基于内容过滤的农资电子商务推荐系统①

徐玲玲 [1,2], 孙丙宇 [1,2], 方 薇 [2]

[1](中国科学技术大学 信息科学技术学院, 合肥 230039)
[2](中国科学院合肥智能机械研究所, 合肥 230031)

摘 要: 随着农业信息化的发展, 农业类网站已经成为农业用户、合作社和农资公司等获取信息的重要渠道. 结合中国现代化农资经营电子商务平台, 提出了基于内容过滤的推荐技术, 采用四元组构建用户偏好模型, 引入遗忘因子挖掘和更新偏好模型, 并根据产品模型和用户偏好模型的相似度向用户推荐产品. 实验结果表明, 基于内容过滤的推荐算法可使农资电子商务平台的产品浏览率和购买率得到提高.

关键词: 农资电子商务; 推荐系统; 内容过滤; 用户兴趣; 遗忘因子

Content-Based Filtering Recommendation System in Agricultural E-commerce

XU Ling-Ling[1,2], SUN Bing-Yu[1,2], FANG Wei[2]

[1](School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China)
[2](Institute of intelligent machine, Chinese Academy of Sciences, Hefei 230031, China)

Abstract: As the agriculture information developing, agricultural websites have become an important channel for accessing information for agricultural users, cooperatives and agricultural companies. Combined with Chinese modern agricultural business e-commerce platform, content-based filtering recommendation technology is proposed, adopting four-tuple to construct user interest model, introducing forgetting factor to mining and update user preference, and generating recommendations depending on the similarity of product model and preference model. By the practical tests, the results show that Content-based filtering recommendation algorithm can effectively improve the purchase rate.

Key words: agricultural e-commerce; recommendation system; content-based filtering; user preference; forgetting factor

1 引言

个性化推荐系统是互联网信息过载问题的产物, 它根据用户兴趣、爱好、习惯, 以及各个用户之间的相关性向特定用户在线推荐相关内容, 提供浏览建议, 为用户进行个性化服务[1]. 目前成功的个性化推荐系统包括亚马逊、CDNow.com、Barnes & Noble.com、MovieFinder.com 等[2].

近年来, 推荐算法主要有基于内容过滤和协同过滤. 基于内容过滤推荐技术的基本思想是: 根据用户行为挖掘出用户兴趣, 提取用户的兴趣特征, 形成特征向量, 通过加权的方式使具有较高区分度的特征具有较大的权重, 形成用户的兴趣模型. 当需要对某个用户进行推荐时, 把该用户的用户兴趣模型协同所有项目的特征矩阵进行相似度计算, 系统通过相似度推荐产品[3]. 如何充分挖掘用户的兴趣并向用户有效推荐产品是当前研究的热点. 文献[4]研究如何利用基于内容与协作过滤模型建立信息推荐系统. 文献[5]利用基于内容过滤技术设计了一个针对航天领域文献的推荐系统.

结合中国现代化农资经营电子商务平台, 基于内容过滤的推荐系统建立了用户兴趣模型和产品模型, 设计了兴趣挖掘算法和基于内容过滤的推荐算法, 其中采用四元组模型建立用户兴趣模型, 将农产品地域性特征引入挖掘过程中, 引入遗忘因子来更新用户兴趣, 并解决了相似性计算问题, 实现产品推荐.

① 基金项目:十二五国家科技支撑计划(2012BAH20B00)
收稿时间:2013-08-13;收到修改稿时间:2013-09-12

2 用户兴趣挖掘

基于内容过滤的推荐技术主要问题包括用户兴趣模型的建立与更新以及相似性计算的问题，整个推荐过程如图 1 所示.

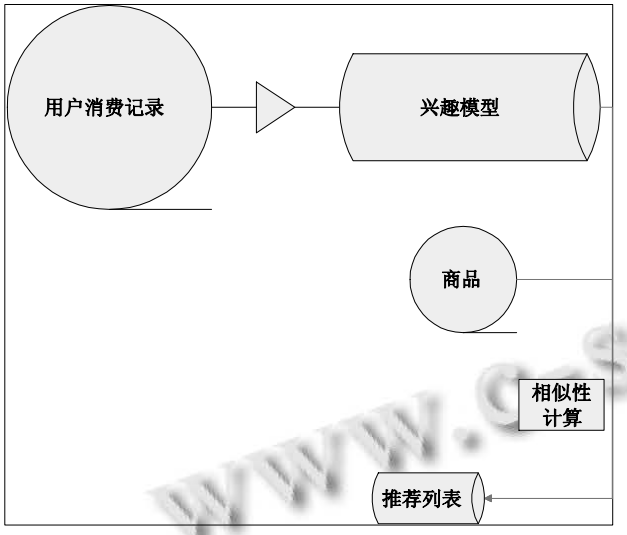

图 1　推荐过程总框图

2.1 用户兴趣的表示

用户建模是指从有关用户兴趣和行为中归纳出可计算的用户模型的过程. 在电子商务平台中，用户的消费记录是体现用户兴趣的重要信息，这里通过机器学习的方法挖掘用户的潜在兴趣，构造用户兴趣模型. 本文使用四元组建模，表示如下:

