基于用户评论评分与信任度的协同过滤算法*

王余斌，王成良，文俊浩
（重庆大学 软件学院，重庆 400044）

摘　要： 针对现有基于评论分析推荐算法中的评论真实度问题和传统协同过滤算法中的数据稀疏问题，通过分析用户评论所包含的主题分布和反馈信息，将改进的用户偏好和信任度引入传统协同过滤算法中，提出了基于用户评论评分与信任度的协同过滤算法。该算法以用户评论为基础，学习物品特征在不同主题上的分布及用户对物品不同特征的偏好程度，生成用户评论主题分布，根据用户评分计算评论差异度来放大主题分布中的突出特征，并利用评论反馈数据生成评论帮助度，进一步矫正用户偏好，以减少虚假评论的影响；引入信任度用于计算更精确的用户相似度，进而对用户进行评分预测和物品推荐。在真实数据集上进行了实验验证，结果表明该算法有效提高了系统的评分预测性能和推荐效果。

关键词： 协同过滤；信任度；主题模型；用户偏好；评论反馈
中图分类号： TP301.6　　**文献标志码：** A　　**文章编号：** 1001-3695(2018)05-1368-04
doi:10.3969/j.issn.1001-3695.2018.05.019

Research on collaborative filtering recommendation algorithm based on ratings, reviews and user trust

Wang Yubin, Wang Chengliang, Wen Junhao
(*School of Software Engineering*, *Chongqing University*, *Chongqing* 400044, *China*)

Abstract: In the interest of solving the review honesty problems in current recommendation algorithms based on review analysis, and the data sparsity problems in traditional collaborative filtering algorithm, this paper involved the modified user preferences and user trust into traditional collaborative filtering algorithm, and proposed a new user-based collaborative filtering algorithm based on user ratings, reviews and user trust by analyzing the topic distributions and feedback information included in reviews from users. In detail, with the consideration of users' reviews, it investigated the distribution of item features on different topics, and the level of preference of different users, to generate topic distribution of users' reviews. To continue, this paper calculated the value of review diversity to magnify distinctive features in topic distribution, and generated the value of review helpful-degree by utilizing review feedback data to rectify user preferences and reduce the influences from spam reviews. On this basis, it further introduced user trust to enhance the accuracy of user similarity. The results of experiment on real dataset show that the proposed algorithm can improve the quality of prediction performance.

Key words: collaborative filtering; user trust; topic model; user preference; review feedback

0　引言

信息技术的高速发展，在为人们提供巨大便利的同时，也带来了信息过载的问题。为解决这一问题，已有大量的研究人员提出了很多有效的方案。推荐系统作为其中极具代表性的一类解决方案，通过个性化的推荐算法，将不同的用户和不同的物品联系起来，既方便用户轻松发现自己感兴趣的物品，又能高效地将物品展示到感兴趣的用户面前。传统协同过滤算法包含了基于用户的协同过滤[1]和基于物品的协同过滤[2,3]两种方法。在众多的推荐算法中，协同过滤算法以其可利用群体智慧进行推荐的特点受到了广泛的关注。随着电子商务的蓬勃发展，用户逐渐成为互联网的主动参与者，开始越来越多地对网购物品、网络店铺、在线服务等发表个性化的评论，这些评论包含了用户的个人情绪、用户偏好、物品特征等丰富信息，不仅能指导其他用户进行网络消费活动，而且能帮助网络店铺搜集和获取反馈意见，提升自身的商品质量和服务体验。传统的推荐算法通常会忽略用户的评论信息，依靠用户评分数据进行评分预测和物品推荐，所以推荐系统数据稀疏性问题也就越来越受到研究人员的重视和关注，而一条真实可靠的评论文本中所包含的信息往往要比一条评分数据中包含的信息要多得多[4]，因此，基于用户评论信息的分析能够更加准确、具体地挖掘用户偏好，从而构建更为精确的用户偏好模型。

Goldberg 等人[5]提出一种新型的协同过滤算法 Eigentaste，将 PCA(principal component analysis)应用到评分矩阵的稠密子集上，对矩阵进行降维，从而在保证算法准确度的同时降低算法复杂度。但在从评论中获得特征信息时，采用人工的方式进行提取，这不仅需要丰富的专业知识，也需要耗费大量的人力和物力，难以取得很好的扩展性。Titov 等人[6]提出了一种基于多特征情感的 MAS(multi-aspect sentiment)模型。在 MAS 模型中，首先对用户评论中的一系列特征进行了预先设定；然后对主题建模，与这些预先设定的特征进行匹配。但该模型中同样需要对大量的数据进行人工训练，缺乏一定的灵活

收稿日期： 2017-01-05；**修回日期：** 2017-02-23　　**基金项目：** 国家自然科学基金资助项目(61379158)

作者简介： 王余斌(1992-)，男，江苏泰州人，硕士研究生，主要研究方向为 Web 数据挖掘、个性化推荐(wyb_alfred@163.com)；王成良(1964-)，男，教授，主要研究方向为 Web 技术研究与开发；文俊浩(1969-)，男，教授，主要研究方向为服务计算与面向对象的软件工程.

