计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
50(Suppl.): 163-169, 2013

基于联合聚类平滑的协同过滤算法

韦素云　肖静静　业　宁
(南京林业大学信息科学技术学院　南京　210037)
(weisuyun@163.com)

Collaborative Filtering Algorithm Based on Co-Clustering Smoothing

Wei Suyun, Xiao Jingjing, and Ye Ning
(*College of Information Science and Technology, Nanjing Forestry University, Nanjing* 210037)

Abstract　Collaborative filtering-based recommender systems have become extremely popular in recent years due to the increase in web-based activities such as e-commerce and online content distribution. However, there exist some bottleneck problems, such as sparsity, cold-start and scalability, which limit the development of collaborative filtering. To address the matter, a novel collaborative filtering algorithm based on co-clustering is proposed. First, co-clustering algorithm is used to simultaneously obtain user and item neighborhoods, and then a smooth filling technique is used on rating matrix based on the average ratings of the co-clusters while taking into account the individual biases of the users and items. Lastly, the similarities between the various items are computed based on the smoothing matrix to identify the set of items to be recommended. The experiment results illustrate that item-based collaborative filtering according to co-clustering smoothing the item correlation matrix will become more accurate, which can effectively relieve the impact of sparse data and improve the quality of recommendation.

Key words　recommendation systems; collaborative filtering; item similarity; co-clustering; data smoothing

摘　要　协同过滤是电子商务推荐系统中被广泛采用的技术,但还存在诸如稀疏性、冷启动、可扩展性等制约其进一步发展的瓶颈问题.针对上述问题,提出一种基于联合聚类平滑的协同过滤推荐算法.在该算法中,首先对原始矩阵中的评分模式进行用户和项目2个维度的联合聚类;然后采用联合聚类平滑的方法预测用户对未评分项目的评分值,分别从用户聚类簇、项目聚类簇和联合聚类簇多方面对评分矩阵空缺项进行平滑填充;最后结合基于项目的协同过滤算法查找项目最近邻并进行推荐.实验结果表明,该算法可以有效缓解用户评分数据稀疏带来的不良影响,一定程度上解决冷启动问题,提高预测准确率和推荐质量.

关键词　推荐系统;协同过滤;项目相似性;联合聚类;数据平滑

中图法分类号　TP391.4

智能推荐系统是个性化信息服务的重要组成部分,可以实现主动地为用户推荐感兴趣的信息.随着互联网上信息的增长和用户个性化需求的提高,推荐系统的应用日益广泛,成为电子商务、社会网络、

收稿日期:2013-11-26
基金项目:江苏省"六大人才高峰"基金项目(2011DZXX043);江苏省自然科学基金项目(BK2012815);江苏省高等学校大学生实践创新项目(201310298036Z)

视频/音乐点播等个性化服务的核心技术. 协同过滤[1]是个性化推荐中研究和应用最多的一种技术,使用统计技术寻找与目标用户有相同或相似兴趣偏好的邻居用户,根据邻居用户的评分来预测目标用户对商品项的评分值,选择预测评分最高的前 N 项商品作为推荐集反馈给目标用户.

随着电子商务站点中用户和商品项的数量不断增加,协同过滤面临严峻的用户评分数据稀疏性和推荐实时性挑战,导致推荐质量迅速下降. 为了产生精确的推荐,保证推荐系统的实时性要求,研究人员将统计分析、机器学习等领域的方法与协同过滤相结合,产生了基于模型的协同过滤算法,主要思想是通过已知的评分数据来确定模型中的参数,用参数确定的模型来预测未知的评分数据. 最常见的模型有贝叶斯(Bayesian)模型[2]、聚类(Clustering)模型[3]、回归(Regression)模型[4]和隐含语义(Latent Semantic)模型[5]等. 其中聚类模型的主要优势在于降低数据维度,提高计算效率,数据之间的簇信息在一定程度上解决了数据稀疏性和数据动态性的问题. 目前对协同过滤中应用聚类方法的研究比较多,通常采用的有 K-means 聚类、模糊聚类[6-11]等算法. 文献[6]提出在协同过滤推荐算法中使用融合聚类技术,一定程度上缓解了数据稀疏性问题,但也存在对处于聚类边缘的用户推荐精度比较低的问题. 文献[7]中根据聚类信息进行平滑处理,对用户未评分的项目赋予初始的平滑值,一定程度上解决数据稀疏的影响,提高推荐精度. 文献[8]提出了一种基于项目聚类的协同过滤推荐算法,根据用户对项目评分的相似性对项目进行聚类,在与目标项目最相似的前若干个聚类簇中搜索它的最近邻居. 文献[9]通过模糊聚类的方法将用户和项目属性特征的相似性与协同过滤推荐技术相结合,在对用户和项目进行模糊聚类后得到用户和项目在属性特征上的相似关系簇. 文献[10]采用聚类方法对稀疏矩阵进行划分,使用基于划分的 K-means 聚类算法,将用户兴趣划分到不同类别中,以缩小近邻搜索范围和需要预测的资源数目,减少数据稀疏性. 文献[11]提出了进化联合聚类方法,提高了预测的性能,同时保持联合聚类在联机阶段的可扩展性.

