维普资讯 http://www.cqvip.com

基于商品特征的个性化推荐算法

李 峰,李军怀,王瑞林,张 璟
LI Feng,LI Jun-huai,WANG Rui-lin,ZHANG Jing

西安理工大学 计算机科学与工程学院,西安 710048
School of Computer Science and Engineering,Xi'an University of Technology,Xi'an 710048,China

LI Feng,LI Jun-huai,WANG Rui-lin,et al.Personalized recommendation algorithm based on product features.Computer Engineering and Applications,2007,43(17):194-197.

Abstract: In the field of personalized recommendation,current algorithms have the disadvantages of lower precision and deficiency on recommendation.This paper presents an algorithm based on the features of product,the customer's purchased logs and real-time browsing action.Firstly the content of on-line browsing is collected to deduce the purchase preference of current customer,then contrasts the purchased logs and the database of product features and analyzes them,by means of which,the preference degree of the product features and corresponding commendation reference groups can be obtained,therefore according to the similarity matrix of features entity,reference groups are recommended.Finally,integrated with the purchase preference and the former results,products are recommended to the customer.

Key words: product features;personalized recommendation;preference degrees;similarity

摘 要:针对现有个性化商品推荐算法精度不高、新商品不能及时推荐等缺点,提出了一种基于商品特征、用户购买日志及用户实时浏览行为的个性化推荐算法。算法首先根据客户的在线浏览情况获取当前客户的购买倾向,然后将客户的购买日志与商品特征数据库进行对比分析,获得客户对商品特征的偏爱度及推荐参照组,依据特征实体的相似度矩阵进行特征推荐组推荐,最后结合当前的购买倾向向客户推荐商品。

关键词:商品特征;个性化推荐;偏爱度;相似度
文章编号:1002-8331(2007)17-0194-04 **文献标识码**:A **中图分类号**:TP311

1 引言

随着互联网的不断发展,在线购物已经成为许多企业的主要销售方式之一。当用户在线浏览 Web 站点的时候,进行商品推荐已经成为了一种主要的促销方式。

目前应用比较广泛的个性化推荐算法主要有:基于内容的个性化推荐[1,2]、协同过滤个性化推荐[3]以及基于 Web 日志的个性化推荐。随着网站内容的不断细化以及客户对推荐内容要求的不断提高,上述算法的不足日益显现,比如推荐精度不高、推荐效率低、新上市或购买率较低的商品不能及时地推荐给客户等。如何满足客户的需求,向他们推荐符合其购买习惯和爱好的商品已经成为当前推荐算法的首要问题之一。

本文针对上面各种算法的不足,引入了商品特征的概念,通过将商品特征数据库、用户的购买日志以及用户的在线浏览对象三者相联系,提出了一种新的推荐算法。该算法可以弥补现行推荐算法推荐精度不高、推荐效率低、新上市或购买率较低的商品不能及时地推荐给客户等不足,推荐结果有了明显的改善,更加符合客户的购买倾向。其主要优点如下:

(1)给用户提供个性化的服务

与以往的推荐不同,本算法是以单一的客户个体作为推荐对象的。对客户的购买日志进行分析,了解用户的兴趣、偏好等,并以此作为基础信息向用户推荐符合其兴趣的商品。实现了针对不同客户不同推荐的个性化推荐。

(2)提高推荐的精确度

通过对客户购买日志的分析,了解客户的购买偏好,根据客户的购买习惯进行推荐,提高了推荐商品的精度。

(3)推荐更加符合客户购买倾向的产品

在了解用户的兴趣的同时,分析其浏览内容,进一步了解其购买倾向,通过其购买倾向和客户的爱好向其推荐符合客户购买倾向的商品。

2 算法模型

2.1 算法描述

首先,根据客户的在线浏览情况获取当前客户的购买倾向,为推荐做准备。然后将客户的购买日志与商品特征数据库进行对比分析,获得客户对商品特征的偏爱度及推荐参照组,结合商品特征数据库及特征的相似度矩阵进行推荐组推荐,最

基金项目:国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2002AA414060);陕西省自然科学基金(the Natural Science Foundation of Shaanxi Province of China under Grant No.2005F05)。

