbol{Z}_2(N),\cdots,\boldsymbol{Z}_K(N))$作

为 K 个簇的中心.

步骤 2:当且仅当满足 $\|\boldsymbol{X}_i-\boldsymbol{Z}_j\|<\|\boldsymbol{X}_i-\boldsymbol{Z}_q\|(q=1,2,\cdots,K,且\ q\neq j)$,则将 $\boldsymbol{X}_i(i=1,2,\cdots,n)$归入簇 $C_j(j=1,2,\cdots,K)$.

步骤 3:计算得到簇的新中心点 $\boldsymbol{Z}_1(N+1)$,$\boldsymbol{Z}_2(N+1),\cdots,\boldsymbol{Z}_K(N+1)$,计算公式为

$$\boldsymbol{Z}_i(N+1)=\frac{1}{n_i}\sum_{\boldsymbol{X}_j^{(i)}\in C_i}\boldsymbol{X}_j^{(i)},\ i=1,2,\cdots,K$$

式中:n_i 是处于簇 C_i 的点的数量,且令平均误差准则函数

$$E(N+1)=\sum_{j=1}^{K}\left(\sum_{\boldsymbol{X}_j^{(i)}\in C_i}|\boldsymbol{X}_j^{(i)}-\boldsymbol{Z}_i(N)|^2/\frac{1}{n_i}\right)$$

步骤 4:给定的算法精度 ξ,如果 $|E(N+1)-E(N)|<\xi$,则算法结束,否则 $N=N+1$,返回步骤 2 继续.

3.1.3 聚类用户的共同兴趣提取

利用 K-MEANS 算法,客户训练样本集被聚为 K 类,每个注册用户都归为 K 类中的某一类. 每类兴趣也采用以属性值词汇、权值、类别为元素的向量来表示,作为每一类用户的总体特征.

步骤 1:建立用户兴趣池 U_{kj},表示属于第 k 类的 j 用户的兴趣原子集合. 兴趣池中每个元素被称为原子,每个原子由四元组的形式表示:<F- words, Value, Data, Class>,其具体含义为:① F-words. 从用户浏览网页文档或从购买产品中抽取的属性值词汇,即树形结构最底层的词汇. ② Value. F-words 的原始权值,是网站一次推荐的用户感兴趣的网页文档中或所购买的产品中出现的 F-words 次数与 F-words 的全局权重的乘积. ③ Data. 是一个时间戳,表示创建该四元组原子的日期,同时也是用户产生此次网络行为的时间. ④ Class. 是 F-words 所属的类别,即树形结构中最上层词汇. 兴趣池是用户偏好的基本兴趣来源,它的更新对获得用户即时兴趣显得尤其重要.

步骤 2:把属于第 k 类($k=1,2,\cdots,K$)训练样本中稳定用户的兴趣原子聚在一起,生成一个类兴趣池 $U_i=\sum_{j=1}^{n_i}U_{ij}$,$n_i$ 为属于第 i 类的用户个数.

步骤 3:确定 U_i 中每个 F-words 的最终权值. 假设对属于第 k 类用户的属性值词汇 f_j,类兴趣池中存在以下四元组:$<f_j,\ v_1,\ d_1,\ c_i>,<f_j,\ v_2,\ d_2,\ c_i>,\cdots,<f_j,\ v_n,\ d_n,\ c_i>$. 其中,$f$ 是关键字,即属性值词汇;v 是权值;d 是原子创建日期;c_i 是类别. 属于 i 类的属性值词汇 f_j 的最终权值为:$v_j=v_1+v_2+\cdots+v_n$.

在所有属性值词汇的当前实际权值计算出来后,从$<f_j,\ v_j,\ c_i>$中选择大于一定阈值的权值 v_j 作为第 k 类用户共同兴趣,即

$$\begin{aligned}\boldsymbol{P}_{\mathrm{L}}=\{&<f_1,\omega_{11},c_1>,<f_2,\omega_{21},c_1>,\cdots,<f_g,\\&\omega_{g1},c_1>,<f_1,\omega_{12},c_2>,<f_2,\omega_{22},c_2>,\cdots,\\&<f_h,\omega_{h2},c_2>,\cdots,<f_j,\omega_{ji},c_i>\}\end{aligned}\tag{4}$$

式中:g,h,j 为提取的每一类兴趣的属性值词汇个数,$i=1,2,\cdots,K$ 为兴趣类别,此时的 ω_{ji} 是 v_j 标准化后的权值,且 $\sum\omega_{ji}=1$.

