文章编号:0253-2778(2015)10-0804-09

一种解决稀疏数据和冷启动问题的组合推荐方法

郭晓波[1],赵书良[1],牛东攀[2], 王长宾[1],逄焕利[3]

(1.河北师范大学数学与信息科学学院,河北石家庄 050024; 2.长春工业大学人文信息学院,吉林长春 130000;
3.长春工业大学计算机科学与工程学院,吉林长春 130000)

摘要:针对传统推荐算法所面临的冷启动与稀疏数据问题以及现有ARM(association rule mining)算法大多用于购物篮顾客行为分析,并不适用于特定用户推荐业务且效率较低等现象,提出一种基于相似度的关联推荐模式,实现一种新的结合关联规则推荐与协同过滤推荐方法.采用基于指定后件项的关联规则推荐,直接对目标用户和目标项目进行关联规则挖掘,并利用兴趣因子对活跃用户(或项目)与非活跃用户(或项目)进行权值均衡,以加权方法推荐最优解(规则).同时,采用相似度测量方法,过滤低相似度的项目,为用户推荐既有高评分又具有较高相似度的项目集合.最后,结合规则推荐与CF(collaborative filter)推荐形成最终推荐结果,实现基于用户(或项目)的协同过滤推荐.在MovieLens数据集上的实验结果表明,同已有成果相比本文方法能够更好地处理稀疏数据和冷启动问题,推荐质量明显提高.

关键词:关联推荐;组合相似度;协同过滤;冷启动;稀疏数据

中图分类号:TP391.13 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2015.10.002

引用格式:Guo Xiaobo, Zhao Shuliang, Niu Dongpan et al. A novel combination recommendation method for solving sparse and cold start problems[J]. Journal of University of Science and Technology of China, 2015,45(10):804-812.

郭晓波,赵书良,牛东攀,等. 一种解决稀疏数据和冷启动问题的组合推荐方法[J]. 中国科学技术大学学报,2015,45(10):804-812.

A novel combination recommendation method for solving sparse and cold start problems

Guo Xiaobo[1], Zhao Shuliang[1], Niu Dongpan[2], Wang Changbin[1], Pang Huanli[3]

(1. *Mathematics & Information Science Colledge, Hebei Normal University, Shijiazhuang* 050024, *China*;
2. *College of Hmanities & Information, Changchun University Of Technology, Changchun* 130000, *China*;
3. *School of Computer Science and Engineering, Changchun University of Technology, Changchun* 130000, *China*)

Abstract: Considering the problems resulting from the traditional recommended approaches which are powerless to address the well-known cold-start and data sparseness, and the fact that most currently existing association rule mining(ARM) algorithms were designed with basket-oriented analysis in mind, which are inefficient for collaborative recommendation because they mine many rules that are not relevant to a given user, this paper introduces a novel association recommendation method based on combination similarity, and proposes a solution to the cold start problem by combining association rules and

收稿日期:2015-08-27;**修回日期**:2015-09-29

基金项目:国家自然科学基金(71271067), 国家社会科学基金(13BTY011).

作者简介:郭晓波,男,硕士/研究员,研究方向:智能信息处理. E-mail: xb_guo@163.com

通讯作者:赵书良,博士/教授. E-mail:zhaoshuliang@sina.com

collaborative filtering techniques. The proposed method focuses on mining rules for only one target user or target item at a time, while utilizing the interest factor to balance the weight between active users (or items) and non active users (or items), which in order to recommend an optimal solution (rules) via weighted method. To recommend both high ratings and collection of items with high similarity, the similarity measurement method was used to filter low similarity items, and to provide the final results by combining the association rules and CF recommendation, realizing user-based or item-based collaborative filtering recommendation. Experiments on the MovieLens data set reveals that the results obtained from employing this method has significantly better than the publishecl results and that it is better able to deal with sparse data and cold start problems.

Key words: association recommend; combination similarity; collaborative filtering; cold-start; data sparseness

0　引言

随着大数据应用的快速增长,在海量数据中如何快速而准确地为用户推荐可能感兴趣的商品或信息已经引起国内外研究人员的广泛关注. 作为一种为客户提供优质产品和服务的技术,推荐系统(recommendation system)已经成为人们日常生活中不可或缺的一部分,被应用到各行各业中. 推荐系统能够筛选可用的网页、图书、文章、电影、音乐、餐馆、杂货产品等,为我们找到最有趣和有价值的信息. 第一个推荐系统—Tapestry[1],给出一个新术语—协同过滤(collaborative filtering,CF),电子邮件分类过滤,解决 Xerox 公司在 Palo Alto 的研究中心资讯过载问题. 协同过滤基本假设是:若用户 U 和 V 对于 n 个项目的评分类似,或有类似行为(如阅读),可为用户 U 推荐用户 V 感兴趣的项目. 其中,最受欢迎的 CF 系统包括 Amazon[2]、eBay、Taobao 等网站.