作者简介:李峰,男,硕士研究生,主要研究 Web 应用;李军怀,男,副教授,主要研究方向为分布式计算,CSCW;王瑞林,硕士研究生,主要研究 Web 应用;张璟,教授,博士生导师,主要研究方向为 Internet 技术及应用。

维普资讯 http://www.cqvip.com

后结合当前的购买倾向进行商品的推荐。推荐算法的结构如图1所示。

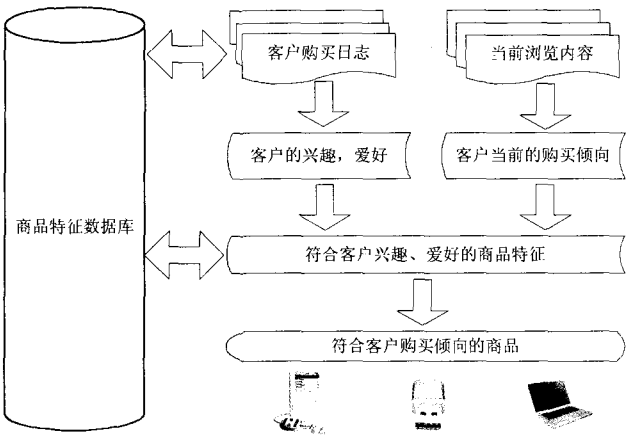

图1　个性化推荐算法结构

2.2　产品特征的描述

本算法建立在商品的客观和可识别的特征之上，这些特征和记录在基本特征数据库中的特征是匹配的，比如价钱和商标等。为了能够进行产品特征模型的描述，建立商品的特征数据库，可以用向量表示，如下：

$P(m)=(f_{11},\cdots f_{1k},\cdots,f_{ik},\cdots f_{ij}),m=1,\cdots,M$

M 是某一实体产品（比如手机）m 的总数量，i 是产品特征的编号（比如颜色），j 表示为特征 i 的第 j 个特征值（比如银色）。参数 k 意味着每一个产品特征的特征域是不固定的，每一个特征值用二进制表示，例如，如果商品有该特征，特征值为1，否则为0[4]。

2.3　用户当前的购买倾向

因为用户在浏览商品的时候总是对自己关心的内容进行高频度的点击和浏览。因此，我们认为用户当前浏览内容中用户点击较多的商品种类为当前用户的潜在购买倾向。商品的种类也是一种商品特征，所以，用户当前的购买倾向也是商品特征的一个特征实体。对于种类特征的特征实体集 $K=\{k_1,k_2,\cdots,k_n\}$，客户当前的购买倾向 K_b 定义如下：

$K_b=k_i$，k_i 为客户点击数量最多的特征实体

2.4　用户对商品特征的偏爱度

根据客户某一时间段内的购买日志，分析其对商品特征的偏爱程度，然后根据客户的偏爱程度对这些特征赋相应的值，即当前客户对于特征的偏爱度。

偏爱度的定义如下：将客户某时间段内购买商品的部分主要商品特征进行统计，特征 i 的偏爱度 C_i 为客户购买商品中包含特征 i 的数量与购买商品中总特征数量的比值。计算公式如下：

$$C_i=\frac{m_i}{\sum_{j=1}^{n}m_j} \tag{1}$$

其中，n 表示统计的商品特征的数量（例如商品的特征中的品牌、颜色等），m 为某时间段内客户购买商品中包含某特征的数量。

2.5　商品特征实体的相似度矩阵

2.5.1　商品特征实体的相似度

对于某一特征的特征实体之间的相似度，通过用户对该特征的评价来进行相似度的计算。用户集 $U=\{u_1,u_2,\cdots,u_m\}$，特征实体集 $I=\{i_1,i_2,\cdots,i_n\}$，用户对各实体的评分如矩阵 $\boldsymbol{E}$。

