不确定近邻的协同过滤推荐算法

黄创光[1)] 印 鉴[1)] 汪 静[1),2)] 刘玉葆[1)] 王甲海[1)]

[1)](中山大学信息科学与技术学院 广州 510006)

[2)](南海东软信息技术职业学院 广东 佛山 528225)

摘 要 文中围绕传统的协同过滤推荐算法存在的局限性展开研究,提出一种不确定近邻的协同过滤推荐算法 UNCF.根据推荐系统应用的实际情况,对于推荐的每一种场景其实都是不可预先确定的,而文中算法基于用户以及产品的相似性计算,自适应地选择预测目标的近邻对象作为推荐群,同时计算推荐群中推荐把握概率较高的信任子群,最后通过不确定近邻的动态度量方法,来对预测结果进行平衡的推荐.通过实验结果表明,该算法可以有效平衡用户群以及产品群推荐结果所带来的不稳定影响,有效缓解用户评分数据稀疏的情况所带来的问题,并在多个实验数据中,提高了推荐系统的预测准确率.

关键词 不确定近邻;协同过滤;推荐系统;相似性度量;信任子群

中图法分类号 TP301 **DOI** 号:10.3724/SP.J.1016.2010.01369

Uncertain Neighbors' Collaborative Filtering Recommendation Algorithm

HUANG Chuang-Guang[1)] YIN Jian[1)] WANG Jing[1),2)] LIU Yu-Bao[1)] WANG Jia-Hai[1)]

[1)](*School of Information Science and Technology, Sun Yat-Sen University, Guangzhou* 510006)

[2)](*Neusoft Institute of Information, Nanhai, Foshan, Guangdong* 528225)

Abstract To overcome several limitations in the research area of collaborative filtering (CF), this paper presents a CF recommendation algorithm, named UNCF(Uncertain Neighbors' Collaborative Filtering Recommendation Algorithm). In the reality, the scene of recommendation is uncertain. The similarities computations of both user-based and item-based are considered to choose the neighbors dynamically as the recommendation set. This set can be used to select the trustworthy subset which is the most effective objects to the predicted result. Moreover, this paper defines a new prediction algorithm that combines the advantages of trustworthy subset for this uncertain recommendation method. Through experimental results, the UNCF algorithm can consistently achieve better prediction accuracy than traditional CF algorithms, and effectively leverage the result in the uncertain environment. Furthermore, the algorithm can alleviate the dataset sparsity problem.

Keywords uncertain neighbors; collaborative filtering; recommendation system; similarity criterion; trustworthy subset

收稿日期:2010-06-11.本课题得到国家自然科学基金(60773198, 60703111)、广东省自然科学基金(7300272, 8151027501000021)、国家科技计划项目(2008ZX10005-013)、广东省科技计划项目(2008B050100040, 2009A080207005, 2009B090300450)、新世纪优秀人才支持计划(NCET-06-0727)资助.黄创光,男,1978 年生,博士研究生,研究方向为数据挖掘在客户行为分析和推荐系统上的应用.E-mail:13316096336@189.cn.印 鉴,男,1968 年生,博士,教授,博士生导师,主要研究领域为机器学习和数据挖掘.汪 静,女,1980 年生,博士研究生,研究方向为个性化推荐.刘玉葆,男,1975 年生,博士,副教授,主要研究方向为数据库、数据仓库和数据挖掘.王甲海,男,1977 年生,博士,副教授,研究方向为人工智能和数据挖掘.

1 引 言

互联网的迅速发展以及信息量的暴涨，为用户提供主动推荐的功能越来越多地被应用到各种门户网站和电子商务系统中，来帮助人们更准确、高效地寻找到所需要的信息，这些系统通常会包括用户过去的点击日志、评分和评论等信息，通过这些历史的信息来分析用户的特征并提供推荐服务，给访问者主动推荐最可能感兴趣的信息.