$$U_i=\left((c_1,w_1,t,V_1),(c_2,w_2,t,V_2),\ldots,(c_n,w_n,t,V_n)\right) \quad (1)$$

式中: U_i 为用户兴趣模型; c 为特征项; w 为特征项的权重,且满足约束条件 $\sum_{k=1}^{n} w_k=1$; t 为模型建立的时间; 其中第 i 个特征项的取值集合为 V_i[6].

2.2 产品模型的表示

与用户兴趣模型所不同的是，产品资源模型的属性没有权重. 这里将该类别中所有产品资源的属性取并集，若共有 n 个属性，则 V_i' 表示资源 R_i 第 i 个属性的属性值，若资源没有该属性则 V_i' 为 0, R_i 表示如下[6]:

$$R_i=\left((a_1,V_1'),(a_2,V_2'),\ldots,(a_n,V_n')\right) \quad (2)$$

2.3 用户兴趣模型的更新机制

2.3.1 遗忘函数

兴趣模型只有一直获取用户新的兴趣数据，才能保证系统对用户的适应性，因此用户的历史偏好应被遗忘，当前的偏好更应重视. 本文参考[7]中的指数型遗忘函数来衰减用户原有兴趣的权重，形式如下:

$$F(t)=e^{-\frac{\ln 2(t-est)}{hl}} \quad (3)$$

式中:

$F(t)$—遗忘因子，表示兴趣权重衰减到原来的比例;

t—用户消费记录建立的时间;

est—兴趣模型建立的时间;

hl—兴趣衰减的半衰期.

当 $t-est=hl$ 时兴趣恰好衰减到原来的 1/2. 兴趣权重衰减的速度是由参数 hl 决定和调节的. hl 越大，衰减速度越慢, hl 越小，衰减速度越快[7].

2.3.2 兴趣模型属性值的更新

对于评分、销售量等数值型属性值，属性值集合 $V_i\{\}$ 仅存储一个数值，通过遗忘因子动态更新用户偏好，其公式为:

$$V_{inew}=V_{iold}*F(t)+\frac{1}{m}V_i' \quad (4)$$

式中:

m—消费记录总数;

V_i'—第 i 条消费记录对应的属性值;

V_{inew}—新的属性值.

对于大部分非数值枚举型属性值，设定集合 $V_i\{\}$ 的大小为阈值 C, $V_i\{\}$ 中属性值的更新按照队列先进先出原理. 例如用户兴趣中某一属性值集合为(*a*,*b*,*c*,*d*,*e*)，现有新增记录 f，则属性值集合在 f 第一次出现时变为(*b*,*c*,*d*,*e*,*f*)，即当前兴趣值连续出现 C 次才可以完全替代历史兴趣值.

2.3.3 兴趣模型属性权重的更新

属性权重的更新一直以来是研究的难点. 依据某属性取值上变化越小则权重越大的原则，自适应调整属性权重. 设置属性值不为空的权重初始值为 w_a，空属性值的权重初始值为，权重之和满足约束条件[8]. 利用遗忘因子，用户偏好模型中若属性改变，则属性权重下降 $F(t)$; 若属性值不变，则在原有基础上调 $F(t)$，然后做归一化处理:

$$W_k=W_k'\Big/\sum_{k=1}^{n}W_k' \quad (5)$$

得到更新后的用户偏好模型 U_i'[6].

例如用户消费记录如图 2 所示，其中属性 ID 与用户偏好无关，则 U_i=((类型，化肥, w_a), (名称，复合肥二铵, w_a), (品牌，辉隆, w_a), (类别，国产化肥/国产复混肥/氮基复合肥, w_a), (发布时间，2012-11-10, w_a)),

w_a=1/5, 且满足约束条件.

该用户第二次消费记录如图 3 所示, 其中改变的属性包括名称、类别和时间, 调整权重为 w_a*(1+$F(t)$); 未改变的属性包括类型和品牌 w_a*(1-$F(t)$), 调整权重为然后做归一化处理得到偏好模型U_i'.