表1　用户评分矩阵 $\boldsymbol{E}$

	i_1	i_2	$\cdots$	i_{n-1}	i_n
u_1	V_{11}	V_{12}	$\cdots$	$V_{1(n-1)}$	V_{1n}
u_2	V_{12}	V_{22}	$\cdots$	$V_{2(n-1)}$	V_{2n}
$\cdots$	$\cdots$	$\cdots$	V_{ij}	$\cdots$	$\cdots$
u_{m-1}	$V_{(m-1)1}$	$V_{(m-1)2}$	$\cdots$	$V_{(m-1)(n-1)}$	$V_{(m-1)n}$
u_m	V_{m1}	V_{m2}	$\cdots$	$V_{m(n-1)}$	V_{mn}

通过不同的相似性度量方法计算特征实体 i 和 j 之间的相似性，记为 $s(i,j)$，即为实体特征的相似度。根据用户评分矩阵的稀疏程度，可以采用不同的相似性计算方法，达到尽量提高计算精度的目的。计算相似性的方法主要包括余弦相似性、相关相似性以及修正的余弦相似性。

余弦相似性：项目评分看作为 m 维用户空间上的向量，如果用户对特征实体没有进行评分，则将用户对该对特征实体的评分设为0，对特征实体间的相似性通过向量间的余弦夹角度量。设客户对特征实体 i 和特征实体 j 在 m 维客户空间上的评分分别表示为向量 $\boldsymbol{i},\boldsymbol{j}$，则特征实体 i 和特征实体 j 之间的相似性 $s(i,j)$ 为：

$$s(i,j)=\cos(\boldsymbol{i},\boldsymbol{j})=\frac{\boldsymbol{i}\,\boldsymbol{j}}{\|\boldsymbol{i}\|*\|\boldsymbol{j}\|} \tag{2}$$

其中，分子为两个特征实体评分向量的内积，分母为两个特征实体评分向量模的乘积。

相关相似性：设对特征实体 i 和特征实体 j 共同评分过的用户集合用 U_{ij} 表示，则特征实体 i 和特征实体 j 之间的相似性 $s(i,j)$ 通过 Pearson 相关系数度量：

$$s(i,j)=\frac{\sum_{c\in U_{ij}}(V_{c,i}-\bar{V}_i)(V_{c,j}-\bar{V}_j)}{\sum_{c\in U_{ij}}(V_{c,i}-\bar{V}_i)^2\sqrt{\sum_{c\in U_{ij}}(V_{c,j}-\bar{V}_j)^2}} \tag{3}$$

其中，$V_{c,i}$ 表示客户 c 对实体特征 i 的评分，V_i 和 V_j 分别表示对实体特征 i 和实体特征 j 的平均评分。

修正的余弦相似性：在余弦相似性度量方法中没有考虑不同用户的评分尺度问题，修正的余弦相似性度量方法通过减去客户对项目的平均评分改善上述缺陷，设对特征实体 i 和特征实体 j 共同评分过的用户集合用 U_{ij} 表示，U_i 和 U_j 分别表示特征实体 i 和特征实体 j 评分过的用户集合，则特征实体 i 和特征实体 j 之间的相似性 $s(i,j)$ 为：

$$s(i,j)=\frac{\sum_{c\in U_{ij}}(V_{c,i}-\bar{V}_c)(V_{c,j}-\bar{V}_c)}{\sqrt{\sum_{c\in U_i}(V_{c,i}-\bar{V}_c)^2}\sqrt{\sum_{c\in U_j}(V_{c,j}-\bar{V}_c)^2}} \tag{4}$$

其中，$V_{c,i}$ 表示客户 c 对特征实体 i 的评分，V_c 表示客户 c 对特征实体的平均评分。

2.5.2　商品特征实体的相似度矩阵

对某一特征的特征实体集 $I=\{i_1,i_2,\cdots,i_n\}$，用矩阵 $\boldsymbol{S}$ 表示其相似度矩阵。相似度值由2.5.1相关公式进行计算得出。

通过观察，不难发现矩阵 $\boldsymbol{S}$ 有如下特征：

（1）$S_{(i,j)}=s_{(j,i)}$

（2）$S_{(i,i)}=1$

因此，在计算时可以直接只考虑上三角矩阵或者下三角矩阵，有利于简化计算。

维普资讯 http://www.cqvip.com

表2 商品特征实体相似度矩阵 S

I	i_1	i_2	…	i_{n-1}	i_n
i_1	S_{11}	S_{12}	…	$S_{1(n-1)}$	S_{1n}
i_2	S_{12}	S_{22}	…	$S_{2(n-1)}$	S_{2n}
…	…	…	S_{ii}	…	…
i_{n-1}	$S_{(n-1)1}$	$S_{(n-1)2}$	…	$S_{(n-1)(n-1)}$	$S_{(n-1)n}$
i_1	S_{n1}	S_{n2}	…	$S_{n(n-1)}$	S_{nn}