3.2 基于 web 日志挖掘的短暂兴趣偏好构建

web 日志挖掘是通过对 web 服务器日志中大量的用户访问记录深入分析,发现用户的访问模式和兴趣爱好等有趣、新颖、潜在有用的以及可理解的未知信息和知识,用于分析站点的使用情况,从而辅助管理和支持决策. 当前,web 日志挖掘主要被用于个性化服务与定制,改进系统性能和结构,站点修改,商业智能以及 web 特征描述等诸多领域[9-12]. web 日志挖掘包括三个阶段:数据收集与预处理、模式发现、模式分析[13].

3.2.1 web 日志的数据预处理

web 日志的数据预处理分为三个部分:

(1) 数据清理. web 日志中包含一些对数据挖掘不重要的数据,需要删减和整合.

(2) 用户识别. 最常用的识别方式是使用客户端的 cookies 信息、IP 地址.

(3) 事务识别. 事务标识可以理解为用户一次访问的目的. 用户访问网站可能有多个目的,例如访问携程网,先查看上海到北京的机票,再关注北京的住房信息,最后退出携程网,那么该用户访问携程网的过程就可以划分为两个事务.

3.2.2 web 日志挖掘的数据建模

(1) 建立用户事务矩阵

预处理后,得到页面访问集 $\boldsymbol{P}=\{p_1,\ p_2,\cdots,p_n\}$和用户事务集 $\boldsymbol{T}=\{t_1,\ t_2,\cdots,\ t_m\}$,$\boldsymbol{T}$ 中元素均为 $\boldsymbol{P}$ 的子集. 将 m 个事务作为一个长度为 l 的有序对序列

$$\begin{aligned}t_j=<&(p_{j1}^{t_j},\ \omega(p_{j1}^{t_j})),(p_{j2}^{t_j},\ \omega(p_{j2}^{t_j})),\cdots,(p_{jl}^{t_j},\\&\omega(p_{jl}^{t_j}))>,\quad j=1,2,\cdots,m\end{aligned}$$

式中:$\omega(p_{ji}^{t_j})$是事务 t_j 中的页面访问 $p_i^{t_j}$ 的权重,表示该页面的重要性. 权重采用布尔型,即一个页面访问在事务中出现,则 $\omega(p_{ji}^{t_j})=1$,否则 $\omega(p_{ji}^{t_j})=0$.

页面在事务中的顺序不重要,故可以将每个用户事务表示成一个 n 维空间的页面访问向量,即

$$t_j = (\omega_{p_1}^{t_j}, \omega_{p_2}^{t_j}, \cdots, \omega_{p_n}^{t_j})$$

若 p_k 在事务 t_j 中出现，则 $\omega_{p_k}^{t_j}=1(k=1,2,\cdots,n)$，否则 $\omega_{p_k}^{t_j}=0$.

用户的每一次会话都可能包含多项事务，所以需要建立以单个用户为单位的页面访问矩阵，即把规定时间内(如 1 周)属于同一个用户的事务合并起来，表示成 n 维空间的页面访问向量 $t^*=(\omega_{p_1}^{t^*}, \omega_{p_2}^{t^*}, \cdots, \omega_{p_n}^{t^*})$，$\omega_{p_j}^{t^*}$ 为某个用户多个事务叠加后的值. 例如，p_j 在用户 A 的事务中出现 3 次，那么 $\omega_{p_j}^{t*} = 3\Big/\sum_{j=1}^{n}\omega_{p_j}^{t*}$. 所有用户事务集合可以表示成 $m\times n$ 的用户页面访问矩阵(也称为事务矩阵)，用 $\boldsymbol{U}$ 表示.