然而基于聚类的协同过滤算法大多都只是从用户或者项目单方面角度出发,没有把客户和项目结合起来作为一个目标. 本文提出基于信息论联合聚类的协同过滤推荐方法,以一种更加全面的方式来克服协同过滤各方面的不足,得到较满意的实验效果. 本文的创新点主要体现在以下方面:1)使用基于信息论的联合聚类算法同时对项目和用户进行聚类,加强用户和项目之间的联系,提高预测准确性;2)采用联合聚类平滑技术预测用户对未评分项目的评分值,分别从用户聚类簇、项目聚类簇和联合聚类簇多方面对评分矩阵空缺项进行平滑填充,缓解协同过滤中数据稀疏性和冷启动问题;3)预测推荐阶段,在平滑填充后的评分矩阵上搜寻目标项目的最近邻居并进行推荐,一定程度上缓解实时推荐问题.

1 相关工作

1.1 协同过滤

协同过滤算法根据基本用户的观点产生对目标用户的推荐列表,它基于这样的假设:如果用户对一些项目的评分比较相似,则他们对其他项目的评分也将会比较相似. 协同过滤推荐系统首先搜索目标用户的若干个最近邻居,然后根据最近邻居对项目的评分预测目标用户对项目的评分,从而产生推荐列表.

定义 1. 推荐系统中的数据源 $D=(U,P,R)$,其中 $U=\{u_i\}_{i=1}^{m}$ 是用户集合,$|U|=m$;$P=\{p_j\}_{j=1}^{n}$ 是项目集合,$|P|=n$;$m\times n$ 阶矩阵 $\boldsymbol{R}$ 是用户对各项目的评分矩阵,其中的元素 r_{ij} 表示用户 u_i 对项目 p_j 的评分.

在现实环境中,由于 m 和 n 的值会比较大,用户不可能对所有的项目都逐个进行评分,因而在用户评分数据矩阵 $\boldsymbol{R}$ 中存在大量的空值,即对应用户没有对该项目进行评分. 随着系统规模的扩大,用户数和项目数迅速增加,$\boldsymbol{R}$ 中空值的数目也将急剧增加,从而导致评分数据极端稀疏.

为了找到目标用户的最近邻居,必须度量用户之间的相似性,选择相似性最高的若干用户作为目标用户的最近邻居. 常见的最近邻居度量标准包括余弦相似性、相关相似性和修正的余弦相似性.

定义 2. 相关相似性,又称为 Pearson 相关相似性,设 S 表示对用户 u_i 和 $u_{i'}$ 有共同评分的项目集合,则用户 u_i 和用户 $u_{i'}$ 之间的相似性 $sim(u_i,u_{i'})$ 如式(1)所示:

$$sim(u_i,u_{i'})=\frac{\sum_{p_j\in S}(r_{ij}-\bar{r}_i)(r_{i'j}-\bar{r}_{i'})}{\sqrt{\sum_{p_j\in S}(r_{ij}-\bar{r}_i)^2}\sqrt{\sum_{p_j\in S}(r_{i'j}-\bar{r}_{i'})^2}},\tag{1}$$

其中,r_{ij} 表示用户 u_i 对项目 p_j 的评分,$r_{i'j}$ 表示用户 $u_{i'}$ 对项目 p_j 的评分,$\bar{r}_i$ 和 $\bar{r}_{i'}$ 分别表示用户 u_i 和 $u_{i'}$ 对项目的平均评分.

定义 3. 已知数据源 $D=(U,P,R)$,给定目标用户 u_a,对于 $\forall u_i \in U$,将相似性 $sim(u_a,u_i)$ 最大的 Kn 个用户组成集合 $Nu_a=\{u_{i1},u_{i2},\cdots,u_{iKn}\}$,$u_a \notin Nu_a$,则称该集合中的元素为目标用户 u_a 的 Kn 个最近邻居.