性,很难扩展到其他领域。McAuley 等人[7]通过 latent-factor 和 LDA(latent Dirichlet allocation)模型将潜在的评分维度与评论主题相结合,提出了 HFT(hidden factors as topics)模型。在 HFT 模型中,通过整合评分数据中的潜在因素和评论信息中的潜在主题,生成用户模型和物品画像,再进行评分预测。但在 HFT 模型中,每条评论只能属于用户或商品维度中的一个,另一维度则必须要与潜在因素空间匹配,存在一定的限制性。

区别于传统的协同过滤算法,本文利用 LDA 模型分析用户评论中包含的主题分布和反馈信息,将改进的用户偏好和用户信任度引入协同过滤算法中,提出了基于用户评论评分与信任度的协同过滤(user-based collaborative filtering with ratings, reviews and user trust, UserCF-RR)算法。UserCF-RR 算法通过将用户评论中的主题分布、用户的评论差异度与评论帮助度相结合,构建更具个性化的用户偏好模型,并在此基础上引入用户信任度,用于计算用户之间更精确的相似度,进一步提升系统的评分预测性能。

1　LDA 模型基本思想

LDA 模型是由 Blei 等人[8]在 2003 年提出的,目前已被成功应用到信息检索、评论分析、垃圾检测等多种与文本分析相关的领域[9~12]。LDA 模型可通过聚类文档中共现的词来识别大规模的文档集或语料库中潜藏的主题信息[13],其模型如图 1 所示。

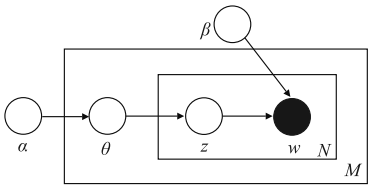

图1　LDA模型

LDA 模型包含文档、主题和词三个层次。其中一个文档集包含了 M 篇文档,一篇文档由 N 个词按照一定的顺序组成。LDA 采用词袋法(bag of words),将每一篇文档视为一个词频向量,与一个 K 维主题分布 θ 相对应。因此不同的文档生成不同主题 k 的概率也不同,在生成每篇文档时会对 θ 重采样一次,参数 α 和 β 针对每篇文档都是一样的,生成语料过程中只会被采样一次。α 用于生成一个主题 θ 向量,β 则代表着不同主题下各个单词对应的概率分布 $p(w|z)$,变量 z 代表主题,通过主题分布 θ 生成,而变量 w 代表单词,由 z 和 β 共同产生,每一个单词 w 都与一个主题 z 相对应。

在 LDA 模型中,生成一篇文档的主要步骤可归纳如下:

a)从狄利克雷分布(Dirichlet distribution)α 中采样,生成文档 d 的主题多项式分布 θ_d;

b)从主题分布 θ_d 中采样,生成文档 d 的第 i 个单词对应的主题 $z_{d,i}$;

c)从狄利克雷分布 β 中采样,生成每个主题对应的词语多项式分布 $\varphi_{z_{d,i}}$;

d)从词语分布 $\varphi_{z_{d,i}}$ 中采样,生成词语 $\omega_{d,i}$。

LDA 相比于其他文本分析模型,因其属于生成式概率模型,计算的过程更高效;同时 LDA 模型又是一种非监督学习技术,针对大规模的语料库或文档集有着更好的表现。因此,本文采用 LDA 模型对用户评论信息进行分析,可有效揭示评论文本中所讨论的主题特征及其分布情况。

2　基于用户评论评分与信任度的协同过滤算法

传统的协同过滤算法主要通过研究用户的历史行为来对用户进行兴趣建模,并作出相应的推荐[14]。在基于用户的协同过滤算法中,用户 u 对物品 i 的预测评分可通过式(1)进行计算。

$$\mathrm{pred}(u,i)=\overline{r_u}+\frac{\sum_{v\in N_u}\mathrm{sim}(u,v)\times(r_{v,i}-\overline{r_v})}{\sum_{v\in N_u}\mathrm{sim}(u,v)} \tag{1}$$