(2) 建立页面访问特性矩阵

在 web 日志上的数据还包括内容数据，而且这些数据大多数情况下由文字材料和图片组成，另外还包括嵌入在网站或单独页面中的含语义的或结构化的元数据，例如描述性关键字、文档属性、语义标签或者 HTTP 变量. 所以，可以利用这些信息抽取每个页面访问 $\boldsymbol{P}$ 的属性值词汇，属性值词汇为图 2 所示的属性值词汇本体库中的词汇. 它可以用一个 r 维特性向量来表示，其中 r 是属性值词汇的数量，表示如下：

$$\boldsymbol{F} = (\omega_{c_1}^{p}(f_1), \omega_{c_2}^{p}(f_2), \cdots, \omega_{c_i}^{p}(f_r)) \quad (5)$$

其中，$\omega_{c_i}^{p}(f_j)(1\leqslant j\leqslant r)$是页面访问 $\boldsymbol{P}$ 中的第 j 个属于 $c_i(1\leqslant i\leqslant K)$类属性值词汇的权重，此时权重为该属性值词汇出现的次数. 此时，对于页面访问的所有集合就有了一个 $n\times r$ 页面访问特性矩阵 $\boldsymbol{F}$.

(3) 建立内容增强型事务矩阵

将事务中的页面访问映射到一个或多个内容特性上，也就是使用以单个用户为单位的页面访问矩阵 $\boldsymbol{U}$ 和页面访问特性矩阵 $\boldsymbol{F}$ 的乘积表示，形成新的矩阵 $\boldsymbol{E}=\{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n\}$，其中 $\boldsymbol{e}_i$ 是特性空间上的 r 维向量. 因此，用户的一次会话可以被表示成一个内容特性向量.

3.2.3　建立短暂兴趣偏好

通过上述过程，计算出每个用户对于不同类别的产品属性值词汇的偏好权值. 由于需要提取用户的短暂兴趣偏好，而且短暂兴趣对用户当前的网络行为产生较大的影响，所以在实际中通常采用近期(如最近 15 天)的 web 日志来对客户短暂兴趣偏好进行提取. 最终得到用户的短暂兴趣偏好

$$\begin{aligned}\boldsymbol{P}_S=\{&<k_1,\omega_{11},c_1>,<k_2,\omega_{21},c_1>,\cdots,<k_g,\omega_{g1},c_1>,<k_1,\omega_{12},c_2>,<k_2,\omega_{22},c_2>,\cdots,<k_h,\omega_{h2},c_2>,\cdots,<k_i,\omega_{ij},c_j>\}\end{aligned} \quad (6)$$

其中：g,h,i 为提取的每一类兴趣的产品属性值词汇个数，$j=1,2,\cdots,N$ 为兴趣类别标号，且 $\sum\omega_{ij}=1$.

4　实例分析

4.1　实验准备

4.1.1　实验条件

硬件配置：ASUS 电脑一台，主频 2.30 GHz，内存 2 G，硬盘容量 160 G.

软件配置：Windows 2007 Professional，Microsoft Internet Explorer 8.0，Microsoft Excel 2007 及其他.

4.1.2　数据集

实验采用问卷调查中 147 个样本的数据，即由淘宝网 147 位用户所提供的用户注册个人信息，这些信息包括：性别、年龄、婚姻状况、学历、职业、收入. 各个兴趣组分别是：服装、鞋、运动健身、珠宝手表、数码产品、家用电器、美容护发、母婴用品、美食特产、生活服务、图书杂志、日用百货、汽车车品. 由于淘宝商业网站的服务器日志数据难以获得，于是选取每个用户提供最近 7 天的客户端日志. 生成用户偏好之后，为用户推荐产品，收集用户的产品评分表.

4.1.3　实验标准

电子商务过滤系统尚无统一的评价方法，但大多数文献和过滤系统都采用信息获取中的评估标准，即查准率(σ). 查准率是检测产品之中真正符合用户意图的产品所占的比率，定义如下：

$$\sigma = \frac{l}{e} \quad (7)$$

式中：l 为符合用户兴趣的、正确的产品数，e 为系统实际推荐出的产品数.

除了查准率，用户体验也非常重要，即用户在获得一系列推荐后需要花费多长时间才能真正确定拟购的产品. 显然，这一决策时间越短，说明用户的体验感越好.

4.2　实验过程解析

4.2.1　用户注册信息与构建用户长期兴趣

(1) 利用 4.1.2 数据集，通过 SPSS(statistical product and service solutions)的双变量相关分析(Pearson 系数)发现，不同的兴趣类别与不同的个人因素显著相关，如表 1 所示.

(2) 利用上面分析结果，构建用户个人兴趣的判别机制，如表 2 所示.