定义 4. 已知数据源 $D=(U,P,R)$,给定目标用户 u_a 及其最近邻居集合 Nu_a,则用户 u_a 对项目 p_j 的预测评分为

$$P_{aj}=\bar{r}_a+\frac{\sum_{u_i\in Nu_a} sim(u_a,u_i)\times(r_{ij}-\bar{r}_i)}{\sum_{u_i\in Nu_a}|sim(u_a,u_i)|}, \quad (2)$$

其中,$\bar{r}_a$ 和 $\bar{r}_i$ 分别表示用户 u_a 和 u_i 对项目的平均评分值.

1.2 联合聚类

联合聚类(co-clustering)算法是在数据矩阵的行和列 2 个方向上同时进行聚类,其目的是发现高度相关子空间内的簇集,即对相似的局部子模式进行聚类.联合聚类在基因表达数据分析、电子商务数据分析、协同过滤[12-15]等领域有着很好的应用前景.

定义 5. 给定推荐数据源 $D=(U,P,R)$,如果 $I\subseteq U$ 和 $J\subseteq Y$ 分别表示用户集和项目集的子集,则矩阵 $\boldsymbol{R}_{IJ}=(I,J)$ 构成了原矩阵 $\boldsymbol{R}$ 的子矩阵.如果矩阵 $\boldsymbol{R}_{IJ}$ 中元素满足一些特定的同源特性,则矩阵 $\boldsymbol{R}_{IJ}=(I,J)$ 为原矩阵的一个联合聚类.

联合聚类的基本原理是通过行聚类和列聚类 2 个步骤进行循环迭代直至收敛.Cheng 等人[16]首次提出了联合聚类模型,并以最小均方残值为聚类标准,该算法每次运行只能得到部分聚类结果,并且聚类结果是不确定的,被称为非确定性算法.Dhillon 等人[17]引入信息熵理论,提出了一种以 Kullback-Leibler(KL)距离最小为标准的联合聚类方法.Banerjee 等人[13]提出了一种同时考虑类别内部均值和行与列向量全局均值的联合聚类方法,并且给出了该方法在 MovieLens 数据集上进行评分预测的结果.通常情况下,联合聚类问题是 NP-Hard 问题[17],穷举全部局部簇集的时间复杂度一般是指数级,因此如何有效发现更有意义的局部聚类和提高算法性能始终是联合聚类问题而临的挑战.

基于信息论的联合聚类算法[17]是将标准化后的矩阵看作一个联合概率分布,将联合聚类问题转换为信息论的最优化问题,最优的结果为最大化聚类后的随机变量间的互信息聚类.反之,最优的聚类结果就是最小化初始的随机变量与聚类后的随机变量间的互信息损失,聚类过程以最小化一个损失函数为目标.

假设 X,Y 表示 2 个随机变量,其值的集合分别用 $\{x_u\}$,$[u]_1^m$ 和 $\{y_v\}$,$[v]_1^n$ 表示,其中 $[u]_1^m$ 表示 u 的变化范围是 $\{1,\cdots,m\}$,$[v]_1^n$ 表示 v 的变化范围是 $\{1,\cdots,n\}$,假定 $\{\hat{x}_g\}$,$[g]_1^k$ 表示各个行簇,$\{\hat{y}_h\}$,$[h]_1^l$ 表示各个列簇,$\{\hat{x}_g\}$,$[g]_1^k$,$\{\hat{y}_h\}$,$[h]_1^l$ 均不相交,即:

$$\bigcup_{i=1}^{k}\hat{x}=\{x_1,\cdots,x_m\},\ \hat{x}_i\bigcap\hat{x}_j=\varnothing,\ i\neq j;$$

$$\bigcup_{i=1}^{l}\hat{y}=\{y_1,\cdots,y_n\},\ \hat{y}_i\bigcap\hat{y}_j=\varnothing,\ i\neq j.$$

聚类的目标是结果尽可能地保留原始矩阵的"信息量",即尽量保留随机变量 X,Y 之间的互信息,使得聚类前后随机变量 X,Y 之间的互信息损失最少.那么目标函数定义如下:

$$\min_{\hat{X},\hat{Y}}(I(X,Y)-I(\hat{X},\hat{Y})). \quad (3)$$

因此,基于信息论的联合聚类算法寻找最优解的问题转化成了寻找一个对 $I(X,Y)$ 最逼近的概率分布 $I(\hat{X},\hat{Y})$ 的问题.