2.6 商品特征推荐组的推荐值

用各特征相似矩阵中的相似度乘以推荐组中该商品特征的偏爱度,然后进行相加,即得到该推荐组的推荐值,以 RV 表示。计算公式如下:

$$RV_{i,j}=\sum_{i=1}^{n} S_{m,n}*C_m \tag{5}$$

其中,i 表示第 i 组参照组,j 表示第 j 组推荐组,$S_{m,n}$ 为某特征的特征实体 m 与特征实体 n 的相似度,即为商品特征实体相似度矩阵中的值,C_m 为客户对特征 m 的偏爱度,n 为所取主要特征的数量。

3 基于商品特征的个性化推荐算法

3.1 获取用户当前购买倾向及特征偏爱度

通过用户在网站上面对商品种类的点击次数,取点击频度最高的种类特证的特征实体作为用户当前的购买倾向。比如对于种类特征的实体特征集 K={PC,DV,MP3,Notebook,Cell-Phone,… },根据对当前客户的浏览统计,得出当前客户的购买倾向为:$K_b=k_i$=Mp3。

通过对客户购买日志的分析,由公式(1)计算出各特征的偏爱度 C_i,然后根据偏爱度的值进行排序,出于对计算精度的考虑,在实际推荐时可以只取偏爱度排在前 n 位的特征用于推荐。比如通过分析客户某时间段的购买日志并按照公式(1)进行计算后,按照偏爱度值进行计算排序的结果为:品牌,颜色,类别,重量…,相应特征的偏爱度依次为 0.286,0.191,0.132,0.096,…。

3.2 获取特征的推荐参照组

在得到客户对特征的偏爱度后,根据偏爱度进行横向排序,通过对客户购买日志的进一步分析,对每一个特征的实体特征根据客户的选择次数进行纵向排序,每一组特征实体定义为一组基础推荐参照组。选取前 n 组基础参照组作为推荐依据,这 n 组特征实体本身也是 n 组推荐参照组,为了进一步扩大推荐范围,可以将每一组相似度最高的特征实体和其他组偏爱度排在第二位以后的特征实体进行特征交叉后组成新的特征推荐组。如果取前 n 组,理论上可以得到 n^2 组推荐参照组。

例如,根据偏爱度值排序后的特征依次为:品牌,颜色,类别,重量…。根据特征排序对各特征进行纵向排序后,发现客户购买的品牌中三星最多,联想次之,颜色中蓝色最多,红色次之等等,当 n=2 时,即取前两组基础推荐参照组,可以得到 2^2=4 组推荐参照组。第一组和第二组推荐参照组为基础推荐参照组本身,分别为:

1 三星,蓝色

2 联想,红色…。

进行特征交叉后,可以得到第三组和第四组推荐参照组,分别如下:

3 三星,红色…;

4 联想,蓝色…。

3.3 推荐推荐组并计算其推荐值

得到推荐参照组后,就可以进行推荐并计算推荐组的推荐值了。

以参照组为依据,通过各特征的相似矩阵来进行推荐组的推荐。因为各特征的实体本身在相似度矩阵中相似度最高(相似度为 1),所以参照组本身是得到的第一组推荐组,然后在相似度矩阵中找到相似度次之的特征,可以得到第 2 组、第 3 组、第 i 组等等,考虑到推荐精度,这里 i 可以根据实际情况进行限制。然后依据推荐值的计算公式进行相关计算,即可得到该推荐组的推荐值。