表1 用户个人信息与用户兴趣偏好相关性

Tab.1 The correlation of the users' personal information and preference

个人信息	服装	鞋	运动健身	珠宝手表	数码产品	家用电器	美容护发
性别	0.226**		−0.282**	0.165*	−0.430**		0.421**
年龄							
婚姻状况							
学历	−0.222**						
职业							
收入					0.180*		

个人信息	母婴用品	美食特产	生活服务	图书杂志	日用百货	汽车用品
性别	0.165*			−0.213**		
年龄	0.415**					0.179*
婚姻状况	−0.395**					
学历						
职业	0.220**					
收入	0.179*					

注：** 表示在0.01水平(双侧)上显著相关，* 表示在0.05水平(双侧)上显著相关.

表2 用户的个人兴趣判别机制

Tab.2 The decision mechanism of the users' personal preference

个人信息			服装	鞋	运动健身	珠宝手表	数码产品	家用电器	美容护发	母婴用品	美食特产	生活服务	图书杂志	日用百货	汽车车品
性别	1	男			+		+						+		
	2	女	+			+			+	+					
年龄	1	0～10							−						−
	2	10～20							−						
	3	20～30							+						−
	4	30～40							−						+
	5	40～50							+						−
	6	50+							−						−
婚姻状况	1	结婚并且有孩子							+						
	2	结婚并且无孩子							−						
	3	单身													
学历	1	高中或中专	+												
	2	大专	+												
	3	大学本科	+												
	4	研究生	+												
	5	博士													
	6	博士后													
就业	1	学生							−						
	2	国企													
	3	外企							−						
	4	个体经营或创业者							−						
	5	自由职业者													
收入	1	低于2 000													
	2	2 000～4 000					+								
	3	4 000～6 000					+								
	4	6 000～8 000					+								
	5	8 000～10 000					+								
	6	>10 000													

注：“+”表示该个人信息与该类别正相关，“−”表示该个人信息与该类别负相关.

(3) 利用第3.1节聚类分析法，对4.1.2数据集中100个样本进行训练，建立用户兴趣池，给出用户的长期兴趣 $\boldsymbol{P}_{\mathrm{L}}$.

以用户的注册信息为 $\boldsymbol{X}$=(女，20～30，单身，研究生，学生，低于2 000)的样本为例，对上述训练结果进行验证. 根据用户注册信息 $\boldsymbol{X}$，可以得到该用户的偏好类(服装，珠宝手表，美容护发，母婴用品). 于是在淘宝网的数据库里随机抽取了100位与其偏好类相同或相似的稳定用户聚为一类. 提取该类用户的共同兴趣偏好向量为

$\boldsymbol{P}_L$=((咖啡色,0.2,颜色,服装),(西瓜红,0.1,颜色,服装),(休闲,0.3,风格,服装),(气质,0.1,风格,服装),(100～200 RMB, 0.3,价格,服装),(银,0.1,质地,珠宝),(我的美白日志,0.3,品牌,美容护发),(贝佳斯,0.1,品牌,美容护发),(0～100 RMB,0.3,价格,美容护发))

权重经过标准化后等于

$\boldsymbol{P}_L$=((咖啡色,0.11,颜色,服装),(西瓜红,0.06,颜色,服装),(休闲,0.17,风格,服装),(气质,0.06,风格,服装),(100～200 RMB, 0.17,价格,服装),(银,0.06,质地,珠宝),(我的美白日志,0.17,品牌,美容护发),(贝佳斯,0.06,品牌,美容护发),(0～100 RMB, 0.17,价格,美容护发)

取 ω=0.1,得到

$\boldsymbol{P}_L$=((咖啡色,0.11,颜色,服装),(休闲,0.17,风格,服装),(100～200 RMB, 0.17,价格,服装),(我的美白日志,0.17,品牌,美容护发),(0～100 RMB, 0.17,价格,美容护发))

4.2.2　web 日志与构建用户短期兴趣

为了获得上述 $\boldsymbol{X}$ 类别用户的 web 日志,可在用户 IE 浏览器中的历史记录中,抽取用户近 7 天之内的淘宝网站的历史记录.根据第 3.2 节所述的方法,可以获得以下用户的短期兴趣偏好向量:

$\boldsymbol{P}_S$=((za,0.33,品牌,美容护发),(洗面,0.33,用途,美容护发),(美肤宝,0.13,品牌,美容护发),(口红,0.13,用途,美容护发),(滋补,0.2,用途,美食特产),(高丽参,0.13,名称,美食特产))