2 基于联合聚类平滑的协同过滤算法

2.1 用户和项目联合聚类

推荐系统的主要问题是预测目标用户对项目的未知评分,给定数据源 $D=(U,P,R)$,即预测评分矩阵 $\boldsymbol{R}$ 中缺失元素的值.可以将此问题转换成加权矩阵逼近问题,并且利用基于联合聚类的方法去解决它.设 $\rho:\{1,\cdots,m\}\mapsto\{1,\cdots,k\}$ 和 $\gamma:\{1,\cdots,n\}\mapsto\{1,\cdots,l\}$ 表示对用户和项目进行聚类,其中用户被聚成 k 类,项目被聚成 l 类.定义 $m\times n$ 阶评分标识矩阵 $\boldsymbol{W}$,当用户 u_i 对项目 p_j 已做出评分,即 $r_{ij}\neq 0$ 时,矩阵 $\boldsymbol{W}$ 中的元素 $w_{ij}=1$,否则 $w_{ij}=0$.为了进一步提高推荐的精确性,本文在定义 $m\times n$ 阶逼近矩阵 $\hat{\boldsymbol{R}}$ 时,除了考虑联合聚类均值之外,还融合了用户和项目的偏好,包括用户(项目)均值和用户(项目)簇均值,元素 $\hat{r}_{ij}$ 定义如下:

$$\hat{r}_{ij}=r_{gh}^{COC}+(r_i^{R}-r_g^{RC})+(r_j^{C}-r_h^{CC}), \quad (4)$$

其中,$g=\rho(i)$,$h=\gamma(j)$,r_i^{R},r_j^{C} 分别是用户 u_i 和项目 p_j 的平均评分值,r_{gh}^{COC},r_g^{RC} 和 r_h^{CC} 分别是用户-项目联合聚类簇、用户聚类簇和项目聚类簇的平均评分值,即:

$$r_{gh}^{\mathrm{COC}}=\frac{\sum_{i'|\rho(i')=g}\sum_{j'|\gamma(j')=h}r_{i'j'}}{\sum_{i'|\rho(i')=g}\sum_{j'|\gamma(j')=h}w_{i'j'}},\ r_i^{\mathrm{R}}=\frac{\sum_{j'=1}^{n}r_{ij'}}{\sum_{j'=1}^{n}w_{ij'}},$$

$$r_j^{\mathrm{C}}=\frac{\sum_{i'=1}^{m}r_{i'j}}{\sum_{i'=1}^{m}w_{i'j}},r_g^{\mathrm{RC}}=\frac{\sum_{i'|\rho(i')=g}\sum_{j'=1}^{n}r_{i'j'}}{\sum_{i'|\rho(i')=g}\sum_{j'=1}^{n}w_{i'j'}},$$

$$r_h^{\mathrm{CC}}=\frac{\sum_{i'=1}^{m}\sum_{j'|\gamma(j')=h}r_{i'j'}}{\sum_{i'=1}^{m}\sum_{j'|\gamma(j')=h}w_{i'j'}}.$$

将预测未知评分作为联合聚类问题,找到最优的用户和项目聚类(ρ,γ),使得评分矩阵 $\boldsymbol{R}$ 与逼近矩阵 $\hat{\boldsymbol{R}}$ 之间的近似误差最小化,即:

$$\min_{(\rho,\gamma)}\sum_{i=1}^{m}\sum_{j=1}^{n}w_{ij}(r_{ij}-\hat{r}_{ij})^2. \tag{5}$$

具体来说,为了对用户-项目评分矩阵进行联合聚类,首先扫描评分矩阵,计算出联合聚类簇均值 r^{COC}、用户聚类簇均值 r^{RC}、项目聚类簇均值 r^{CC}、用户评分均值 r^{R} 和项目评分均值 r^{C},然后采用双向聚类交叉迭代算法,调整用户聚类和项目聚类结果直至收敛.

算法 1. 用户和项目联合聚类算法.

输入:用户-项目评分矩阵 $\boldsymbol{R}$、评分标识矩阵 $\boldsymbol{W}$、用户(行)聚类数目 k、项目(列)聚类数目 l;

输出:用户和项目联合聚类(ρ,γ)、各个聚类簇均值 $r^{\mathrm{COC}},r^{\mathrm{RC}},r^{\mathrm{CC}},r^{\mathrm{R}}$ 和 r^{C}.