从研究对象方法的角度来说,传统推荐系统主要分为[3-4]:基于内容的推荐(content-based recommendations),即基于用户上次喜欢的项目集合推荐相似项目;协同过滤推荐(collaborative recommendations),即基于寻找相同行为与偏好的用户对其物品进行推荐;混合推荐(hybrid approaches),即融合基于内容以及协同过滤推荐方法. 随着多种推荐技术的出现,是否具有提供个性化商品和服务的能力已经成为决定网上业务能否成功的关键. 面对网上的众多商品,客户一般会倾向于能够为自己提供个人喜好信息的网站. 基于信息的相似性和不同用户口味之间的异同,协同推荐系统为用户提供潜在感兴趣的商品. 一般 CF 方法采用某种相似度或权重测量方法 $w_{i,j}$,找出与该用户 U(或物品)最相似的 Top K 个用户(或物品). 由于该类方法过于依赖用户对物品的评分,当遇到稀疏数据[5]或新用户和物品[6,7]时,推荐效率有所下降,因此迫切需要研究新的推荐方法来解决稀疏数据和冷启动问题.

1　相关工作

目前,协同过滤是推荐系统中最流行的技术,但是传统 CF 方法仍然存在很多挑战,例如,如何有效处理海量数据中的稀疏数据与冷启动、提高算法效率及可扩展性等问题[8-10]. 针对上述问题,国内外研究人员已经研究出不同的方法来克服该类问题,并由此给出了高质量推荐结果. 在最近研究中,研究学者发现在 CF 中利用关联规则挖掘(association rule mining,ARM)预测 Top-K 推荐的任务,将其用于处理稀疏数据与冷启动问题,可提供更准确的推荐建议. García 等[11]介绍使用交互式迭代关联规则挖掘和推荐系统协同过滤,帮助教师改进的电子学习课程. Sarwar[12]等介绍了一种利用关联规则挖掘的推荐方法,该方法将满足最小支持度与置信度的规则进行排序,通过频繁项集或近邻规则取 Top-K 进行推荐. Leung 等[13]描述了一种基于模糊关联规则和多层次的相似性(FARAMS)CF 框架,该方法可用于解决稀疏数据问题. 同时 Leung 在[14-15]中给出了另一种跨层关联规则推荐方法(cross level association RulEs,CLARE),用于解决冷启动问题. Lin 等[16]给出一种高效的自适应支持 ASARM 的推荐系统,该方法通过挖掘指定规则后件的关联规则完成用户(或项目)推荐任务,并且不需要事先给出最小支持度. Shaw 等[17]采用 ARM 推荐方法处

理冷启动问题,该方法利用用户属性信息对物品进行评分. Sobhanam 等[18]介绍了一种基于关联规则的推荐算法,该方法结合了关联规则和聚类分析技术来解决冷启动问题. Khanzadeh 等[19]利用关联规则改进协同过滤的性能. Shweta 等[20]试图通过多目标粒子群(MOPSO)算法,在 CF 推荐框架运用 ARM 来改善推荐结果的质量,但该方法仅用支持度与置信度作为衡量参数. 同时,文献[21]给出了基于定量关联规则挖掘来提高新用户的推荐结果. Ye[22]结合关联规则挖掘和自我组织图提出了一种个性化协同过滤推荐方法,该方法可以缓解数据稀疏问题. Yang[23]给出了一种基于加权关联规则的改进协同过滤推荐算法,该方法无法解决活跃项目与非活跃项目的推荐不均衡问题. 上述研究内容主要集中在利用关联规则挖掘进行推荐,仅通过调整支持度与置信度对潜在推荐物品进行综合评分,由于置信度度量忽略了规则后件的项集支持度,高置信度的规则有时可能出现误导;另外,上述方法中并未结合相似度或权重测量方法来提高推荐结果的准确度.

传统 ARM 算法大多用于购物篮顾客行为分析,需要发掘多模式关联规则[24]服务大众用户,因此普通规则挖掘方法不适用于特定用户推荐业务. 相比之下,推荐系统旨在为特定用户提供个性化建议. 就关联规则而言,这意味着我们需要指定规则后件(antecedents →RecomObject)来完成关联规则挖掘. 为了提高挖掘过程的效率,许多现有规则挖掘算法将无法满足这种限制. 本文给出一种基于相似度的关联推荐方法. 首先,采用基于指定后件项的关联规则推荐,同时利用兴趣因子对活跃用户(或项目)与非活跃用户(或项目)进行权值均衡,直接对目标用户(或目标项目)进行关联规则挖掘,以加权方法推荐最优解(规则),该方法能够改善推荐结果的准确性,并且有利于处理稀疏数据和冷启动问题. 其次,由于仅需挖掘特定目标用户的规则信息,且不需要提前设置最小支持度和最小置信度,从而减少整个规则挖掘时间. 同时,采用相似度的 CF 测量方法,过滤低相似度的项目(只保留 $\mathrm{sim}(i,i_l)\geqslant\delta$ 项目,而不是对所有或 Top-K 相似项目进行加权求和),为用户推荐既有高评分又具有较高相似度的项目集合. 最后,结合规则推荐与 CF 推荐形成最终推荐结果,实现基于用户(或项目)协同过滤推荐.