ID	类型	名称	品牌	类别	浏览次数	销售量	是否特价	是否热卖	是否新品	评分	发布时间
1	化肥	复合肥二铵	辉隆	国产化肥/国产复混肥/氮基复合肥	10	222	是	是	否	7	2012-11-10

图 2 农产品资源消费记录 1

ID	类型	名称	品牌	类别	浏览次数	销售量	是否特价	是否热卖	是否新品	评分	发布时间
2	化肥	多肽尿素 富尔康	辉隆	国产化肥/国产氮肥/尿素	1	5	是	是	否	5	2012-11-23

图 3 农产品资源消费记录 2

2.4 兴趣挖掘算法

根据[6]中所述, 用户兴趣挖掘算法如下:

输入: 当前类别 *ID*, 更新前的用户兴趣模型 U_i, 自最近一次更新后用户关于该类别所有消费记录 $R_j(j=1,2,\ldots,m)$, 其中按时间升序排列.

输出: 更新后的用户兴趣模型 U_i.

算法步骤:

(1) 判断是否已建立用户兴趣模型 U_i.

(2) 若 U_i 已建立, 首先更新属性值 *V*, 分数值型和非数值型两类.

(3) 其次更新属性权重 *W*.

(4) 重复(2)、(3), 直到最后一条消费记录.

(5) 若不存在历史兴趣模型, 则根据第一条消费记录来建立用户兴趣模型 U_i, 然后针对剩余的 *m*-1 消费记录依次执行步骤(2)~(4).

(6) 对于新注册用户, 不存在历史兴趣模型且消费、浏览记录为空, 则根据用户注册信息获得用户地理位置如: 省、市信息, 然后寻找同一省市的用户, 若存在, 则将其兴趣模型作为新注册用户的兴趣模型; 若不存在, 则该用户兴趣模型为空.

3 基于内容过滤的农资产品推荐算法

3.1 相似度计算

根据文献[6]的算法计算用户兴趣模型与产品资源模型之间的相似度, 计算用户偏好与产品资源对应属性值之间的差异值 D_k 并加权求和. 对于数值型属性, D_k 的计算式为:

$$D_k=\frac{|V_k-V_k'|}{\frac{1}{2}(V_k+V_k')} \tag{6}$$

对于非数值型属性, 根据用户偏好模型中非数值型属性值的更新方法, 若阈值 *C* 为 5, 则用户偏好模型中“品牌”属性取值为集合(辉隆, 江苏苏农, 上海农资, 浙江农资, 天津农资) 最多存储 5 个属性值, 若产品资源中品牌属性值为(辉隆), 则属性值的“差异值”为 1/5. 具体做法为: 将用户偏好模型中“品牌”属性编码(1,3,5,7,9), 则产品资源中“辉隆”编码为 1, 则依次判断 1 是否与用户偏好属性值集合中的值相等, 相等的次数为 $x(x\in[0,C])$, 则的计算式为:

$$D_k=\frac{V_k\cap V_k'}{V_k\cup V_k'}=\frac{x}{n} \tag{7}$$

式中分母 *n* 为用户偏好模型中属性值集合与产品资源属性值取并集后的元素个数.

则相似度计算公式为:

$$sim(U_i,Rc_j)=\sum_{k=1}^{n}(w_k\times D_k) \tag{8}$$

3.2 推荐算法

依据文献[6]中算法可得:

输入: 当前类别 *ID*, 用户兴趣模型 U_i, 相似度域值 *Z*, 产品资源模型 $Rc_j(j=1,2,\ldots,M)$.

输出: 与 U_i 最相似的产品集合 R_u.

推荐算法的步骤如下:

(1) 从兴趣模型 U_i 中获取它的属性权重向量 $W=(W_1,W_2,\ldots,W_n)$.

(2) 从兴趣模型 U_i 中获取它的属性值向量 $V=(V_1,V_2,\ldots,V_n)$.

(3) 从类别 *ID* 下某一产品资源模型中获取它对应属性值的向量$V'=(V_1',V_2',\ldots,V_n')$.

(4) 计算当前用户兴趣模型U_i与产品资源模型Rc_j相似度, 若相似度大于或等于 *Z*, 则 Rc_j 归入到集合 R_u; 否则, 舍去.

(5) 重复步骤(3)、(4), 计算该类别中每一个产品

资源与用户偏好模型的相似度.

算法说明: 相似度域值通过实验不断地修正, 最终达到一个合理的值, 算法流程图如图 4 所示.