步骤 1. 初始化. 随机划分行和列,产生初始用户和项目联合聚类(ρ,γ);

步骤 2. 计算各个聚类簇均值. $r^{\mathrm{COC}},r^{\mathrm{RC}},r^{\mathrm{CC}},r^{\mathrm{R}}$ 和 r^{C};

步骤 3. 更新行聚类:

$$\rho(i)=\arg\min_{1\leqslant g\leqslant k}\sum_{j=1}^{n}w_{ij}(r_{ij}-r_{g\gamma(j)}^{\mathrm{COC}}-r_i^{\mathrm{R}}+r_g^{\mathrm{RC}}-r_j^{\mathrm{C}}+r_{\gamma(j)}^{\mathrm{CC}})^2,\ 1\leqslant i\leqslant m;$$

步骤 4. 更新列聚类:

$$\gamma(j)=\arg\min_{1\leqslant h\leqslant l}\sum_{i=1}^{m}w_{ij}(r_{ij}-r_{\rho(i)h}^{\mathrm{COC}}-r_i^{\mathrm{R}}+r_{\rho(i)}^{\mathrm{RC}}-r_j^{\mathrm{C}}+r_h^{\mathrm{CC}})^2,\ 1\leqslant j\leqslant n;$$

步骤 5. 终止条件. 计算损失值 $\sum_{i=1}^{m}\sum_{j=1}^{n}w_{ij}(r_{ij}-\hat{r}_{ij})^2$,若损失函数的变化值小于某预先定义的阈值,则停止迭代,输出结果;否则转向步骤 2,继续迭代运行.

2.2 聚类平滑

协同过滤完全依赖于用户评分,通过构建用户-项目评分矩阵,使用统计技术寻找邻居用户并进行推荐. 但是由于电子商务站点用户及商品项的数量庞大且不断增加,使得评分矩阵成为高维矩阵,同时用户给予评分的商品项很少,通常在 1%以下,导致评分数据极端稀疏. 协同过滤的稀疏性(sparsity)问题由此产生,并成为推荐质量下降的主要原因. 冷启动(cold-start)问题是稀疏性的极端情况,当一个新用户(新项目)进入推荐系统后,由于还未提供任何项目(用户)的评分,导致系统无法向新用户推荐其可能喜欢的项目或将新项目推荐给可能喜欢它的用户.

为了缓解数据稀疏对预测性能的影响,使得评分矩阵变得稠密,目前提出了多种评分矩阵填充方法. 最简单的方式是将所有未评分项目用缺省值"0"来代替,这种方法的推荐精度很差. 另一种较为普遍的填充方式是使用矩阵中已评分项的平均值替换未评分项目值,但是填充方法仍然太过笼统. 本文采用联合聚类平滑的方法预测用户对未评分项目的评分值,分别从用户聚类簇、项目聚类簇和联合聚类簇多方面对矩阵空缺项进行平滑填充,克服传统的单一填充方法的片面性. 同时利用联合聚类簇中的各个均值,可以对新用户或新项目预测评分,一定程度上解决了冷启动问题.

算法 2. 联合聚类平滑算法.

输入:用户-项目评分矩阵 $\boldsymbol{R}$,矩阵 $\boldsymbol{R}$ 的均值 $\bar{R}$,用户和项目联合聚类(ρ,γ),各个聚类簇均值 $r^{\mathrm{COC}},r^{\mathrm{RC}},r^{\mathrm{CC}},r^{\mathrm{R}}$ 和 r^{C},用户集 U,项目集 P,用户 u_i,项目 p_j;

输出:填充后的评分矩阵 $\boldsymbol{R}$.

过程:

情况 1. 若 $u_i\in U,p_j\in P$ 且 $r_{ij}=0$,则 $r_{ij}=r_i^{\mathrm{R}}+r_j^{\mathrm{C}}-r_{\rho(i)}^{\mathrm{RC}}-r_{\gamma(j)}^{\mathrm{CC}}+r_{\rho(i)\gamma(j)}^{\mathrm{COC}}$;

情况 2. 若 $u_i\in U$ 且 $p_j\notin P$,则 $r_{ij}=r_i^{\mathrm{R}}$;

情况 3. 若 $u_i\notin U$ 且 $p_j\in P$,则 $r_{ij}=r_j^{\mathrm{C}}$;

情况 4. 若 $u_i\notin U$ 且 $p_j\notin P$,则 $r_{ij}=\bar{R}$.

2.3 基于项目的协同过滤

传统的基于用户的协同过滤算法存在扩展性差的问题,推荐效率随用户数目、项目数目的增多而明显降低. 由于项目间的相似性比较稳定,而且可以离线计算,基于项目的协同过滤算法[18]一定程度上缓解了实时推荐问题. 基于项目的协同过滤的基本思想是:用户对项目的评分可以通过项目之间的相似性以及用户已有的评分进行预测. 本文将联合聚类

平滑与基于项目的协同过滤结合,形成基于联合聚类和项目的协同过滤算法,在离线数据预处理阶段,分别从用户和项目2个方面进行联合聚类,利用聚类结果对原始评分数据进行平滑处理;在预测推荐阶段,搜寻目标项目的最近邻居项目集合,然后根据最近邻居项目的评分来预测其对未评分项目的评分,进而选择预测评分最高的前 N 项作为推荐结果反馈给用户.

