文章编号:1000-2367(2017)01-0079-06 DOI:10.16366/j.cnki.1000-2367.2017.01.013

基于用户评论的协同过滤推荐算法

叶海智,刘骏飞

(河南师范大学 教育学院,河南 新乡 453007)

摘 要:提出融合用户评论的协同过滤推荐算法,通过挖掘电商网站的用户评论信息,获取用户评论中的产品特征和意见,通过计算每个特征意见对的极性,得到特征矩阵,结合用户意见质量形成的用户评分矩阵,求出用户评分的相似度.最后结合特征矩阵和用户评分相似度得出目标用户的综合相似度,并由预测评分得出产品推荐表,对用户进行产品推荐.实验结果表明,提出的算法与常用的推荐算法相比,改善了推荐的质量,同时推荐精度得到提高.

关键词:用户评论;推荐算法;相似度;协同过滤

中图分类号:TP301.6 **文献标志码**:A

随着Web 2.0技术的迅速发展,"信息过载"与用户个性化需求之间的矛盾日益突出,传统的基于搜索引擎的方式已经难以满足用户的需求.推荐系统成为解决用户个性化需求的有效途径,它可以通过系统提供的用户信息,为其推荐个性化内容.目前,推荐系统主要有基于内容的推荐系统、基于产品的推荐系统以及协同过滤的推荐系统,相较于前两类,协同过滤系统应用更为广泛.最早的协同过滤系统是由文献[1]提出的Tapestry系统.随着推荐技术的迅猛发展,文献[2-3]将用户环境信息和基于内容协同过滤算法相结合提出一种混合上下文推荐系统.文献[4]将上下文推荐系统同贝叶斯网络相结合,提出了一个基于贝叶斯网络上下文推荐算法,并设计了上下文资源推荐系统架构.文献[5]提出了一种将传统协同过滤方法和用户情绪特征、用户所在的上下文信息结合在一起的新的推荐系统.文献[6]将用户环境信息和用户服务结合,提出一种新的推荐系统架构.目前在结合评论挖掘的推荐方面,文献[7]通过深入对比研究各种推荐算法,指出用户评论挖掘技术对于推动推荐系统的应用发展具有深刻而广泛的意义,还有巨大的潜力进行研究.文献[3]把协同过滤方法同用户评分技术相结合,构建一个自动推荐系统为用户推荐电影或新闻.文献[8]将图论的有关方法应用到推荐系统中,通过用户节点构成的有向图来进行用户相似度的度量,最后利用深度搜索为用户提供相应的推荐内容.文献[9]将用户所处上下文信息和用户评论相结合,提出了新的混合推荐算法,通过该算法可以将用户的上下文信息以及邻居用户对产品的喜好程度相结合为用户提供推荐产品.

本文提出融合用户评论的协同过滤推荐算法,通过挖掘电商网站的用户评论信息,获取用户评论中关于产品特征以及相关特征的意见,最后结合特征矩阵和用户评分相似度得出目标用户的综合相似度,求出推荐结果.

1 相关知识

1.1 基于用户的协同过滤算法

基于用户的协同过滤主要是利用和推荐用户有一定联系的其他用户所做出的选择或购买行为之类的数据,给出推荐用户需要的结果,相似性用户需要由用户项目的评分数据集计算得出,通常和目标用户相似度

收稿日期:2016-09-22;**修回日期**:2016-12-05.

基金项目:河南省科技攻关重点项目(162102310442);河南省教育厅科学技术研究重点项目(14A880018).

作者简介(通信作者):叶海智(1963-),河南栾川人,河南师范大学教授,博士,主要研究方向为教育数据挖掘,E-mail:yhz87@163.com.

较高的用户对预测得分的贡献也就相应较大. 相似度计算方法主要有皮尔森相关相似度(Pearson Correlation coefficient)[10]、余弦相似度(Cosine Similarity)[11]等计算方法. 本算法采用的相似度

$$w_{uv} = \frac{\sum (i \in N(u) \cap N(v)) \frac{1}{\lg 1 + |N(i)|}}{\sqrt{|N(u)UN(v)|}}. \tag{1}$$

其中 w_{uv} 表示用户 u,v 之间的余弦相似度,$N(u)$ 是用户 u 感兴趣的物品集合,$I(v)$ 是用户 v 感兴趣的物品集合.