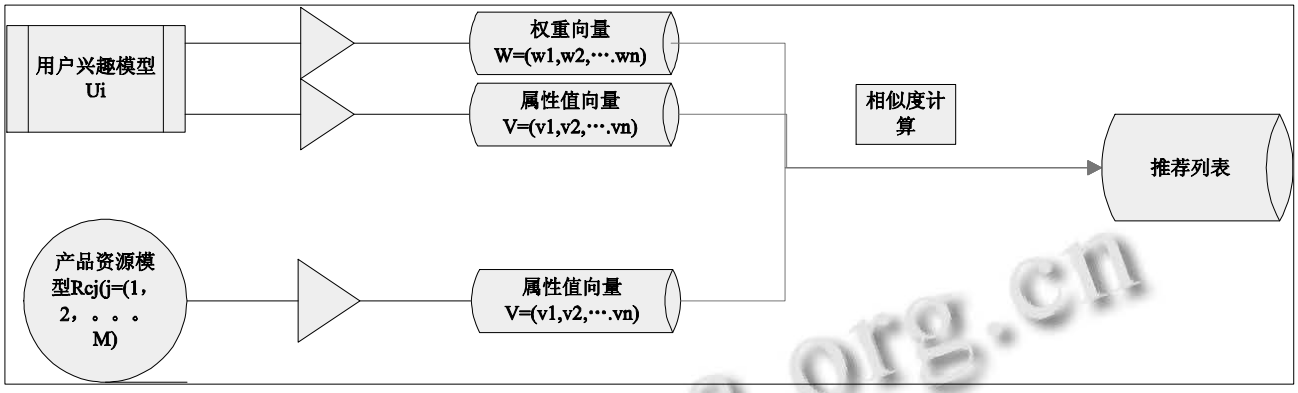

图 4　推荐算法流程图

4　实验系统及结果分析

农资电子商务推荐系统结合中国现代化农资经营电子商务平台, 系统以 MySQL 为数据库, 利用 J2EE 结构开发, 建立了用户兴趣模型、离线数据挖掘和在线推荐等模块, 其中推荐模块效果如图 5 所示.

序号	名称	品牌	类别
1	湛蓝-丰收牌硝硫基复合肥	辉隆	国产化肥/国产复混肥/硫基复合肥
2	中天-纯硫基复合肥1	辉隆	国产化肥/国产复混肥/硫基复合肥
3	衡水湖-硝硫基复合肥	江苏苏农	国产化肥/国产复混肥/硫基复合肥
4	美凯华-硝硫基复合肥	江苏苏农	国产化肥/国产复混肥/硫基复合肥
5	嘉施利高塔尿硫基20-5-20	无锡农资	国产化肥/国产复混肥/硫基复合肥
6	开磷高塔硝硫基	无锡农资	国产化肥/国产复混肥/硫基复合肥
7	丰田硫基复合肥	江苏苏农	国产化肥/国产复混肥/硫基复合肥
8	德沃尔高塔硝硫基	无锡农资	国产化肥/国产复混肥/硫基复合肥

图 5　农资推荐系统

系统采用中国现代化农资经营电子商务平台网站数据库中的客户消费记录评价校验本文的推荐算法. 为了说明算法的准确度, 引入了 ROC 曲线[9-10], ROC 曲线被称为受试者工作特征(ReceiverOperating Characteristic)[11]. 对于推荐系统, Herlocker J 等介绍了一种 ROC 曲线的画法[12]: 1)确定用户对每个产品感兴趣与否. 2)根据预测结果为用户提供一个推荐列表, 从图的原点开始, 如果预测的产品符合用户喜好,画一个竖线; 如果预测的产品不符合实际,画一个横线;如果预测产品还没有被打分,那么抛弃这个产品,并不影响曲线. 本文根据该方法计算推荐算法的准确度, 由于 ROC 曲线需要分类用户喜欢和不喜欢的产品, 本文假设用户浏览产品超过预设时间或购买的产品为喜欢的产品,打分为 1, 则其余的为不喜欢的产品, 打分为-1, 则曲线绘制过程中不存在未打分的产品. 共选取了 100 个用户, 每个用户有 100 条消费记录, 参数 C 和 Z 的取值分别取四组值(C=4,Z=0.5)、(C=4,Z=0.7)、(C=6,Z=0.5 和(C=6,Z=0.7). 实验得到: 1)参数 C 和 Z 取值为(C=6,Z=0.7)的情况下, 算法的准确度最高; 2)挖掘算法中引入地域性特征, 得到推荐列表中用户喜欢的产品占百分比近 55%; 不引入地域性特征, 百分比仅有 17%左右. 实验结果如图 6 所示, 为了避免点重合, 将引入地域性特征的点在纵轴分别上调和下降一个单位.