项目间相似性计算公式与用户间相似性计算公式类似,以相关相似性计算为例,项目 p_j 和项目 $p_{j'}$ 之间的相关相似性 $sim(p_j,p_{j'})$ 如(6)所示:

$$sim(p_j,p_{j'})=\frac{\sum_{u_i\in U}(r_{ij}-r_j^{\mathrm{C}})(r_{ij'}-r_{j'}^{\mathrm{C}})}{\sqrt{\sum_{u_i\in U}(r_{ij}-r_j^{\mathrm{C}})^2}\sqrt{\sum_{u_i\in U}(r_{ij'}-r_{j'}^{\mathrm{C}})^2}},\tag{6}$$

其中,r_j^{C} 是项目 p_j 的平均评分值.

计算项目 p_j 和其余项目间的相似性,相似性最大的 Kn 个项目组成项目 p_j 的最近邻居集 Np_j,则目标用户 u_a 对项目 p_j 的预测评分为

$$P_{aj}=r_a^{\mathrm{R}}+\frac{\sum_{j'\in Np_j}sim(p_j,p_{j'})\times(r_{aj'}-r_{j'}^{\mathrm{C}})}{\sum_{j'\in Np_j}|sim(p_j,p_{j'})|},\tag{7}$$

其中,r_a^{R} 和 $r_{j'}^{\mathrm{C}}$ 分别是用户 u_a 和项目 $p_{j'}$ 的平均评分值.计算用户 u_a 对各个未评分项目的预测评分,取评分值最高的前 N 个项目作为推荐集.

算法3. 基于联合聚类平滑的协同过滤算法(CoC-CF).

输入:用户-项目评分矩阵 $\boldsymbol{R}$、用户集 U、项目集 P、目标用户 u_a、推荐集元素个数 N;

输出:目标用户 u_a 的推荐集 top-N.

步骤1. 用户和项目联合聚类.扫描评分矩阵 $\boldsymbol{R}$,创建辅助矩阵 $\boldsymbol{W}$,根据算法1生成用户和项目联合聚类(ρ,γ)及各个聚类簇均值 r^{COC},r^{RC},r^{CC},r^{R} 和 r^{C};

步骤2. 对评分矩阵进行平滑填充.利用联合聚类结果,根据算法2对用户-项目评分矩阵 $\boldsymbol{R}$ 中未评分元素进行平滑填充;

步骤3. 计算相似性.对于任意不同项目 p_j 和 $p_{j'}$ 利用式(6)计算项目相似性 $sim(p_j,p_{j'})$;

步骤4. 生成最近邻居.对于每个项目 p_j 寻找最近邻居集 Np_j;

步骤5. 产生推荐.利用式(7)计算目标用户 u_a 对未评分项目 i 的预测评分 P_{ai},取前 N 个值所对应的项目组成推荐集 top-N.

3 实验结果与分析

3.1 数据集和度量标准

本文实验采用欧莱雅旗下的子化妆品牌 THE BODY SHOP 美体小铺提供的数据集.这个数据集包括600多种头发和皮肤的美容保养品,另外还有400多样产品附件,其中包括脸部、身体及头发清洁用品、护肤保养品、香氛、香水油、精油及彩妆等.本文选取475 154个用户对3 856个产品购买近7年总共9 548 825条记录.实际评分数据的密度为9 548 825/(475 154×3 856)=0.52%,说明此数据是相当稀疏的.

评价推荐系统推荐质量的度量标准主要包括统计精度度量方法和决策支持精度度量方法2类.统计精度度量方法中的平均绝对误差(mean absolute error, MAE)是一种常用的衡量推荐结果的度量方法.该标准是通过比较预测值与用户实际的评分值之间的偏差来衡量预测结果的准确性.MAE 越小,表明推荐质量越高.绝大多数实验都采用 MAE 作为衡量推荐结果的参考标准,因此本文也采用平均绝对偏差 MAE 作为度量标准.设预测的用户评分集合表示为$\{p_1,p_2,\cdots,p_n\}$,对应的实际用户评分集合为$\{q_1,q_2,\cdots,q_n\}$,则平均绝对偏差 MAE 定义如下:

$$MAE=\frac{\sum_{i=1}^{n}|p_i-q_i|}{n}.\tag{8}$$

3.2 聚类数目的选取

联合聚类中的主要参数是用户和项目的聚类数目,初始用户聚类数目 k 和项目聚类数目 l 需要事先指定,聚类数目的选取对实验结果有一定的影响.为了选取合适的聚类数,我们选用不同的用户聚类数目和项目聚类数目,采用5折交叉验证,同时与文献[12]提出的直接利用联合聚类进行评分和预测的推荐算法(COCLUST)进行对比实验,分析平均绝对误差 MAE 的变化情况.图1是在项目聚类数目给定为5的情况下,用户聚类数目 k 从4递增到35时基于联合聚类的协同过滤算法(CoC-CF)的平均绝对偏差 MAE 的变化情况.图2是在用户聚类数目给定为5的情况下,项目聚类数目 l 从4递增到35时 MAE 的变化情况.

由图1,2可见,使用联合聚类对原始矩阵进行平滑,进而进行基于项目的协同过滤推荐,实验结果随着聚类数目的不同而变化.当聚类数过小时,类信

息过于普遍化无法表示不相似用户和项目之间的不同性;而当聚类数过大时,类信息过于个性化无法表示相似用户和项目间的相似性. 当用户聚类数和项目聚类数为 10 时,预测性能最优.

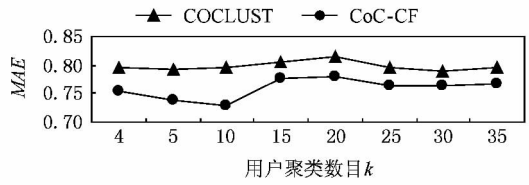

图 1 用户聚类数目 k 对 MAE 的影响

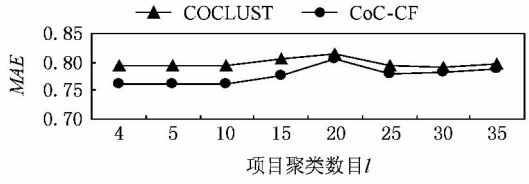

图 2 项目聚类数目 l 对 MAE 的影响

3.3 实验结果比较

最近邻居的个数会在很大程度上影响算法的性能,在实验中,将最近邻居的个数从 5 递增到 200. 计算本文提出的基于联合聚类平滑的协同过滤推荐算法(CoC-CF)的 MAE 值和基于聚类平滑的利用相关相似性度量的协同过滤算法(SCBPCC)[7]、传统的基于项目的协同过滤算法(IBCF)的 MAE 值,并进行比较分析,实验参数按 3.2 节讨论的最优值来设置.

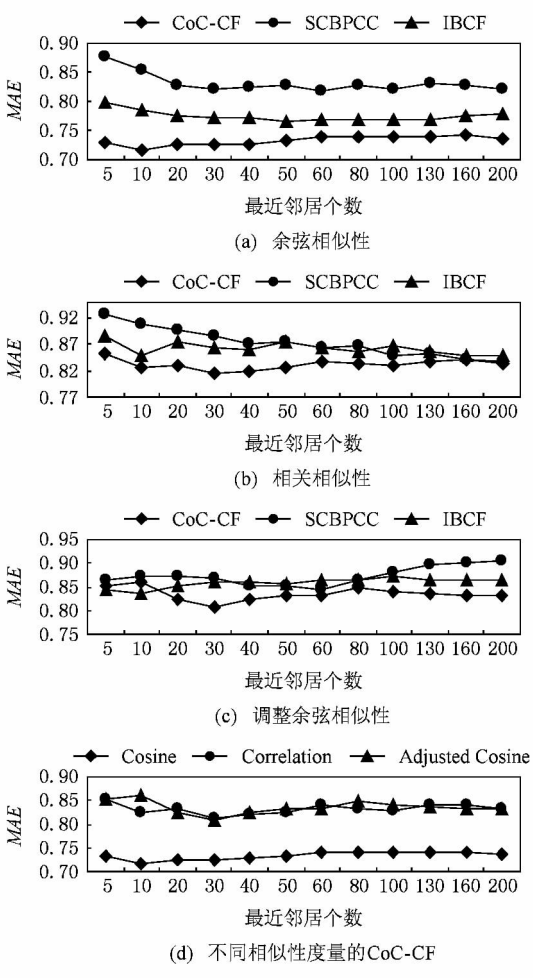

图 3 不同推荐算法的 MAE 结果比较

本文分别从基于项目的余弦相似性度量方法、基于项目的相关相似性度量方法和基于项目的调整余弦相似性度量方法 3 个方面分别进行对比实验. 由图 3 可知,本文提出的基于联合聚类平滑的协同过滤推荐算法与其他相关的协同过滤算法相比,具有较小的 MAE,推荐质量有了较大的提高,其中基于余弦相似性度量的优势最为明显,具有一定的稳定性. 这是由于在计算项目相似性之前,利用用户聚类簇、项目聚类簇和联合聚类簇多方面对矩阵空缺项进行平滑填充,克服传统的单一填充方法的片面性,缓解数据稀疏性对推荐质量的影响,提高预测结果的准确性.