采用协同过滤推荐技术的应用领域非常广泛，Tapestry[1] 是最早的推荐系统之一，在这个系统中，记录了每个用户对他们阅读文章的观点，这些观点可以被其他用户进行获取. GroupLens[2]、Ringo[3] 和 Video Recommender[4] 也是较早期的推荐系统，通过协同过滤思想，通过其他人的意见来给用户分别提供新闻、音乐和视频等推荐服务. 推荐系统中普遍存在的 3 大问题：数据稀疏性[5]、冷启动[6] 和可扩展性[7] 等问题，随着协同过滤推荐系统应用的不断深入，许多研究者提出了一些新的方法来改进基于规则或基于模型过滤的推荐系统的不足，如 Park[8] 等通过协同过滤结合搜索工具提升Yahoo !的推荐结果；Tomoharu[9] 等通过基于最大熵原理协同过滤来推荐用户最可能购买的商品；Yang[10] 等提出的基于推论的协同过滤方法；Chen[11] 等结合推荐结果的收益进行协同过滤推荐目标用户可能感兴趣的商品. 推荐系统近期一个重要的研究趋势是研究个体与个体之间关系的拓展和延伸，例如 Chen[12] 等通过建立 k 最近邻及其影响集来计算预测的评分；Liu[13] 等通过建立信任群的方式来构建一个Beta 分布，来预测用户的相似程度；Jamali[14] 等将用户之间的信任关系进行深度搜索，寻找更深层次的相似用户来进行推荐.

当前基于协同过滤的推荐研究，主要是针对用户群对产品群的评分预测. 往往由于用户与用户个体之间的差异，造成基于用户推荐效果的差异. 单对用户群而言，对于每一个预测场景的某个用户而言，用户不固定的，同时预测的产品对象也是不固定的，由于预测场景的不确定性，传统方法存在一些局限性：(1)协同过滤研究中，很多采用 kNN 方法[2, 11, 15] 来为预测的目标选择推荐对象，也就是选择 k 个最近的邻居作为推荐对象，这个方法选择的参数 k 通常具有一定的普遍性，但缺乏特殊性，对具体某个推荐对象而言，这 k 个推荐对象有可能属于另一个群，而不一定都属于预测目标的近邻群，而且未知用户或产品的相似性近邻数量也不是可以固定的；(2)传统的推荐往往只考虑了某一群的影响作用而忽视另一个关联群的影响作用，或者过分关注某一群体的影响作用[8-9, 15-16]，由于我们在对未知目标进行推荐之前，我们对他们相关联的近邻对象缺乏足够的分析和认识，从而导致推荐质量的降低.

本文主要的贡献：围绕解决上述问题展开研究，并在已有研究的基础上，基于动态规划思想，提出了一种不确定近邻的协同过滤推荐算法(Uncertain Neighbors' Collaborative Filtering，UNCF)，根据基于用户以及产品的相似性计算，来自适应地选择预测目标的近邻作为推荐群，同时计算推荐群中对预测目标的信任子群，在充分结合推荐群和信任子群的基础上，提出一种不确定近邻因子分析度量来计算预测目标的推荐结果. 通过实验结果表明，该算法相比较传统的基于用户以及基于产品的协同过滤推荐算法，可以有效平衡用户群以及产品群推荐结果所带来不确定的影响，有效缓解用户评分数据极端稀疏情况使用传统性度量方法带来的问题，并显著提高推荐系统的推荐质量. 本文第 2 节给出问题定义以及协同过滤领域研究的基本方法；第 3 节详细介绍本文主要的贡献：不确定近邻的协同过滤算法，并对算法进行形式化描述以及运算时间复杂度分析；第 4 节是针对提出的算法进行实验验证，并对实验结果的比较进行分析；最后是本文的小结.

2 问题定义及基本方法

在推荐系统中，用户对所有产品的评价数据库中包含 s 个用户的集合 $U=\{U_1, U_2, \cdots, U_s\}$ 和 t 个产品的集合 $I=\{I_1, I_2, \cdots, I_t\}$. 用户评分数据集可用一个 $s\times t$ 阶矩阵表示，见表 1.