其中:N_u 表示用户 u 的最近邻居集合;$\mathrm{sim}(u,v)$ 表示用户 u 与 v 之间的相似度;$r_{v,i}$ 表示用户 v 对物品 i 的评分;$\overline{r_u}$ 和 $\overline{r_v}$ 分别表示用户 u 和 v 对所有物品的平均评分。

UserCF-RR 算法利用 LDA 主题模型对用户评论进行主题分析,生成用户偏好,并用评论差异度和评论帮助度对其进行优化;同时利用用户评分数据进行信任度计算,生成用户之间的信任关系;最后将用户偏好相似度与用户信任度相结合,对用户进行评分预测和物品推荐。其算法流程如图 2 所示。

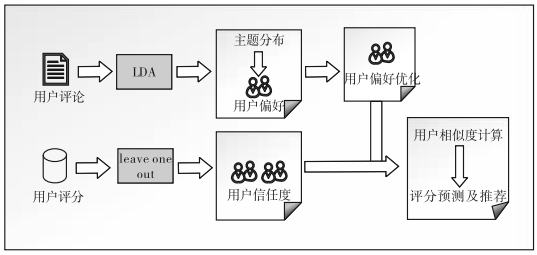

图2　UserCF-RR算法框架

2.1　用户偏好计算

在利用 LDA 模型生成用户偏好时,将用户 u 对物品 i 的每条评论视做一篇文档 $d_{u,i}$,利用 LDA 主题模型对评论内容进行处理分析,生成对应的 K 维主题分布 $\theta_{u,i}$。在主题分布 $\theta_{u,i}$ 中,包含了评论文档 $d_{u,i}$ 中所讨论内容的主题分布情况,可以用来指示用户偏好的特征。为进一步放大用户偏好特征,本文引入评论对应的评分值 $r_{u,i}$ 和评论获得的有帮助的反馈(helpful feedback)数量 $h_{u,i}$,作为用户偏好计算过程中的权重参数。$r_{u,i}$ 和 $h_{u,i}$ 的值越高,表明物品 i 的特征与用户 u 的偏好越吻合,评论的帮助度越高,对应的主题分布也就越符合用户偏好。

评论差异度 $D_{u,i}$ 定义如下:

$$D_{u,i}=\frac{2}{1+\mathrm{e}^{-(r_{u,i}-\overline{r_u})}}-1 \tag{2}$$

其中:$D_{u,i}$ 的取值为(-1,1)。当 $r_{u,i}>\overline{r_u}$ 时,用户评分较高,此时 $D_{u,i}>0$,表示用户 u 对物品 i 有喜好的偏向;反之,则表示用户 u 对物品 i 有讨厌的偏向。

评论帮助度 $H_{u,i}$ 定义如下:

$$H_{u,i}=\frac{1}{1+\mathrm{e}^{-(h_{u,i}-\overline{h_u})}} \tag{3}$$

其中:$\overline{h_u}$ 表示用户 u 所有评论获得有帮助的反馈的平均数量;$H_{u,i}$ 的取值为(0,1)。$h_{u,i}$ 值越大,评论帮助度越高,也就表示该条评论的参考价值越高。

结合上述两项权重值,本文将用户偏好 P_u 定义如下:

$$P_u=\frac{\sum_{i\in I_u}\theta_{u,i}D_{u,i}H_{u,i}}{|R_u|} \tag{4}$$

其中:$P_u=(P_{u1},P_{u2},\cdots,P_{uk})$;$I_u$ 表示用户 u 评论过的所有物品的集合;R_u 表示用户 u 的所有评论的集合。本文通过评论