用户 u 对未评分物品 i 的预测得分为:

$$P(u,i) = \sum_{v \in S(u,G) \cap N(i)} w_{uv} r_{vi}, \tag{2}$$

其中 $S(u,G)$ 表示和用户 u 兴趣相似的 G 个物品集合,r_{vi} 表示用户 v 对物品 i 的兴趣得分.

1.2 基于物品的协同过滤

基于产品的协同过滤是利用用户曾经做出的物品评价或者购买物品等行为来计算物品之间的相似度,如(3)式所示,w_{ij} 是物品品 i,j 的相似度,$N(i)$ 是表示对物品 i 感兴趣的用户集合,而 $N(j)$ 是所有对物品 j 有兴趣的用户集合,

$$w_{ij} = \frac{|N(i) \cap N(j)|}{|N(i)|}. \tag{3}$$

通过计算物品之间的相似度,就有用户 u 对产品 j 的预测得分如(4)式所示,其中 $S(j,G)$ 表示和用户 j 兴趣相似的 G 个物品集合,r_{ui} 表示用户 u 对物品 i 的兴趣得分,

$$P_{uj} = \sum_{i \in N(u) \cap S(j,G)} w_{ji} r_{ui}. \tag{4}$$

2 产品特征意见抽取与量化

基于用户评论的协同过滤推荐算法的整体框架是:首先获取用户评价数据中产品质量评价,抽取产品的特征及相应的评价词,然后计算产品特征评价词的极性,从而形成相应的产品特征矩阵,结合由用户的意见质量生成的用户评分矩阵,联系用户技能和用户经验,从而得到用户评分的相似度,最后将产品特征矩阵和邻居用户评分相似度相结合,求出目标用户的综合相似度,进而预测评分,形成推荐列表,整体流程如图 1 所示.

2.1 用户评论的获取

这些评论为推荐系统提供了有价值的信息资源. 而自动获得此类信息需要创新的科技解决方案. 这些评论是文本的、松散的资源,以至于很难获取信息. 有效的用户意见选择、检索需要几个步骤,如:以常见的格式陈述信息(文本生成),然后根据意见计算产品等级,最后挑选最符合要求的意见,并将其作为推进回复给用户.

2.2 用户评论表述

评论首先映射到本体论上来计算排名. 在这个应用中,意见质量和产品质量分别在评论中总结了用户技术水平和产品使用经验. 其中意见质量包括几种变量,用来测量意见提供者的产品专业知识. 产品质量则表现了意见提供者对于产品特色的价值.

2.3 评定用户技术水平等级

评论意见由具有不同背景和专业水平的人给出. 一般来讲,长期使用产品的人更能提供专业的意见. 因此,经验丰富的人相较于对产品了解较少的人提供的意见水平更高. 意见质量(O)规定要根据意见提供者的专业水平评估其意见价值,如(5)式所示,

$$O_i = \frac{\sum_{j=i}^{n} w_j}{n}, \tag{5}$$

其中 w_j 表示权重,i 表示用户,j 表示技能和经验,n 表示用户技能和专业知识的信息的数量.

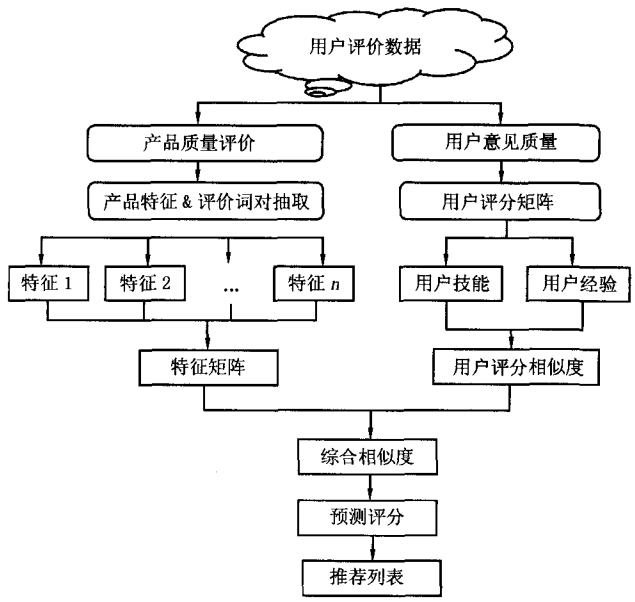

图 1　推荐算法整体框架

2.4　产品质量等级

根据每条用户评论的特色,产品被划分不同等级. 由于来自文本数据的用户评价量化难度较高,每条对产品某个特征评论均被赋值为"好"、"差"或"中性",分别记为"1","-1","0"这里用 r 表示,评分规则如下:

$$r(f) = \begin{cases} 1, \text{若评论为}''\text{好}'', \\ -1, \text{若评论为}''\text{差}'', \\ 0, \text{其他}. \end{cases} \tag{6}$$