2 基于指定后件项的规则推荐

2.1 规则度量方法

传统 AR 挖掘一般面向大众推荐而非个性化推荐,由于传统的关联规则挖掘算法局限于设置最小支持度和最小置信度两个参数进行规则挖掘,执行效率较差,并且这些参数阈值的选取将会显著影响关联规则的质量. 本文采用推荐项为后件的关联规则推荐方法,不需要提前设置最小支持度和最小置信度,而是直接对目标用户和目标项目进行关联规则挖掘. 其中,这两种类型的关联规则的目标用户(target user)或目标项目(target item)被称为目标对象(target object). 对于 CF 推荐方法而言,该目标对象需要被预先指定推荐用户(或物品),因此该规则是根据推荐系统的自身特点来挖掘目标用户或目标项目.

设 $U=\{u_1,u_2,\cdots u_j,\cdots,u_n\}(0<j<n)$ 表示一个用户集合,u_j 称为一个用户,代表一个具体用户;$I=\{i_1,i_2,\cdots i_j,\cdots,i_n\}$ 为 n 个不同项目的集合,i_j 称为一个项目,代表一个具体项目. 设定事务集 $T=\{t_1,t_2,\cdots t_j,\cdots,t_n\}$,其中 $t_j=(u_j,I_j)$ 代表一个用户 u_j 对所有项目的评分记录(事务),I_j 属于 I;t_j 表示对评分记录的标识(事务),(u_j,i_j) 代表一个用户 u_j 对 i_j 的评分,如表 1 所示.

表 1 用户/项目信息表

Tab. 1 Users/items information

user	item				
	i_1	i_2	i_3	i_4	
u_1	3	5	4	—	—
u_2	3	—	2	1	—
u_3	4	2	2	4	—

支持度度量值(support)用来评估规则的一般性(重要性). 给定一个关联规则 $x\Rightarrow y$ 并且满足 $x\cap y=\varnothing$,可知满足条件的事务集中同时包含 x 和 y,支持度计数 $\sigma(x)$ 表示为 $\sigma(x)=|\{t_i\mid x\subseteq t_i, t_i\in T\}|$.

$$\mathrm{support}(x\Rightarrow y)=\mathrm{support}(x\cup y) \tag{1}$$

置信度度量值(confidence)用来评估规则的有效性. 对于一个关联规则,它用来确定事务集 T 中包含后件 y 同时又包含前件 x 的频繁程度.

$$\mathrm{confidence}(x\Rightarrow y)=\frac{\mathrm{support}(x\cup y)}{\mathrm{support}(x)} \tag{2}$$

由于支持度-置信度框架的局限性，可能导致高置信度的规则有时出现误导的情况. 本文引入兴趣因子(interest factor)对活跃用户(或物品)与非活跃用户(或物品)进行权值均衡，避免出现置信度度量忽略规则后件项集的支持度的情况.

$$\text{interest}(x \Rightarrow y) = \frac{\text{support}(x \Rightarrow y)}{\text{support}(y)} \tag{3}$$

2.2　规则评分方法

基于推荐项为后件的关联规则推荐可描述为：k-item - RecomObject ⇒ RecomObject，其中 k-item=(k−1)item + RecomObject，即(k−1) item ⇒ RecomObject，RecomObject 表示目标对象集合，UserObject 和 ItemObject 属于 RecomObject，(k−1)item 表示目标对象 o_j 与目标对象 o_i 共同拥有项的集合. 推荐项为后件的关联规则表示为图 1 和图 2. 其中，u_r 和 i_r 表示推荐目标. 在本文中，支持度、置信度以及兴趣因子是 ARM 的三个重要参数，最终以加权方法推荐最优解(规则). 与普通 ARM 的任意形式规则相比，由于每次仅需要推荐相关的目标对象 k-item 的规则，仅需分析子集数据而非全集数据；同时推荐项为后件的关联规则，使得整个挖掘过程节省了大量执行时间，有利于提高推荐效率.