差异度 $D_{u,i}$ 的值来放大用户偏好特征，并通过评论帮助度 $H_{u,i}$ 的值来矫正用户偏好。当 $D_{u,i}$ 越大，$H_{u,i}$ 越大，表明用户对物品的偏好程度越高，评论的帮助度也越高，也就表示评论的主题分布更有价值，更能反映用户的偏好特点。而当评论差异度 $D_{u,i}$ 很低，但评论帮助度 $H_{u,i}$ 很高时，表明用户对物品的评分很低，其他用户的认可度却很高，表明用户对物品的喜好程度较低，且符合物品的实际情况，即用户生成的是有效差评，这种情况下通过 $H_{u,i}$ 放大用户对物品的厌恶偏好，评论对应的主题分布也就更能体现用户实际偏好，进而更准确地进行物品推荐；反之，当评论差异度 $D_{u,i}$ 很高，但评论帮助度 $H_{u,i}$ 却很低时，表明用户对物品评分很高，其他用户的认可度却很低，即存在虚假评论的可能性，通过 $H_{u,i}$ 拉低最终的用户偏好 P_u，可以有效降低虚假评论对评分预测的影响。本文认为用户评论中的多用户合谋生成虚假评论的成本远高于单用户产生虚假评论的成本，用户合谋产生虚假评论概率较低，因此本文中只对单用户虚假评论内容进行过滤，不考虑虚假评论中的多用户合谋现象。

2.2 用户信任度计算

利用传统的协同过滤算法为用户推荐物品，往往会受到用户评分矩阵数据稀疏性的影响，无法计算用户相似度，利用用户之间的信任度来代替用户相似度的计算，可在一定程度上缓解数据稀疏的问题。本文中信任度将作为评判一个用户预测准确能力的度量指标，信任值的计算在每两个用户之间进行，且信任关系是不可逆的。即用户 A 高度信任用户 B 的情况下，用户 B 仍有极大可能对用户 A 是不信任的。针对某个目标用户，该用户可能会同时收到来自很多其他用户推荐的物品，但最终选择购买和评分物品的只会是其中的一部分，当目标用户对物品评分后，就可将目标用户的实际评分和预测评分进行比较，并获得两者之间的信任值。整个过程中，只关注目标用户 u 实际购买并评分的物品 i 对应的推荐用户 v 的评分预测精度。为了评估推荐用户 v 的评分预测精度，首先利用式(5)生成推荐用户 v 的评分预测 $\hat{r}_{u,i}$。

$$\hat{r}_{u,i} = r_{v,i} + (\overline{r_u} - \overline{r_v}) \tag{5}$$

其中：$r_{v,i}$ 表示推荐用户 v 对物品 i 的实际评分。得到推荐用户 v 的预测评分后，通过式(6)计算出推荐用户 v 相对于目标用户 u 在物品 i 上的评分预测精度 $p_v(u,i)$。

$$p_v(u,i) = 1 - \frac{|\hat{r}_{u,i} - r_{u,i}|}{4} \tag{6}$$

其中：$r_{u,i}$ 表示目标用户 u 对物品 i 的实际评分。由于推荐系统的评分采取 5 分制，5 即代表最高评分 5 分，1 即代表最低评分 1 分，这里使用最高分和最低分的差值 4 作为分母，以保证 $p_v(u,i)$ 的值保持在[0,1]内，并且 $p_v(u,i)$ 越大，说明评分预测精度越高。再将评分预测精度放大到全局，通过式(7)生成用户 u 对用户 v 的信任度 $\mathrm{trust}(u,v)$。

$$\mathrm{trust}(u,v) = \frac{\sum_{i \in I_{u,v}} p_v(u,i)}{|I_{u,v}|} \tag{7}$$

其中：$I_{u,v}$ 表示所有用户 v 推荐给用户 u 的物品集合；$|I_{u,v}|$ 表示集合 $I_{u,v}$ 中物品的总数。信任度 $\mathrm{trust}(u,v)$ 的取值在[0,1]，并且 $\mathrm{trust}(u,v)$ 的值越高，表示用户 v 的评分预测精度越高，用户 u 对用户 v 的信任度也就越高。随着目标用户实际评分数量的增加，相应信任度计算的准确度也会随之提高。

2.3 用户相似度计算

本文算法需对用户评论和评分双重信息进行计算，所以针对每个用户，相似度被分解为用户偏好相似度和用户信任度。

基于用户偏好的特征分布，使用余弦相似度[2]计算任意两个用户 u 与 v 之间的偏好相似度 $\mathrm{usim}(u,v)$，如式(8)所示。

$$\mathrm{usim}(u,v) = \frac{\sum_{j=1}^{k} P_{uj} \times P_{vj}}{\sqrt{\sum_{j=1}^{k} P_{uj}^2} \times \sqrt{\sum_{j=1}^{k} P_{vj}^2}} \tag{8}$$

其中：P_{uj} 和 P_{vj} 分别表示用户 u 和 v 在第 j 个主题上的用户偏好分布。本文认为两个用户同时对某个物品充满喜好或厌恶，都属于用户相似度的一种表现。