通常每一个特性都是考虑了用户的价值和意见,这里以质量特性(F)表示,如(7)式所示

$$F_f = r(f) O_i. \tag{7}$$

3　融合用户评论的协同过滤推荐算法

3.1　评论整体的极性和强度预测

当一个用户要求在某些特性上评价某个产品时,需要从评论中获取产品的各种特性,要取得描述这些评价词就要从评论中提取产品在各个层面的特征,并获取用户评论中表示用户观点的评价性语句,这些就涉及语句的词性 POS(part of speech)层面,本文采用统计的方法对相关词性进行数据统计,这里使用总体特性质量来表示. 总体特性质量(Q)是来自所有评论特性的全局价值量,是通过特性的平均价值量计算的,如(8)式所示.

$$Q_f = \frac{1}{m} \sum S \cdot F_f, \tag{8}$$

这里 S 表示比例因子,m 表示特征个数,Q_f 值的大小则表示推荐或不推荐的强度.

3.2　用户评分相似度计算

协同过滤算法的相似度计算主要是兴趣接近的用户对同一个项目的评分数据,共同评分项目越多,精度就越高,但是计算复杂度会大量增加,反之相似度的结果可能就存在偶然性,可信度较差,为改进这种情况,利用欧氏距离公式计算用户 u 和 v 对物品评分的偏差:

$$d_r(R(u,i),R(v,i)) = O_i \sqrt{\sum_{i=0}^{k} | r_{ui} - r_{vi} |^2} + (1 + O_i) | Q_f(u,i) - Q_f(v,i) |, \tag{9}$$

这里，$R(u,i) = (r_{u1}, r_{u2}, \cdots, r_{uk}, Q_f(u,i))$，$R(v,i) = (r_{v1}, r_{v2}, \cdots, r_{vk}, Q_f(v,i))$，且 $r_{ui}, r_{vi}(i \in [1,k])$ 表示用户 u 和 v 对物品 i 的评分，$Q_f(u,i)$，$Q_f(v,i)$ 分别表示用户 u 和 v 对产品 i 的推荐情况. $d_r(R(u,i), R(v,i))$ 表示用户 u 和 v 对产品 i 的评分偏差.

用户 u 和 v 评分偏差由(9)式来计算：

$$d(u,v) = \frac{1}{|I(u,v)|} \sum_{i \in I} d_r(R(u,i), R(v,i)), \tag{10}$$

这里，$I(u,v)$ 表示用户 u 和 v 都进行评分的物品集合. 则用户 u,v 的评分相似度为：

$$s(u,v) = \frac{1}{1 + d(u,v)}. \tag{11}$$

3.3 评分预测与产品推荐

协同过滤算法中，基于用户的协同过滤通常设定目标用户也会接受和他喜好相似用户喜欢的物品，这样通过计算同用户爱好相似的邻居用户来预测用户对物品的评分. 通常与用户相似度较高的邻居用户对预测得分的贡献也就相应的较大. 选择与用户相似度最高的 N 个邻居用户集合. 可以得出预测产品评分公式.

$$P_{u,i} = \overline{R_N} + \frac{\sum_{n \in N} R_{n,i} - \overline{R_n} \cdot s(u,n)}{\sum_{n \in N} s(u,n)}, \tag{12}$$

这里，$\overline{R_N}$ 表示用户 U 对所有物品评分均值，N 表示与用户 U 兴趣相近的邻居用户结合，$R_{n,i}$ 表示邻居用户 n 对产品 i 的评分，R_n 表示邻居用户 n 对物品评分的均值，$s(u,n)$ 表示用户 U 和邻居用户 n 的评分相似度.