经过标准化以及 ω=0.1 的处理之后得

$\boldsymbol{P}_S$=((za,0.26,品牌,美容护发),(洗面,0.26,用途,美容护发),(美肤宝,0.10,品牌,美容护发),(口红,0.10,用途,美容护发),(滋补,0.16,用途,美食特产),(高丽参,0.10,名称,美食特产))

再经过 $\boldsymbol{P}_L+\boldsymbol{P}_S=\boldsymbol{P}$ 和 ω=0.1 处理之后得

$\boldsymbol{P}$=((咖啡色,0.11,颜色,服装),(休闲,0.17,风格,服装),(100～200 RMB, 0.11,价格,服装),(我的美白日志,0.11,品牌,美容护发),(0～100 RMB, 0.11,价格,美容护发),(za,0.16,品牌,美容护发),(洗面,0.16,用途,美容护发),(滋补,0.10,用途,美食特产))

4.3　实验结果分析

为了分析本文所建立的模型的效果,把基于长短期兴趣偏好的新模型和仅基于注册信息的旧模型进行对比,分析这两种模型在“查准率”和“决策时间”上的优劣.

根据上述实验得出的结论用户兴趣 $\boldsymbol{P}$,根据用户兴趣和用户搜索关键词信息“T 恤”,为用户提供推荐服务.通过用户的反馈信息,其为用户推荐的产品 top10 中,八项是符合用户需求的,于是

$$\sigma=\frac{8}{10}=80\% \tag{8}$$

在对 100 名淘宝用户进行实验后,得到的查准率的数据如表 3 和图 4 所示.

表 3　查准率对比

Tab.3　The experimental accuracy

用户	σ	用户	σ	用户	σ	用户	σ
1	0.8	26	0.9	51	0.8	76	0.7
2	0.7	27	0.7	52	0.7	77	0.9
3	0.8	28	0.8	53	0.9	78	0.8
4	0.6	29	0.9	54	0.8	79	0.9
5	0.7	30	0.7	55	0.9	80	0.7
6	0.9	31	0.9	56	0.7	81	0.7
7	1.0	32	0.7	57	0.8	82	0.8
8	0.8	33	0.7	58	0.7	83	0.7
9	0.7	34	0.8	59	0.9	84	0.9
10	0.8	35	0.7	60	0.7	85	0.7
11	0.7	36	0.9	61	0.6	86	0.6
12	0.9	37	0.7	62	0.7	87	0.8
13	0.9	38	0.6	63	0.9	88	0.7
14	0.8	39	0.7	64	1.0	89	0.9
15	0.6	40	0.9	65	0.8	90	0.7
16	0.5	41	0.9	66	0.7	91	0.9
17	0.9	42	0.8	67	0.8	92	1.0
18	0.7	43	0.6	68	0.7	93	0.8
19	1.0	44	0.8	69	0.9	94	0.7
20	0.9	45	0.7	70	0.7	95	0.8
21	0.8	46	0.9	71	0.7	96	0.9
22	0.6	47	0.7	72	0.9	97	0.8
23	0.7	48	0.9	73	0.8	98	0.6
24	0.8	49	0.6	74	0.7	99	0.5
25	0.7	50	0.8	75	0.9	100	0.9

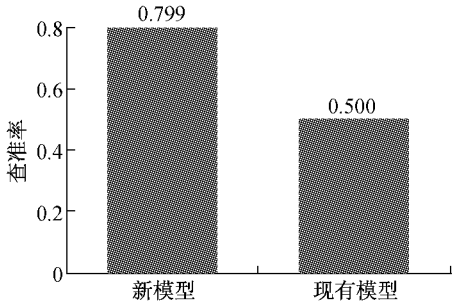

图 4　模型查准率对比

Fig.4　Comparison of model accuracy

从实验结果来看,采用基于长期兴趣和短暂兴趣的综合模型构建用户偏好可以获得更好的查准率.对比没有长短期兴趣结合模型的查准率要高出 27.9%,提高了了解用户偏好的准确率.

图 5 为新旧模型的用户决策时间的对比,可以

直观地看出在应用了新模型之后的用户决策时间有了明显缩短.现有系统背景下,在做出推荐后用户的决策时间大多在3～10 min,而在应用新的模型试验下,用户的决策时间主要集中在1～5 min.于是,可以得出该模型是有效可行的,它提高了用户的体验感.