最后，利用参数 α 来平衡用户偏好相似度、信任度两者之间的重要性，计算出用户相似度 $\mathrm{sim}(u,v)$。对于任意两个用户 u 和 v，用户相似度 $\mathrm{sim}(u,v)$ 的计算如式(9)所示。

$$\mathrm{sim}(u,v) = \alpha \cdot \mathrm{usim}(u,v) + (1-\alpha) \cdot \mathrm{trust}(u,v) \tag{9}$$

当 $\alpha = 1$ 时，用户相似度退化为用户偏好相似度；当 $\alpha = 0$ 时，用户相似度退化为用户信任度；当 α 在(0,1)时，则依据用户偏好和用户信任度双重信息进行计算。

推荐系统的目标就是以用户的不同偏好为基础，为其推荐个性化的物品。本文通过评分差异度来放大用户偏好，并利用评分帮助度来矫正用户偏好，最终引入用户信任度，对用户之间的相似度进行多层面的综合评价，建立更加精准的用户偏好分布，进而实现高质量的个性化推荐。

3 实验

3.1 数据集

本文选用文献[15]中所使用的数据(源自 2011 年亚马逊电子设备(Amazon electronic device)评论数据集的数据)来进行实验。数据集中共包含 Coffee Machines、Canister Vacuums、Digital SLRs、Space Heaters、Laptops、MP3 Players 以及 Air-Conditioners 七类电子设备的评论信息。其中，每条评论都有对应的评分和反馈数据。为更好地反映数据集的稀疏程度，定义稀疏度 sparsity 的计算式如式(10)所示。

$$\mathrm{sparsity} = 1 - \frac{|\mathrm{reviews}|}{|\mathrm{users}| \times |\mathrm{items}|} \tag{10}$$

其中：$|\mathrm{reviews}|$ 表示数据集中的评论数量；$|\mathrm{users}|$ 表示数据集中的用户数量；$|\mathrm{items}|$ 表示数据集中的物品数量。数据集的稀疏度越高，表明其稀疏程度越高，数据也就越稀少。表 1 描述了该数据集的基本信息。

表 1 Amazon 电子设备数据集信息

数据集	用户数	物品数	评论数	稀疏度
Air-Conditioners	507	121	551	0.991
Space Heaters	3 721	366	3 822	0.993
Coffee Machines	4 016	284	4 054	0.993
Canister Vacuums	3 420	143	3 474	0.996
Digital SLRs	3 150	188	4 014	0.997
MP3 Players	3 239	375	3 504	0.997
Laptops	3 286	623	4 065	0.998

从表 1 可看出，每个数据集的数据都非常稀疏，稀疏度达到 99% 以上，若按照传统的协同过滤方法，仅依靠单一的用户评分数据进行个性化推荐，推荐系统的推荐质量将很难得以保证。

本文算法中引入了用户信任度来提高推荐质量。在使用 LDA 模型对评论信息进行主题分析的同时，需要对数据集中的评分数据进行预处理，建立用户的信任关系。本文使用差一

法(leave one out)进行计算。在数据集中,隐藏特定用户 v 的评分值,然后让其他每个用户对用户 v 进行评分预测,最后通过比较隐藏值(用户实际评分)和预测值的差异来计算两个用户之间的信任度。具体计算方法在 2.2 节中已经给出。

3.2　评价指标

本文采用常见的评价指标 MAE(mean absolute error)来评价系统评分预测的准确度。MAE 定义为系统预测评分与用户实际评分之间误差的平均值。

$$\mathrm{MAE}=\frac{\sum_{u,i\in N_t}|r_{u,i}-\mathrm{pre}(u,i)|}{|N_t|} \tag{11}$$

其中:N_t 表示数据集中所有评分对应的用户—物品的集合;$|N_t|$ 表示该集合的大小。MAE 的值越小,表示推荐系统的预测准确度越高;反之,则预测准确度越低。

3.3　实验及结果分析

本文通过两组不同的实验来验证算法的质量。实验 1 中,针对算法设置不同的参数来计算 UserCF-RR 在不同环境下的 MAE 值。实验 2 中,针对不同的算法,使用相同的数据集进行对比实验,分析算法的质量。

3.3.1　在不同参数下的实验结果比较

UserCF-RR 既是基于用户,也是基于主题分析的协同过滤算法,因此,本实验需针对算法中的主题数 K、最近邻居数(nearest neighbor,NN)两种参数进行对比实验。实验中用网格搜索确定式(9)中参数 α 的最优值,对主题数分别取 5、10、15、20,对用户最近邻居数分别取 5、10、15、20、25、30、35。表 2 显示了 UserCF-RR 在不同最近邻居数情况下的实验结果($K=10$)。