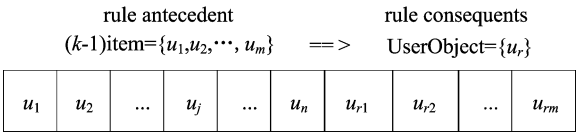

图 1　基于用户推荐

Fig. 1　User-based recommendation

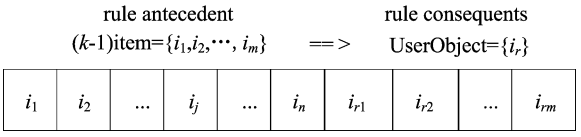

图 2　基于项目推荐

Fig. 2　Item-based recommendation

对于某个用户 u 的推荐目标项目 i 的预测评分，具体计算公式为：

$$\text{score}(\text{rule}) = \text{support} * \text{confidence} * \text{interest} \tag{4}$$

$$\text{pred}(\text{rule}) = \bar{r}_u + \frac{\sum_{\text{rule} \in R_u} \text{score}(\text{rule})}{|R_u|} \tag{5}$$

式中，R_u 表示给用户 u 关联规则集合，$\bar{r}_u$ 表示用户 u 对所有评分项目的平均评分. 当 $\bar{r}_u = 0$ 时，式(5)也能对一个历史信息较少的新用户提供预测分，有效地处理稀疏数据和冷启动问题.

3　基于 AR 和 CF 的推荐方法

3.1　相似性度量方法

传统 CF 推荐方法一般基于对相似用户或项目记录的喜好项来预测目标用户的喜好. 如表 1 所示，我们将其看作是由用户集合 U 和项目集合 I 构成的 $m \times n$ 评分矩阵 $\boldsymbol{R}$，其中 $r_{u,i}$ 表示用户 u 对项目 i 的评分值. CF 推荐算法首先找出用户中每个用户的评分集合，计算得到项目集合 I 中项目 i 的最近邻项目作为候选项目集合 C，并且从中除去目标用户已经访问过的项目. 对于集合 C 中的每个备选项目 c，预测目标用户对其的评分. 人们通常情况采用以下两种相似性度量方法.

(Ⅰ)Pearson 相关系数：给定评分矩阵 $\boldsymbol{R}$，用户 u_a 和用户 u_b 之间的相似度 $\text{sim}(u_a, u_b)$ 可表示为[25]：

$$\text{sim}(u_a, u_b) = \frac{\sum_{i \in (I_a \cap I_b)} (r_{u_a, i_a} - \bar{r}_{ua})(r_{u_g, i_g} - \bar{r}_{ub})}{\sqrt{\sum_{i \in (I_a \cap I_b)} (r_{u_a, i_a} - \bar{r}_{ua})^2} \sqrt{\sum_{i \in (I_a \cap I_b)} (r_{u_g, i_b} - \bar{r}_{ub})^2}} \tag{6}$$

式中，$\bar{r}_{ua}$ 和 $\bar{r}_{ub}$ 表示用户 u_a 和用户 u_b 的平均评分.

(Ⅱ) 为了找到相似项目，余弦相似度通常用于计算两个 n 维向量之间的相似度. 设 U 表示对项目 i, j 均有评分的用户集合，向量 $\vec{i}$ 和 $\vec{j}$ 分别表示用户 u 在 i, j 上的评分，则两个项目 i, j 之间的改进余弦相似性 $\text{sim}(i, j)$ 如下[26]：

$$\text{sim}(i, j) = \frac{\sum_{u \in U} (R_{u,i} - \bar{r}_u)(R_{u,j} - \bar{r}_u)}{\sqrt{\sum_{u \in U} (R_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{u \in U} (R_{u,j} - \bar{r}_u)^2}} \tag{7}$$

式中，$R_{u,i}$ 和 $R_{u,j}$ 分别表示用户 u 对项目 i 和 j 的评分.

3.2　预测评分方法

若对用户 u 进行项目推荐，其评分向量如表 2 所示.

表 2　用户 u_a 评分向量

Tab. 2　Score vector of user ua

user	item					
	i_1	i_2	i_3	i_4	…	i_n
u_a	—	4	1	3	—	—

首先删除目标用户 u 访问过的项目 $I_u=\{i_2,i_3,i_4\}$ 后，剩余的所有为备选项目集合 $C-I_u=\{i_1,i_5,i_6,\cdots,i_n\}$. 若我们预测 u 对于 i_1 的评分，从而决定是否对 u 推荐 i_1，由于 u 只对 $I_u=\{i_2,i_3,i_4\}$ 进行了评分，所以只需要对照 R 中 i_1 与 i_2、i_3、i_4 所在行的数据，这里我们采用公式(7)计算相似度，如表3所示.