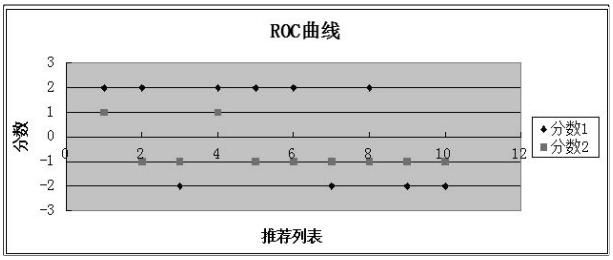

图 6 ROC 曲线

由实验结果可以看出: 用户兴趣挖掘算法中引入地域性特征的推荐算法比传统的推荐算法的推荐效果要好, 因为引入地域性特征, 使得挖掘用户兴趣模型的速度和准确率得到了提高.

5 结语

为了有效地提高个性化推荐服务的质量, 本文结合中国现代化农资经营电子商务平台, 综合用户消费记录, 引入农资产品的地域性特征, 挖掘用户兴趣模型, 并引入遗忘因子动态更新兴趣模型的属性及权重, 最后通过计算相似性产生推荐集合, 实现了一个农资电子商务推荐系统. 随着农业信息化和农资电子商务的不断发展和进步, 基于农资的个性化推荐系统将会越来越完善.

参考文献

1 曾春,邢春晓,周立柱.个性化服务技术综述.软件学报,2002, 13(10):1952−1960.

2 Amir A, Mohammad SB. A hybrid recommendation technique based on product category attributes. Expert Systems with Applications, 2009, (36): 11480−11488.

3 庄景明,王明文,叶茂盛.基于内容过滤的农业信息推荐系统.计算机工程,2012,38(11):38−40.

4 Kim BD, Kim SO. A new recommender system to combine content-based and collaborative filtering systems. Database Marketing, 2001, 6(3): 244−252.

5 Rao KN, Talwar VG. Content-based document recommender system for aerospace grey literature: System design. Journal of Library & Information Technology, 2011, 31(3): 189−201.

6 聂规划,孟洁,陈冬林.基于内容过滤的数字家庭服务资源推荐技术.武汉理工大学学报,2013,35(2):219−221.

7 布红艳,王国胤,董振兴.邮件系统中的兴趣偏移混合型.计算机工程与设计,2011,32(12):4026−4027.

8 陈基漓,牛秦洲.用户兴趣模型在图书馆个性化推荐服务中的应用.情报杂志,2009,28(1):190−193.

9 Swets JA. Information retrieval systems. Science, 1963, 141: 245−250.

10 Swets JA. Effectiveness of information retrievalmethods. Amer. Doc., 1969, 20: 72−89.

11 Hanley JA, Mcneil BJ. Themeaning and use of the area under a receiveroperating characteristic (ROC) curve. Radiology, 1982, 143: 29−36.

12 Herlocker J, Konstan JA, Terveen L, et al. Evaluating collaborative filtering recommender systems. ACM Trans. on Information Systems, 2004, 22(1): 5−53.

(上接第 185 页)

4 Vilnrotter VA, Hinedi S, Kumar R. Frequency estimation techniques for high dynamic trajectories. IEEE Trans. on Aerospace and Electronic Systems, 1989, (25): 559−577.

5 田园,吴长奇,牛晓斋.低信噪比高动态信号的载波同步研究.电子测量技术,2009,(5):40−43.

6 Kandeepan S, RJ Evans. Bias-free phase tracking with linear and nonlinear systems. Wireless Communication, 2010, (10): 3779−3789.

7 田甜,安建平,王爱华.高动态环境下无数据辅助的扩展 Kalman 滤波载波跟踪环.电子与信息学报,2013,(35):63−67.

8 Chen X, Wang WJ, Meng WX, Zhang ZZ. High dynamic GPS signal tracking based on UKF and carrier aiding technology. communications and mobile computing (CMC). 2010 International Conference. 2010. 476−480.

9 Vanneea DJR, Coenen JRM. New fast GPS code-acquisition technique using FFT. Electronics Letters, Jan. 1991, (27): 158−160.

10 徐佳鹤.基于 UKF 的滤波算法设计分析与应用[学位论文].沈阳:东北大学,2010.

11 Zhao HL, Li XL, Zhang J, et al. A novel frequency tracking algorithm in high dynamic environments. IEEE International Conference on WCNIS, 2010, 8: 31−34.

12 Julier S, Uhlmann JK. Unscented filtering and nonlinear estimation. IEEE, 2004, 92(3): 401−422.