表 2　UserCF-RR 算法实验结果

数据集	$NN=5$	$NN=10$	$NN=15$	$NN=20$	$NN=25$	$NN=30$	$NN=35$	平均值
Air-Conditioners	1.386	1.231	1.106	0.951	0.963	1.107	1.122	1.124
Space Heater	1.291	1.039	0.956	1.023	0.897	0.845	0.902	0.993
Coffee Machine	1.279	1.138	1.061	1.087	1.046	1.073	1.102	1.112
Canister Vacuum	1.193	0.975	0.921	0.893	0.897	0.903	0.906	0.955
Digital SLR	0.816	0.672	0.615	0.528	0.498	0.535	0.552	0.602
MP3 Player	1.078	0.976	0.877	0.793	0.813	0.839	0.852	0.89
Laptops	0.877	0.834	0.816	0.703	0.71	0.699	0.686	0.761
平均值	1.131	0.981	0.907	0.854	0.832	0.857	0.875	

图 3 显示了 UserCF-RR 算法在不同最近邻居数情况下的实验结果。

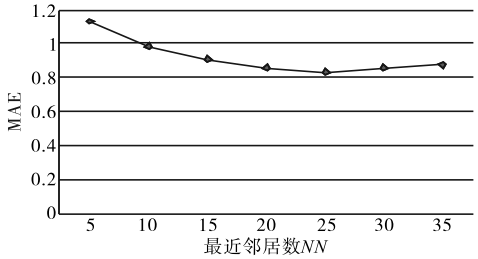

图3　UserCF-RR算法实验结果(不同最近邻居数)

图 4 显示了 UserCF-RR 算法在数据集中不同物品类别上的实验结果。

从表 2 实验结果可看出,UserCF-RR 算法在数据集中 Air-Conditioners 上的平均 MAE 值最大,推荐效果最差,而在数据集中 Digital SLR 上的效果最好。由于数据集中 Air-Conditioners 的用户数和评论数都很少,用户平均评论数也是七个数据集中唯一一个小于 1 的,实验结果与预期较为符合。从图 3 可看出,算法在最近邻居数为 25 时,取得最小 MAE 值 0.832,推荐效果最好。结合表 2 和图 4,可看出算法在数据集中 Air-Conditioners 上的最小 MAE 值为 0.951($NN=20$),在数据集中 Coffee Machines 上的最小 MAE 值为 1.046($NN=25$),在数据集中其余五个电子产品类别上的最小 MAE 值分别为 0.845、0.893、0.498、0.793 和 0.686。

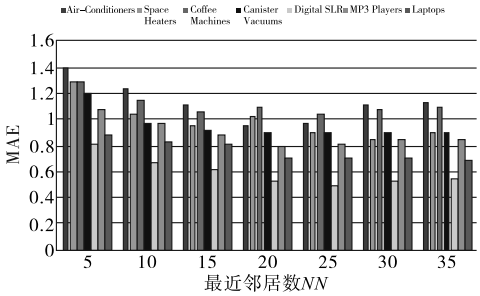

图4　UserCF-RR算法实验结果(不同物品类别)

3.3.2　对比不同算法的实验结果比较

本文使用传统的基于用户的协同过滤算法 UserCF(user-based collaborative filtering)、基于用户的主题分析过滤算法 ULCF(user-level LDA collaborative filtering),以及文献[16]中的 TMCF 模型进行对比实验。TMCF 模型同样是基于主题模型的协同过滤方法。实验 2 的结果如图 5 所示。

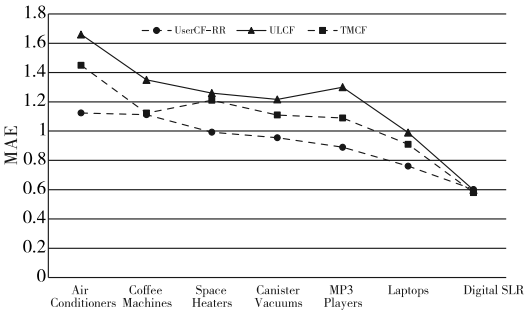

图5　对比算法实验结果

图 5 中 UserCF-RR 算法的整体效果较好,在实验数据集上的结果均优于 ULCF 算法和 TMCF 模型。相比于 ULCF 算法,本文在改进的用户偏好基础上,引入用户信任度,有效提升了算法的性能。而在针对 UCF 算法进行实验时,发现只有在 MP3 Players、Laptops 以及 Digital SLRs 数据集上才有 MAE 值,很明显这是由于数据的稀疏性所造成的数据缺失。实验表明,相比于传统的 UCF 算法,UserCF-RR 算法在稀疏数据集上拥有更好的推荐效果,能有效降低虚假评论对推荐系统的影响,评分预测能力更强。