以第一组参照组为例,首先因为各特征之间相同的实体特征的相似度为 1,所以得到的第一组推荐组就是第一组参照组本身,即三星,蓝色…,根据公式(5)计算其推荐值为

RV_{11}=1*0.286+1*0.191+…

接着,针对第一组参照组,在各特征的相似度矩阵中找相似度次之的特征,比如在品牌相似度矩阵中,通过对比查找后发现与三星次之的是 LG,其值为 0.83,在颜色相似度矩阵中,与蓝色次之的是银色,其值为 0.92,其它特征依此类推,于是通过第一组参照组得到其第二组推荐组,即 LG,银色…,根据公式(5)计算其推荐值为:

RV_{12}==0.83*0.286+0.92*0.191+…

类似地,可以通过第一组参照组得到其它的推荐组及其推荐值 $RV_{1,i}$。

同样的方法,可以根据第 i 组参照组得到其推荐组及其推荐值 $RV_{i,j}$。

3.4 根据推荐组进行商品推荐

根据上面得到的特征推荐组,对所有推荐组按照其推荐值进行排序,取排序在前 N 位的特征推荐组作为对用户进行推荐的推荐依据。此时将用户当前的购买倾向(即种类特征的特征实体)与推荐的 N 组特征推荐组进行组合,得到 N 种商品,然后根据得到的 N 种商品去商品特征书库一一进行匹配,如果该商品存在,推荐之,如果不存在商品实体,从推荐值在 N 位以后推荐组依次进行填充,然后继续推荐和匹配,直至填充到推荐值排序在最后一位的推荐组。N 值需要根据实际情况进行设置。

比如 RV_{11} 排在第一位,此时计算得出客户当前购买倾向的种类实体为 MP3,于是根据组合后得到的商品为,三星 MP3,蓝色,…。然后根据特征去商品特征库进行匹配,如果有符合特征条件的商品,推荐,如果该商品不存在,转而进行下一组的推荐。

4 试验及结果分析

4.1 试验方法

本文以某在线购物站点为对象,通过本推荐算法对客户进行商品的推荐。为了对推荐结果进行分析,引入推荐命中率这一概念,定义如下:

$$RP=\frac{\sum_{i=1}^{n} ClickNum_i}{\sum_{i=1}^{n} RecItems_i}$$

其中,i 为第 i 次推荐,ClickNum 代表客户在第 i 次推荐中点击

维普资讯 http://www.cqvip.com

所推荐商品的数量,RecItems 表示第 i 次推荐推荐的数量。n 为对当前客户推荐的次数。

为了方便统计数据,对各部分取如下参数:

偏爱度:取偏爱度排序前 10 位的特征作为推荐依据。

推荐参照组:为了进行对比,取基础推荐参照组 n=2 和 n=3,即可分别得到 2^2=4 和 3^2=9 组推荐参照组。

推荐组:考虑到推荐的精确度,针对每一组推荐参照组,取前 5 组进行推荐,其中包括推荐参照组本身。

商品推荐:因为根据上述参数可分别获取 20 种和 45 种推荐商品,这里取前 10 种商品进行推荐。

推荐命中率:为了提高测试数据的精度,这里取 n=10,即对每个用户推荐 10 次后进行命中率的计算。

4.2 试验结果及分析

根据客户不同数量的购买纪录进行推荐后,统计客户的推荐命中率,得到如图 2 的推荐命中率曲线。

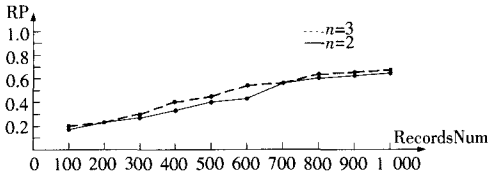

图 2 推荐命中率曲线

在整个推荐试验中,由于是按照特征进行推荐,同时又是将客户的浏览行为作为推荐依据的一部分,推荐的结果解决了传统推荐的一些缺点,比如新商品可以及时推荐给客户,推荐精度较高等。由图 2 可见,随着分析的购买日志数量的增加,推荐的精度在不断地提高,但是超过 700 条之后,推荐命中率的增加微乎其微,这可以作为进行推荐参数设置的一个参考;另外,伴随着基础推荐参照组的增多,推荐的精度也有所提高,但是推荐命中率的增幅也比较小,分析其原因,是由于基础推荐参照组的增多,导致了特征范围的扩大,在一定程度上提高了推荐的精度,但同时也扩大了商品类型,在一定程度上又影响了推荐的精度。