表3 i_1 与 i_2、i_3、i_4 相似性矩阵

Tab. 3 Similarity matrix i_1 and i_2、i_3、i_4

item	item		
	i_2	i_3	i_4
i_1	0.854 198	0.910 182	0.921 635

根据式(8)预测目标用户 u 对备选项目集合 $C-I_u$ 中每个项目 i 的评分.

$$\mathrm{pred}_{\text{user-based}}(u,i)=\bar{r}_u+\frac{\sum\limits_{u_l\in S(u)}\mathrm{sim}(u,u_l)(r_{u,i_l}-\bar{r}_{ul})}{\sum\limits_{u_l\in S(u)}\mathrm{sim}(u,u_l)} \tag{8}$$

$$\mathrm{pred}_{\text{item-based}}(u,i)=\bar{r}_i+\frac{\sum\limits_{i_l\in S(i)}\mathrm{sim}(i,i_l)(r_{u,i_l}-\bar{r}_{il})}{\sum\limits_{i_l\in S(i)}\mathrm{sim}(i,i_l)} \tag{9}$$

式中，S 表示推荐列表，r_{u,i_l} 为用户 u 对于项目 i_l 的评分，$\bar{r}_u$ 表示用户 u 对所有评分项目的平均评分，$\bar{r}_i$ 表示所有用户对于项目 i 的平均评分，$\bar{r}_{i_l}$ 表示所有访问过项目 i_l 的用户对于项目 i_l 的平均评分. 其中 $S(i)$ 为用户 u 评分的项目集合 $I_u=\{i,j,\cdots,k\}$ 中 i 的最邻近项目集合，其标准是 i 与 i_l 之间的相似度 $\mathrm{sim}(i,i_l)$ 大于设定的某个阈值 $\delta\in[0,1]$. 为了提高推荐质量，我们只保留 $\mathrm{sim}(i,i_l)\geqslant\delta$ 的项目，而不是对所有(或 Top-K)相似项目进行加权求和，从而形成一个既有高评分又具有较高相似度的项目集合 I_S.

3.3 相似度线性组合推荐策略

本文采用结合关联规则推荐和 CF 推荐方法，将 $\mathrm{pred}(\mathrm{rule}_{u,i})$ 与 $\mathrm{pred}(u,i)$ 加权总和对用户 u 的项目 i 进行评分预测，即

$$\mathrm{pr}(u,i)=\alpha*\mathrm{pred}(u,i)+(1-\alpha)*\mathrm{pred}(\mathrm{rule}_{u,i}) \tag{10}$$

式中，$\alpha\in[0,1]$ 是一个组合参数，用于指定组合相似性度量的权重. 当 $\alpha=0$ 时，$\mathrm{pr}(u,i)=\mathrm{pred}(\mathrm{rule}_{u,i})$，即采用关联规则推荐 $\mathrm{pred}(\mathrm{rule}_{u,i})$ 作为用户 u 项目 i 的评分预测，这样能够有效地解决新用户问题；相反，当 $\alpha=1$，$\mathrm{pr}(u,i)=\mathrm{pred}(u,i)$，即采用 CF 推荐 $\mathrm{pred}(u,i)$ 作为用户 u 项目 i 的评分预测；当 $\alpha\in[0.1,0.9]$ 时，$\mathrm{pr}(u,i)=\alpha*\mathrm{pred}(u,i)+(1-\alpha)*\mathrm{pred}(\mathrm{rule}_{u,i})$，结合关联规则推荐和 CF 推荐方法. 如何选取 α 值属于一个非平凡过程，它通常高度依赖于数据的特点，因此用户 u 项目 i 的评分预测会随着组合参数 α 变化有所不同，可根据经验设置组合参数的值.

4 实验结果及对比分析

4.1 数据集

本实验数据采用基于 Web 搜索推荐数据集 MovieLens[27]，完整数据集包含 943 个用户在 1 682 部电影上的 100 000 条评分记录，评分值采用 5 分制. 其中，每个用户至少对 20 部电影进行了评分. 实验采用 5 折交叉验证，训练集和测试集占比分别为 80% 和 20%. 该数据集的稀疏等级为 0.936 91[12].

4.2 评价标准

本文采用均方根误差(RMSE)、准确率(precision)、召回率(recall)、覆盖率(coverage)、F 值(F-measure) 作为评价推荐算法预测用户行为能力的度量标准，评估推荐列表是否匹配用户的喜好. 在评分预测问题中，我们用均方根误差评测评分预测的准确度，用准确率、召回率和与 F 值来度量推荐的精度，以覆盖率表示系统能够为用户推荐的商品占所有商品的比例. 实验中，我们对 5 折交叉验证中实验结果取均值. 设 $R(u)$ 是根据用户 u 在训练集上的行为给用户提供的推荐列表，$T(u)$ 是用户 u 在测试集上的行为列表，I 是测试集上所有物品集合.