4　结束语

本文通过分析用户评论中所包含的主题分布和用户反馈信息,并结合用户评分,引入用户信任度,提出了基于用户评论评分与信任度的协同过滤算法 UserCF-RR。该算法基于 LDA 模型生成评论主题分布,借助评论信息中的反馈数据计算评论帮助度,并利用用户评分数据计算评论差异度和用户信任度,结合评分相似度、偏好相似度以及信任度进行用户相似度计算,建立了更为精确的用户偏好模型。实验结果表明,UserCF-RR 算法在稀疏数据集上仍然拥有较好的推荐效果,基于评论帮助度和用户信任度,能有效降低虚假评论对推荐系统的影响。在未来的研究中,将增加对多用户合谋现象的分析,进一步提升系统的虚假评论过滤能力,并考虑用户之间的信任度传递等因素,以建立更精确的用户偏好模型,进一步提升推荐系统的质量。

(下转第 1402 页)

如果存在 x_3 使得 $\frac{m([x_3]\cap X)}{m[x_3]}=0.5$，那么对任何 $\beta>0.5$，$\overline{R}_\beta(X)\supseteq[x_3]$，而 $x_3\notin\underline{R}_\beta(X)$，故 $\mathrm{bnr}_\beta X\neq\varnothing$。

设不存在 x_3，并且 $m_1=\frac{m([x_1]\cap X)}{m[x_1]}>\frac{m([x]\cap X)}{m[x]}$，$x_1$，$x\in Y_1$；

$m_2=1-\frac{m([x_2]\cap X)}{m[x_2]}\geqslant 1-\frac{m([x]\cap X)}{m[x]}$，$x_2$，$x\in Y_2$，$m_1=m_2=\beta>0.5$，$x_2\notin\underline{R}_\beta(X)=\overline{R}_\beta(X)$。

当 $m_1<m_2$ 时，对任何 $\beta>m_1$，$[x_2]\not\subset\overline{R}_\beta(X)=\underline{R}_\beta(X)$；

当 $m_1>m_2$ 时，对任何 $\beta>m_2$，$[x_2]\subseteq\overline{R}_\beta(X)=\underline{R}_\beta(X)$。

所以 $\sup(\mathrm{Ndis}(R,X))=\min\{m_1,m_2\}$。

如果不存在这样的 x_1、x_2，类似可证。

3 结束语

本文提出了一个基于有界无限集合的变精度粗糙集模型——基于 Lebesgue 测度的变精度粗糙集模型，进而对相关性质进行了充分的研究。后续有待解决的问题有很多，主要归纳如下：a）对于无界集合的上下近似空间如何定义？如自然数集合及其子集偶数集合；b）拓展模型的应用研究，如快速分类、属性约简等；c）变精度粗糙集模型的公理化以及基于 Lebesgue 测度的变精度粗糙集模型的公理化是今后要研究的重点。

参考文献：

[1] 王丽娜. 基于粗糙集的数据挖掘改进的属性约简算法研究[D]. 成都：电子科技大学，2015.

[2] 张人上，曲开社. 一种基于新的特征选择的海量网络文本挖掘算法研究[J]. 计算机应用研究，2014，31(9)：2632-2634.

[3] Pawlak Z. Rough sets[J]. International Journal of Information and Compute Sciences，1982，11(5)：341-356.

[4] Ziarko W. Variable precision rough set model[J]. Journal of Computer and System Sciences，1993，46(1)：39-59.

[5] 刘妍琼，钟波. 变精度粗糙集模型中 β 参数范围的确定[J]. 湖南理工学院学报：自然科学版，2008，21(1)：11-13.

[6] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法[M]. 北京：科学出版社，2006：123-132.

[7] Liang Jiye，Wang Feng，Dang Chuangyin，*et al*. A group incremental approach to feature selection applying rough set technique[J]. IEEE Trans on Knowledge & Data Engineering，2014，26(2)：294-308.

[8] Liang Jiye，Mi Junrong，Wei Wei，*et al*. An accelerator for attribute reduction based on perspective of objects and attributes[J]. Knowledge-Based Systems，2013，44(1)：90-100.