综上，利用(12)式求出目标用户对物品的预测评分，通过与用户实际评分进行比对来评测算法的精准度.

4 实验及结果分析

本文算法采用 JAVA 语言在 Eclipse 平台下进行数据测试，使用的数据集来自国内某知名购物网站上所有的用户评论和商品数据，此数据为 2008 - 2013 年间的商品评论数据，共包含用户评论数量 9249 万条，评论涉及的所有商品信息数量 48 万余件，参与评论的所有用户信息数量 56 万条. 经过对本数据进行预处理后，本文提取了从 2008 年 10 月至 2012 年 7 月之间，13 208 个用户对 531 种型号计算机的 236 450 条评论，其中每条评论均包含对产品的评论和评分. 数据稀疏度为 0.875 25，随机的将用户评分数据集进行划分，其中 80% 用于数据训练，20% 用于数据测试.

4.1 算法评价标准

对协同过滤推荐算法准确度进行评价的标准主要有平均绝对误差(MAE)、覆盖率(Cov)以及均方根误差(RMSE)等，本算法采用 MAE 值作为评价算法的标准. MAE 是通过计算训练集求出预测评分值和测试集上实际分数的平均绝对误差来对算法进行评测，推荐精度的高低同 MAE 的大小成反比. 本文实验结果采用平均绝对误差(MAE)来评测算法预测的准确度，MAE 的计算公式 C 为：

$$C = \frac{\sum_{u,i \in S} |\overline{r_{u,i}} - r_{u,i}|}{|I|}, \tag{13}$$

这里，I 表示测试集，$r_{u,i}$ 表示用户 U 对物品 i 的评分预测，$r_{u,i}$ 为测试集 I 中用户 U 对物品 i 的实际评分，C 的值越小，误差就越小，算法精度就越高.

4.2 实验结果及分析

为验证本文算法的推荐质量，选择传统的 UserCF 算法、ItemCF 算法、Slope One 算法、KNN 算法和 SVD 算法等推荐算法作为本文算法的比较对象.

在这里要计算意见质量的值，首先给出不同的权值 W_j，O 的取值范围在 0 到 1 之间，通过 O 的值的变化观察 MAE 值的变化，实验结果如图 2 所示. 从图 2 中可以看出，当 O 的取值不断增加的时候，C 的值不断减少，当 O 的值为 0.56 的时候，C 的值减到最少，此后 C 的值呈现出上升趋势，见图 2. 这是因为用户评论信息在相似度获取上权重降低，推荐的精度自然变差.

将最近邻居数的变化范围设置为 5 到 30,实验结果如图 3 所示. 从图 3 中可以看出,当邻居数增加的情况下,本文提出的算法 MAE 的值均低于其他的 4 种算法,这意味着本文算法的相似度计算结果更为准确,推荐的精度和可靠性都得到了提高.

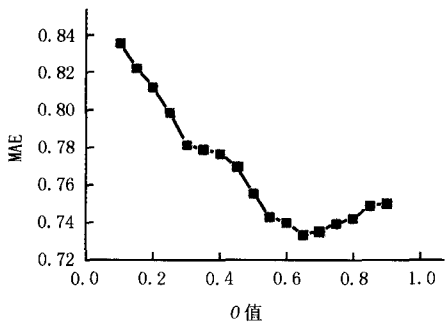

图 2　意见质量 θ 对 MAE 的影响

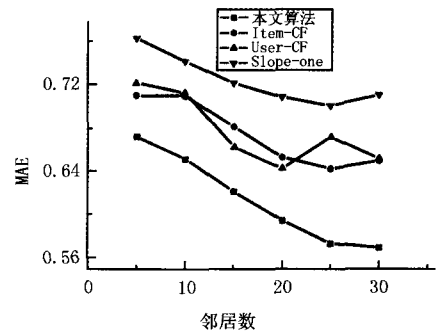

图 3　邻居数对 MAE 值的影响

在本次测试中,每次随机选取 10% 的用户进行预测,共进行 5 次测试,最后取 5 次 MAE 实验结果的均值. 另 KNN 算法选取的邻居用户数为 20,SVD 算法中正则化控制参数设置为 0.005,特征数取 25,迭代次数为 20. 结果如图 4 所示.