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{u,i\in T}(r_{u,i}-\hat{r}_{u,i})^2}{|T|}} \tag{11}$$

$$\mathrm{precision}=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|R(u)|} \tag{12}$$

$$\mathrm{recall}=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|T(u)|} \tag{13}$$

$$\mathrm{coverage}=\frac{|\bigcup_{u,i\in R(u)}R(u_i)|}{|I|} \tag{14}$$

$$F_1\text{-measure}=\frac{2*\mathrm{precision}*\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}} \tag{15}$$

式中，$T(u)$ 对于测试集中的一个用户 u 的项目 i，

$r_{u,i}$ 表示用户 u 对项目 i 的实际评分，$\hat{r}_{u,i}$ 表示预测评分.

4.3　实验对比结果

为了验证本文所提推荐算法的有效性，我们将其与传统推荐方法进行分析. 具体控制参数含义及其缺省值详见表 4.

表 4　测试数据的相关参数

Tab. 4　Related parameters of test data

参数符号	具体含义	设置大小
MaxRuleLeng	最大规则项目数量	[3,8]
RangesRuleNum	推荐规则数量	[10,100]
interest	推荐规则兴趣因子	{yes,no}
δ	相似度阈值	[0,1]
α	组合参数	[0,1]
Top-K	推荐列表数量	[5,200]

4.3.1　指标参数分析

(Ⅰ)最大规则项目数量

本文从最大规则项目数量对规则推荐方法进行分析，推荐规则数量为[10,100]，最大规则项目数量为[3,8]，实验结果如表 5 所示. 当 MaxRuleLeng＝3 时，推荐结果最优. 为避免过拟合问题，后续实验中，我们采用最大规则长度为 3 的推荐方式.

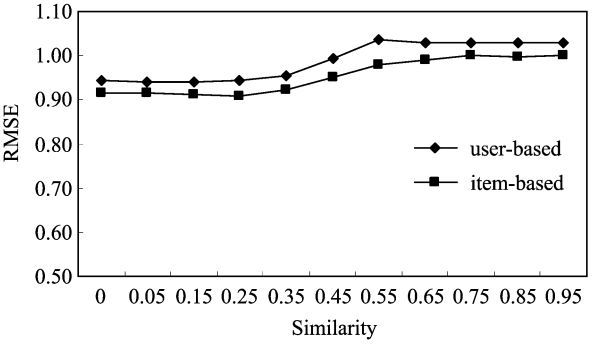

图 3　user-based 与 item-based 推荐方法 RMSE 对比结果

Fig. 3　RMSE comparison results between user-based and item-based

表 5　最大规则项目数量的性能分析

Tab. 5　Performance analysis of maximum rule items

accuracy	MaxRuleLeng					
	3	4	5	6	7	8
RMSE	0.916 515	0.922 497	0.927 146	0.931 052	0.931 051	0.931 051

表 6　不同相似度阈值 δ 的准确度

Tab. 6　Accuracy in different similarity thresholds δ

CF algorithm	δ									
	0	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
user-based	0.942 492	0.939 638	0.941 531	0.942 703	0.955 814	0.992 293	1.034 441	1.030 258	1.030 275	1.030 633	1.030 759
item-based	0.914 127	0.914 023	0.910 469	0.909 306	0.922 920	0.952 074	0.977 827	0.990 245	0.999 144	0.998 750	0.999 997

表 7　不同组合参数 α 下的准确度

Tab. 7　Accuracy in different combined parameters α

accuracy	α									
	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
RMSE	0.962 463	0.945 730	0.930 882	0.918 012	0.907 203	0.898 530	0.892 055	0.887 827	0.885 877	0.886 220

(Ⅱ)相似度阈值 δ 分析

在评分预测问题中，我们用均方根误差评测评分预测的准确度. 实验分析了相似度阈值 δ 选取 0～1 之间的分布情况，user-based 与 item-based 推荐方法在不同相似度阈值 δ 的准确度，如表 6 所示.

由图 3 可知，与 user-based 推荐方法相比，item-based 推荐方法具有更好的推荐性能. 特别是在 $\delta = 0.25$ 时，item-based 推荐方法的 RMSE ＝0.909 306.

(Ⅲ)组合参数 α 分析

实验分析结合关联规则推荐和 CF 推荐方法，由表 7 可知，当 $\alpha = 0.85$ 时推荐质量最优.

(Ⅳ)兴趣因子 interest 分析

本文引入兴趣因子(interest factor)对活跃用户(或物品)与非活跃用户(或物品)进行权值均衡,避免出现置信度度量忽略规则后件项集的支持度.实验采用 item-based CF 算法(user-base CF 亦同),其中,$\delta=0.25$,$\alpha=0.85$.由表 8 可知,采用支持度、置信度以及兴趣因子三个参数的加权值来推荐最优解(规则)比传统支持度-置信度框架的推荐准确度有明显提高.