[9] 苏永华，刘科伟，张进华. 基于粗糙集重心理论的公路隧道塌方风险分析[J]. 湖南大学学报：自然科学版，2013，40(1)：21-26.

[10] Liu Guilong. Axiomatic systems for rough sets and fuzzy rough sets[J]. International Journal of Approximate Reasoning，2008，48(3)：857-867.

[11] Liu Guilong. Using one axiom to characterize rough set and fuzzy rough set approximations[J]. Information Sciences，2013，223(2)：285-296.

[12] 刘耀峰. 基于 Sugeno 测度的粗糙集模型[D]. 保定：河北大学，2010.

[13] 薛占熬，刘杰，朱泰隆，等. 基于覆盖的 Sugeno 测度粗糙集模型及其三支决策[J]. 计算机科学，2016，43(3)：285-290.

[14] Yang Yanyan，Chen Degang，Dong Ze. Novel algorithms of attribute reduction with variable precision rough set model[J]. Neurocomputing，2014，139(9)：336-344.

[15] Chen Degang，Yang Yanyan，Dong Ze. An incremental algorithm for attribute reduction with variable precision rough sets[J]. Applied Soft Computing，2016，45(8)：129-149.

[16] Halmos P R. Measure theory[M]. [S. l.]：World Publishing Corporation，2007：100-152.

（上接第 1371 页）

参考文献：

[1] Resnick P，Iacovou N，Suchak M，*et al*. GroupLens：an open architecture for collaborative filtering of netnews[C]//Proc of ACM Conference on Computer Supported Cooperative Work. New York：ACM Press，1994：175-186.

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

[3] Linden G，Smith B，York J. Amazon. com recommendations：item-to-item collaborative filtering[J]. IEEE Internet Computing，2003，7(1)：76-80.

[4] Hu Guangneng，Dai Xinyu，Song Yunya，*et al*. A synthetic approach for recommendation：combining ratings，social relations，and reviews[C]//Proc of International Conference on Artificial Intelligence. Palo Alto，CA：AAAI Press，2015.

[5] Goldberg K，Roeder T，Gupta D，*et al*. Eigentaste：a constant time collaborative filtering algorithm[J]. Information Retrieval Journal，2001，4(2)：133-151.

[6] Titov I，McDonald R. A joint model of text and aspect ratings for sentiment summarization[C]//Proc of ACL-08 HLT. 2008：308-316.

[7] McAuley J，Leskovec J. Hidden factors and hidden topics：understanding rating dimensions with review text[C]//Proc of the 7th ACM Conference on Recommender Systems. New York：ACM Press，2013：165-172.

[8] Blei D M，Ng A Y，Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research，2003，3：993-1022.

[9] Wei Xing，Croft W B. LDA-based document models for Ad hoc retrieval[C]//Proc of the 29th Annual International SIGIR Conference on Research and Development in Information Retrieval. New York：ACM Press，2006：178-185.

[10] Alsumait L，Barbará D，Domeniconi C. On-line LDA：adaptive topic models for mining text streams with applications to topic detection and tracking[C]//Proc of the 8th IEEE International Conference on Data Mining. Washington DC：IEEE Computer Society，2008：3-12.

[11] Burns N，Bi Yaxin，Wang Hui，*et al*. A twofold-LDA model for customer review analysis[C]//Proc of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Piscataway：IEEE Press，2011：253-256.

[12] Lee K D，Han K，Myaeng S H. Capturing word choice patterns with LDA for fake review detection in sentiment analysis[C]//Proc of the 6th International Conference on Web Intelligence，Mining and Semantics. 2016：1-7.

[13] 张晓. 基于 LDA 主题模型的文本聚类研究[EB/OL]. [2012-02-28]. http://www. paper. edu. cn/releasepaper/content/201202-1066.

[14] 项亮. 推荐系统实践[M]. 北京：人民邮电出版社，2012.

[15] Jo Y，Oh A H. Aspect and sentiment unification model for online review analysis[C]//Proc of the 4th ACM International Conference on Web Search and Data Mining. New York：ACM Press，2011：815-824.

[16] Xu Jingnan，Zheng Xiaolin，Ding Weifeng. Personalized recommendation based on reviews and ratings alleviating the sparsity problem of collaborative filtering[C]//Proc of the 9th IEEE International Conference on e-Business Engineering. Washington DC：IEEE Computer Society，2012：9-16.