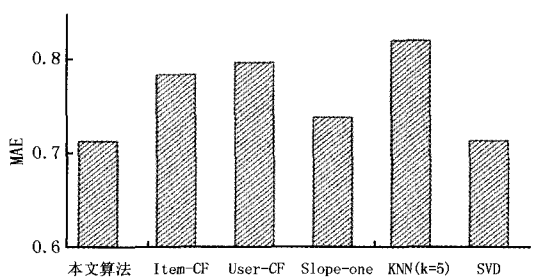

图 4　各算法在稀疏度为 89% 下的 MAE 值

MAE 值越小,预测精度越高,因此从图 4 中可以很直观地看出 KNN 算法的预测精度最差,本文算法在预测精度上与传统的 ItemCF 和 UserCF 算法相比优势也比较明显,与 Slope One 算法和 SVD 算法相比也具有一定的优势,取得了较好的预测精度.

5　小　结

本文提出了一个基于用户评论的协同过滤推荐算法,该算法融合了传统的协同过滤算法和用户评论信息,通过以国内某知名购物网站上的用户评论和商品数据的数据集进行验证对比发现,该算法与传统的推荐算法相比,预测精度得到了很大的提升. 以后的研究中将考虑能够在此算法的基础上,充分考虑纳入上下文信息,进一步提高预测精度.

参　考　文　献

[1] Goldberg D, Nichols d, Oki B M, et al. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, 1992, 35(12):61-70.

[2] Woerndl W, Schueller C, Wojtech R. A hybrid recommender system for context - aware recommendations of mobile applications[C]. Washington: In Proc. of the WPRSIUI IEEE Computer Society, 2007:871-878.

[3] Woerndl W, Brocco M, Eigner R. Context-Aware recommender systems in mobile scenarios[J]. Int'l Journal of Information Technology and Web Engineering, 2009, 4(1):67-85.

[4] 海本斋,解瑞云. 基于贝叶斯网络的上下文推荐算法[J]. 计算机科学,2014,41(7):275-278.

[5] Wang L C, Meng X W, Zhang Y J, et al. New approaches to mood-based hybrid collaborative filtering[C]. New York: ACM Press, 2010:28-33.

[6] Abbar S, Bouzeghoub M, Lopez S. Context-Aware recommender systems: A service-oriented approach[C]. Lyon: In Proc. of the VLDB Workshop on PersDB,2009:1-6.

[7] Adomavicius G,Tuzhilin A. Toward the next generation of recommender systems:A survey of the state-of-the-art and possible extensions[J]. Knowledge and Data Engineering,2005,17(6):734-749.

[8] Aggarwal C C,Wolf J L,WU K L,et al. Horting hatches an egg:A new graph-theoretic approach to collaborative filtering [C]//Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining. San Diego:ACM,1999:201-212.

[9] 张付志,刘赛,李忠华. 融合用户评论和环境信息的协同过滤推荐算法[J]. 小型微型计算机系统,2014,35(2):228-232.

[10] 秦佳,杨建峰,薛彬,等. 基于向量相似度匹配准则的图像配准与拼接[J]. 微电子学与计算机,2013,30(6):22-25.

[11] 陈大力,沈岩涛,谢槟竹,等. 基于余弦相似度模型的最佳教练遴选算法[J]. 东北大学学(自然科学版),2014,35(12):1697-1700.

[12] 张忠林,曹志宇,李元韬. 基于加权欧式距离的 k-means 算法研究[J]. 郑州大学学报(工学版),2010,31(1):89-92.

Collaborative Filtering Recommendation Algorithm Based on User Comments

Ye Haizhi, Liu Junfei

(College of Teachers and Educational Development, Henan Normal University, Xinxiang 453007, China)

Abstract: In this paper, a new collaborative filtering recommendation algorithm based on the fusion of user comments is proposed, which is based on the mining of users' comments on the electricity supplier website to obtain the product features and related features. By using the polarity of each feature opinion pairs, the characteristic matrix is formed, and then the user rating matrix is formed by the user's opinion quality, and the similarity of the user's score is obtained. Finally, according to the characteristic matrix and the user's score similarity, the comprehensive similarity of the target user is obtained, which can predict the score and form a recommendation list. Experimental results show that the proposed algorithm improves the recommendation accuracy compared with the traditional recommendation algorithm, and improves the quality of the recommendation.

Keywords: user comments; recommendation algorithm; similarity; collaborative filtering