表 8 兴趣因子 interest 对比分析

Tab. 8 Comparison of interest factors

accuracy	interest	
	yes	no
RMSE	0.885 877	0.921 909

(Ⅴ)Top-K 分析

在 Top-K 推荐中,相似度 δ 是一个重要的参数,即为每个用户选出满足 $\mathrm{sim}(i_a,i_l)\geq\delta$ 的用户,然后推荐前 K 个用户感兴趣的物品.实验采用 item-based CF 算法(user-base CF 亦同).其中,$\delta=0.25$,$\alpha=0.85$,$K=\{5,10,20,40,80,160,200\}$,推荐列表大小为 20.在实验对比分析中,$K$ 取最优准确率和召回率的值为 5,具体如表 9 所示.实验结果表明,与其他推荐方式[2,26]相比,该方法的准确率和召回率取得了明显的改进,且提高了推荐结果的覆盖率.

表 9 Top-K 分析对比分析结果

Tab. 9 Comparative analysis results of Top-k

type	index			
	precision	recall	coverage	F-measure
item CF	0.443 223	0.104 49	0.635 259	0.169 111
item CF-IUF	0.445 005	0.104 91	0.631 748	0.169 791
ourapproach	0.649 045	0.405 86	0.846 152	0.499 422

4.3.2 冷启动问题比较

如何为新用户推荐合适的物品是每个推荐系统关注的重点问题.为了验证文中方法的科学性,我们随机选择 $n(n=\{5,10,15,20\})$ 项冷启动项目,对于每种类别我们选取 100 个随机样本,采用 5 折交叉验证求均值来测试算法处理稀疏数据和冷启动的性能.图 4 给出的实验结果表明,当 $n=\{5,10,15,20\}$ 时,采用 $\delta=0.25$,$\alpha=0.85$ 组合推荐方法 $\mathrm{pr}(u,i)$ 作为用户 u 项目 i 的评分预测.与其他推荐方法所得到的 RMSE 结果相比,该方法能够很好地解决新用户和稀疏数据问题,即用户仅对很少项目评分.从实验结果可以看出,本文所提出方法取得了明显的改进.

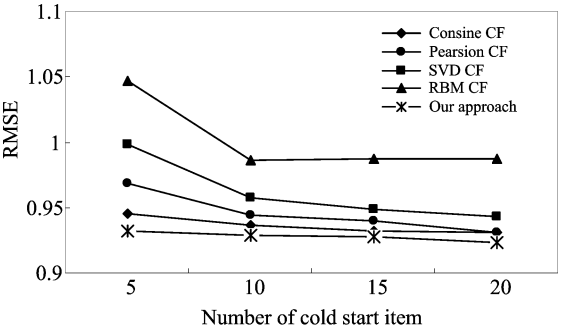

图 4 处理冷启动问题的 *RMSE* 对比结果

Fig. 4 RMSE comparison results dealing cold-start

4.3.3 相似性度量方法比较

在相同的实验条件下,我们对本文所提方法与其他方法[4,16,28,29]进行性能对比分析.实验采用 5 折交叉验证,训练集和测试集占比分别为 80% 和 20%.其中,文中采用 item-based 推荐模式,相似度阈值 $\delta=0.25$,组合参数 $\alpha=0.85$,具体结果如表 10 所示.在相同情况下,文中所提方法的评分预测准确度明显优于 consine CF、pearsion CF、SVD CF、RBM CF 与 LFM CF 方法.此外,实验随机选择相同目标用户,文中方法的 RMSE 取值为 0.885 877,明显低于 ASARM 所得到的 RMSE.

表 10 相似性度量方法比较

Tab. 10 Comparison of different similarity measure methods

accuracy	Type						
	consine CF	Pearsion CF	SVD CF	RBM CF	ASARM	LFM CF	Our approach
RMSE	0.909 306	0.944 341	0.948 379	0.970 895	0.916 515	0.918 332	0.885 877

5　结论

本文提出了一种新的基于组合相似度的关联推荐方法,实现了一种结合关联规则推荐与协同过滤推荐模式.该方法能够有效地处理稀疏数据和冷启动问题,具有很好的实用价值;同时,基于指定后件项的规则推荐方式能够有效地为用户推荐非活跃项目,直接对目标用户或目标项目进行关联规则挖掘,节省大量规则挖掘执行时间,提高推荐效率.实验结果表明,针对用户很少评分的新项目或稀疏数据问题,文中方法不仅为新用户推荐合适的物品,并且对推荐准确率的提高也起到了积极的作用.在下一步的研究中,我们将针对如何利用人口统计学信息、用户兴趣的描述信息、用户聚类信息等方式来改进推荐质量.

参考文献(References)

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

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

[3] Su X Y, Khoshgoftaar T M. A survey of collaborative filtering techniques [J]. Advances in Artificial Intelligence, 2009, 4: 1-19.