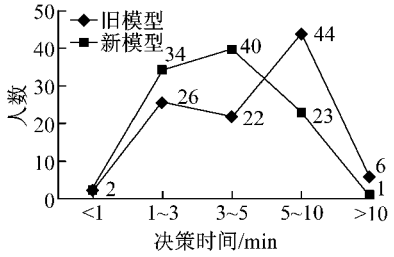

图5 新旧模型的决策时间比较

Fig.5 Comparison of decision-making time by new and old models

5 结语

本文通过对用户web使用记录的数据挖掘,在心理学研究基础上,建立了一种基于长期兴趣与短暂兴趣的用户偏好文档.该偏好文档不仅能够有效反映用户信息需求的变化,还可以充分表现出用户对不同兴趣类别的重视程度,最大深度地表现出用户的需求和偏好,并且能广泛应用于电子商务网站产品推荐或网站网页推荐.在对这一问题研究过程中尚有以下问题有待深入:① 如何利用web访问日志上的结构数据来提取用户兴趣;② 寻找更有效的用户聚类算法;③ 用户偏好文档中偏好权值的计算方法有待改进;④ 实现短暂兴趣向长期兴趣的转换.其中,利用结构数据来提取用户兴趣以及实现短暂兴趣向长期兴趣的转换是两个研究重点.

参考文献:

[1] 宗胜. 基于情境兴趣和个人兴趣的用户偏好模型研究与设计[D]. 上海:上海交通大学,2006.
ZONG Sheng. Research and design of user profile model based on situation interest and individual interest[D]. Shanghai: Shanghai Jiaotong University, 2006.

[2] 周晓兰. WEB数据挖掘中用户兴趣模型设计[J]. 湘潭师范学院学报, 2009(6): 55.
ZHOU Xiaolan. User prefence modeling and designing based on web data mining[J]. Journal of Xiangtan Normal College, 2009(6): 55.

[3] Danilowicz C, Indyka-Piasecka A. Dynamic user profiles based on boolean formulas[C]//17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems. Ottawa: [s.n.] 2004: 779-787.

[4] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]//Communications of ACM, 1975, 18(5): 613.

[5] Chen L, Sycara K. Web mate: a personal agent for browsing and searching[C]//the 2nd International Conference on Autonomous Agents and Multiagent Systerms. Minneapolis: [s.n.], 1998: 132-139.

[6] Kesorn K, Liang Z K, Poslad S. Use of granularity and coverage in a user profile model to personalize visual content retrieval[C]//2009 Second International Conference on Advances in Human-oriented and Personalized Mechanism, Technologies, and Service. Porto: [s.n.], 2009: 79-84.

[7] 邵志峰. 基于中图分类法的用户兴趣模型研究[J]. 计算机应用与软件,2007(8):86.
SHAO Zhifeng. A research on user profile model based on chinese library classification[J]. Computer Applications and Software, 2007(8):86.

[8] 吕新波. 隐式用户兴趣挖掘的研究和实现[D]. 哈尔滨:哈尔滨工业大学,2008.
LÜ Xinbo. Research and implementation of mining implicit user interest[D]. Harbin: Harbin Engineering University, 2008.

[9] Cooley R, Mobasher B, Srivastava J. Web mining: information and pattern discovery on the world wide web[C]//Proceeding of the 9th IEEE International Conference on Tools with Artificial Intelligence. Los Angeles: IEEE, 1997: 558-567.

[10] Wang John. Encyclepedia of data warehousing and mining[M]. Calgary: Idea Group Inc, 2006.

[11] Srivastava J, Cooley R, Deshpande M. Web usage mining: discovery and applications of usage pattern from web data[C]//SIGKDD Explorations. New York: Association for Computing Machinery, 2000: 12-23.

[12] 郭晓磊. 基于WEB日志挖掘的网络用户聚类研究[D]. 北京:北京邮电大学,2009.
GUO Xiaolei. Research on web users clustering based on web log mine[D]. Beijing: Beijing University of Posts and Telecommunications, 2009.

[13] Fayyad M U, Piatetsky-Shapiro G Smyth P. From data mining to knowledge discovery: an overview[C]//Advances in Kownledge Discovery and Data Mining. Menlo Park: AAAI Press, 1996: 1-34.