4 结束语

随着电子商务规模的日益增长,用户和项目数量急剧增加,推荐系统的实时性、数据稀疏性和可扩展性问题也随之加剧. 本文提出联合聚类的协同过滤算法,通过调整用户聚类和项目聚类,可以在一定程度上解决数据稀疏和冷启动问题,提高推荐质量.

下一步的工作是进一步对算法进行改进,由于采用联合聚类的预测评分可靠度不如原始评分高,在基于项目的协同过滤推荐过程中,需要考虑到预测评分的权值,这可以通过评分标识矩阵 W 来实现,保证推荐结果的可靠性. 另外如何利用联合聚类进行降维、增量式更新及降低推荐算法的复杂度,也是十分有趣和极具挑战性的研究方向.

参 考 文 献

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

[2] Xiong L, Chen X, Huang T K, et al. Temporal collaborative filtering with bayesian probabilistic tensor factorization //Proc of SIAM Data Mining. Columbus: SIAM, 2010: 211-222

[3] Ungar L H, Foster D P. Clustering methods for collaborative filtering //AAAI Workshop on Recommendation Systems. Menlo Park, CA: AAAI, 1998: 114-129

[4] Vucetic S, Obradovic Z. Collaborative filtering using a regression-based approach. Knowledge and Information Systems, 2005, 7(1): 1-22

[5] Hofmann T. Latent semantic models for collaborative filtering. ACM Trans on Information Systems, 2004, 22(1): 89-115

[6] Tsai C F, Hung C. Cluster ensembles in collaborative filtering recommendation. Applied Soft Computing, 2012, 12(4): 1417-1425

[7] Xue G R, Lin C, Yang Q, et al. Scalable collaborative filtering using cluster-based smoothing //Proc of the 28th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2005: 114-121

[8] Gong S. An efficient collaborative recommendation algorithm based on item clustering //Proc of Advances in Wireless Networks and Information Systems. Berlin: Springer, 2010: 381-387

[9] Honda K, OH C H, Matsumoto Y, et al. Exclusive Partition in FCM-type Co-clustering and Its Application to Collaborative Filtering. International Journal of Computer Science and Network Security, 2012, 12(12): 52-58

[10] Dakhel G M, Mahdavi M. A new collaborative filtering algorithm using *K*-means clustering and neighbors' voting // Proc of the 11th Conf on Hybrid Intelligent Systems (HIS). Piscataway, NJ: IEEE, 2011: 179-184

[11] Khoshneshin M, Street W N. Incremental collaborative filtering via evolutionary co-clustering //Proc of the 4th ACM Conf on Recommender Systems. New York: ACM, 2010: 325-328

[12] George T, Merugu S. A scalable collaborative filtering framework based on co-clustering //Proc of the 5th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2005: 625-628

[13] Banerjee A, Dhillon I, Ghosh J, et al. A generalized maximum entropy approach to bregman co-clustering and matrix approximation //Proc of the 10th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2004: 509-514

[14] **吴湖, 王永吉, 王哲, 等. 两阶段联合聚类协同过滤算法: 软件学报**, 2011, 21(5): 1042-1054

[15] Honda K, Muranishi M, Notsu A, et al. FCM-type cluster validation in fuzzy co-clustering and collaborative filtering applicability. International Journal of Computer Science and Network Security, 2013, 13(1): 24-29

[16] Cheng Y, Church G M. Biclustering of expression data // Proc of the 8th Int Conf on Intelligent Systems for Molecular Biology. Menlo Park, CA: AAAI, 2000: 93-103

[17] Dhillon I S, Mallela S, Modha D S. Information-theoretic co-clustering //Proc of the 9th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2003: 89-98

[18] Deshpande M, Karypis G. Item-based top-*n* recommendation algorithms. ACM Trans on Information Systems, 2004, 22(1): 143-177

韦素云　女,1981 年生,硕士,讲师,主要研究方向为数据挖掘、个性化推荐技术.

肖静静　女,1993 年生,学士,主要研究方向为数据挖掘.

业　宁　男,1967 年生,教授,主要研究方向为模式识别、智能处理、生物信息学.