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

[5] Su X Y, Khoshgoftaar T M. Collaborative filtering for multi-class data using belief nets algorithms [C]// Proceedings of the International Conference on Tools with Artificial Intelligence. Arlington, USA: IEEE Computer Society, 2006: 497-504.

[6] Yu K, Schwaighofer A, Tresp V, et al. Probabilistic memory-based collaborative filtering [J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(1): 56-69.

[7] Ben J, Dan F, Jon H. The Adaptive Web: Methods and Strategies of Web Personalization [M]. Berlin Heidelberg: Springer, 2004.

[8] Lü L, Medo M, et al. Recommender systems [J]. Physics Reports, 2012, 519(1): 1-49.

[9] Herlocker J L, Konstan J A, Terveen L G, et al. Evaluating collaborative filtering recommender systems [J]. ACM Transactions on Information Systems, 2004, 22(1): 5-53.

[10] Huang Z, Zeng D, Chen H. A comparative study of recommendation algorithms in e-commerce applications [J]. IEEE Intelligent Systems, 2007, 22(5): 68-78.

[11] García E, Romero C, Ventura S, et al. An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering [J]. User Modeling and User-Adapted Interaction, 2009, 19(1-2): 99-132.

[12] Sarwar B, Karypis G, Konstan J, et al. Analysis of recommendation algorithms for E-commerce [C]// Proceedings of the ACM E-Commerce. NewYork, USA: ACM Press, 2000: 158-167.

[13] Leung C W K, Chan S C F, Chung F L. A collaborative filtering framework based on fuzzy association rules and multi-level similarity [J]. Knowledge and Information Systems, 2006, 10(3): 357-381.

[14] Leung CW K, Chan S C F, Chung F L. Applying cross-level association rule mining to cold-start recommendations[C]// Proceeding of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology Workshops. Silicon Valley, USA: IEEE Press, 2007: 133-136.

[15] LeungC W K, Chan S C F, Chung F L. An empirical study of a cross-level association rule mining approach to cold-start recommendations [J]. Knowledge-Based Systems, 2008, 21(7): 515-529.

[16] Lin W, Alvarez S A, Ruiz C. Efficient adaptive-support association rule mining for recommender systems[J]. Data Mining and Knowledge Discovery, 2014, 6(1): 83-105.

[17] Shaw G, Xu Y, Geva S. Using association rules to solve the cold-start problem in recommender systems [C]// Proceeding of the 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Berlin: Springer, 2014: 340-347.

[18] Sobhanam H, Mariappan A K. Addressing cold start problem in recommender systems using association rules and clustering technique[C]// Proceeding of the International Conference on Computer Communication and Informatics. Coimbatore: IEEE press, 2013:1-5.

[19] Khanzadeh Z, Mahdavi M. Utilizing association rules for improving the performance of collaborative filtering [J]. International Journal of E-Entrepreneurship and Innovation, 2012, 3(2): 14-28.

[20] Tyagi S, Bharadwaj K K. Enhancing collaborative filtering recommendations by utilizing multi-objective particle swarm optimization embedded association rule mining[J]. Swarm and Evolutionary Computation, 2013, 13: 1-12.

[21] Tyagi S, Bharadwaj K K. Enhanced new user recommendations based on quantitative association rule mining[J]. Procedia Computer Science, 2012, 10: 102-109.

[22] Ye H W. A personalized collaborative filtering recommendation using association rules mining and self-organizing map[J]. Journal of Software, 2011, 6(4): 732-739.

[23] Yang H. Improved collaborative filtering recommendation algorithm based on weighted association rules [J]. Applied Mechanics and Materials, 2013, (411-414): 94-97.

[24] 郭晓波，赵书良，王长宾，等. 一种新的面向普通用户的多值属性关联规则可视化挖掘方法[J]. 电子学报，2015，43(2)：344-352.
Guo X B, Zhao S L, Wang C B, et al. A new visualizing mining method of Multi-valued attribute association rules for ordinary users [J]. Acta Electronica Sinica, 2015, 43(23): 344-352.

[25] Sarwar B M, Karypis G, Konstan J A, et al. Item-based collaborative filtering recommendation algorithms [C]// Proceedings of the 10th International Conference on World Wide Web. New York, USA: ACM Press, 2001: 285-295.

[26] Breese J, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[C]// Proceeding of the 14th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann, 1998: 43-52.

[27] MovieLensDataset[EB/OL]. http://www.grouplens.org/datasets/movielens/.

[28] Koren Y. Factorization meets the neighborhood: A multifaceted collaborative filtering model [C]// Proceedings of 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA: ACM Press, 2008: 426-434.

[29] Salakhutdinov R, Mnih A, Hinton G. Restricted Boltzmann machines for collaborative filtering[C]// Proceedings of the 24th International Conference on Machine learning. Corvallis, USA: ACM Press, 2007: 791-798.