5 小结

随着 Internet 的不断发展,智能化的商品推荐已经成为企业销售中必不可少的一种推销方法。本文提出的推荐策略以客户的购买日志为出发点,对客户的购买行为进行分析,最后推荐给用户符合自己兴趣的商品,缩小了推荐的范围,提高了推荐效率,更加智能化地满足了用户的需求,使客户对商业站点的满意度增加,从而达到增加站点产品销量的目的。但是由于算法是对客户购买行为进行分析,在提高推荐效率的同时,推荐的响应时间不够理想,为此,我们将不断探索,对算法进行优化,在提高推荐效率的同时保证合理的响应时间。

(收稿日期:2007 年 1 月)

参考文献:

[1] Middleton S E,Shadbolt N R,Roure D C.Ontological userProfiling in recommender systems[J].ACM Transactions on Information Systems,2004:54-88.

[2] Sarwar B,Karypis G,Konstan J,et al.Item-based collaborative filtering recommendation algorithms[C]//Proceedings of the Tenth International Conference on World Wide Web.ACM Press,2001:285-295.

[3] 邓爱林,左子叶,朱扬勇.基于项目聚类的协同过滤推荐算法[J].小型微型计算机系统,2004,25(9):1665-1670.

[4] Weng S,Liu M.Feature-based recommendations for one-to-one marketing[J].Expert Systems with Application,2004,26:493-508.

[5] Han EuiHong(Sam),Karypis G.Feature-based recommendation system[C]//Proceedings of the 14th ACM International Conference on Information and Knowledge Management,2005:446-452.

[6] Yoo Jungsoon,Gervasio M,Langley P.Personalized trading recommendation system[J].The ACM Digital Library,2003,1:446-452.

[7] Wang Pei.Recommendation based on personal preference[M]//Computational Web Intelligence.Singapore:World Scientific Publishing Company,2004:101-115.

[8] Coyle L,Cunningham P.Improving recommendation ranking by learning personal feature weights [C]//7th European Conference,EC-CBR 2004,Madrid,Spain,August 30-September 2,2004:560-572.

[9] 曾春,邢春晓,周立柱.个性化服务技术综述[J].软件学报,2002,13(10):1953-1961.

(上接175 页)

5 结论

本文提出的基于偏最小二乘的支持向量机分类算法,经实验证明具有很好的分类效果。PLS 在样本属性的约简过程中,不仅考虑属性之间的相关信息进行降维,而且充分利用了类信息,因此提取的综合成分更能代表样本的真实情况,再用 SVM 进行分类时,不仅减少了支持向量的数目,而且当样本属性较多时,可以提高一定的识别率,分类效果更佳。笔者在此只是将偏最小二乘的回归方法作为一种降维方法对样本数据进行预处理,为利用支持向量机分类作准备,这种应用显然还不能充分利用 PLS 优良的性能,如何将 PLS 与 SVM 更加有效的结合,将是下一步还要继续研究的方向。

(收稿日期:2006 年 11 月)

参考文献:

[1] 张学工.关于统计学习理论与支持向量机[J].自动化学报,2000,26(1):32-42.

[2] 许建华,张学工,李衍达.支持向量机的新发展[J].控制与决策,2004,19(5):481-484.

[3] 赵广社,张希仁.基于主成分分析的支持向量机分类方法研究[J].计算机工程与应用,2004,40(3):37-38.

[4] 王惠文.偏最小二乘回归方法及其应用[M].北京:国防工业出版社,1999.

[5] 邓乃扬,田英杰.数据挖掘中的新方法——支持向量机[M].北京:科学出版社,2004.

[6] Cristianini N,Shawe-Taylor J.支持向量机导论[M].北京:电子工业出版社,2004.

[7] Bastien P,Vinzi V E,Tenenhaus M.PLS generalized linear regression[J].Computational Statistics & Data Analysis,2005,48:17-46.

[8] Angulo C,Ruiz F J,Gonzalez L,et al.Multi-classification by using tri-class SVM[J].Neural Processing Letters,2006,23:89-101.