表 1 用户-产品评分矩阵 $\boldsymbol{R}(s\times t)$

	I_1	…	I_j	…	I_t
U_1	$R_{1,1}$	…	R_{1j}	…	$R_{1,t}$
…					
U_a	$R_{a,1}$	…	$R_{aj}=?$	…	$R_{a,t}$
…					
U_s	$R_{s,1}$	…	R_{sj}	…	$R_{s,t}$

其中，矩阵共有 s 行代表 s 个用户，t 列代表 t 个产品. 假设某一用户 U_a 对产品 I_j (其中 $U_a\in U$，$I_j\in I$) 的打分为 $R_{a,j}$，这个评分体现了用户 U_a 对产品 I_j 的兴趣和偏好.

2.1 相似性度量方法

相似性计算可以是在用户之间的相似性计算，也可以是产品间的计算. 本节中以用户之间的相似

性研究为例，也就是基于用户协同过滤方法.

(1) 标准的余弦相似性. 通过向量间的余弦夹角计算度量

$$Sim(U_a, U_b) = cosine(\boldsymbol{R}_a, \boldsymbol{R}_b) = \frac{\sum_{k=1}^{t} R_{a,k} \times R_{b,k}}{\sqrt{\sum_{k=1}^{t} (R_{a,k})^2} \times \sqrt{\sum_{k=1}^{t} (R_{b,k})^2}} \quad (1)$$

其中，$R_{a,k}$表示用户 U_a对产品 I_k的打分值.

(2) 修正的余弦相似性. 为了修正不同用户存在不同评分尺度的偏差，修正的余弦相似性度量方法通过减去用户对所有产品的平均评分来改善这一缺陷. 选取用户 U_a和 U_b共同打分的产品集，也就是用户 U_a和 U_b打分的交集($I_{U_a} \cap I_{U_b}$)，定义为 I'.

$$Sim(U_a, U_b) = \frac{\sum_{i_k \in I'} (R_{a,k} - \bar{R}_a) \times (R_{b,k} - \bar{R}_b)}{\sqrt{\sum_{k \in I'} (R_{a,k} - \bar{R}_a)^2} \times \sqrt{\sum_{k \in I'} (R_{b,k} - \bar{R}_b)^2}} \quad (2)$$

其中$\bar{R}_a$是用户a评分的平均值. 计算结果$Sim(U_a, U_b)$的值落在[0, 1] 区间中，$Sim(U_a, U_b)$值越大，则表示用户 U_a和 U_b之间的相似性越大.

(3) 相似性度量计算的改良. 基于用户之间的相似性计算依赖于他们共同评分的产品，如果共同评分的产品数较少，则这个相似性度量存在一定的偶然性，为了消除这种偶然性带来的影响. Herlocker 等[17-18] 提出要增加一个关联权重因子来进行计算相似性. 在这个基础上，Ma 等在[19] 中提出了影响性权重的设置，我们定义用户 U_a和 U_b之间共同打分的产品$I' = I_{U_a} \cap I_{U_b}$，通过设定某个阈值 γ，与用户 U_a和 U_b共同打分的产品数目$|I'|$进行比较.

$$Sim'(U_a, U_b) = \frac{\min(|I'|, \gamma)}{\gamma} \times Sim(U_a, U_b) \quad (3)$$

从式(3)中，我们可以看到满足$\frac{\min(|I'|, \gamma)}{\gamma} \leq 1$，因此改良之后的相似度量 $Sim'(U_a, U_b)$的值域仍然落在[0, 1] 区间上. 如果用户 U_a和 U_b间共同打分的产品数目较多，满足$|I'| \geq \gamma$，那么 $Sim'(U_a, U_b) = Sim(U_a, U_b)$，如果共同打分数较少，那么相似度量值也会相应减少.

2.2 *k*NN 协同过滤推荐的结果

通常推荐系统对某一个用户 U_a的主要任务有两个：

(1) 在产品集中，选择某一个用户 a 未曾评分的产品I_j，$I_j \notin I_{U_a}$，预测用户对它的评分 $R_{a,j}$.

(2) 在用户未评分的产品中，预测用户提供评分中最大的 N 个产品($N \geq 1$)，选择推荐给用户

通过计算用户之间的相似性，基于用户的协同过滤算法为未知的 I_j寻找 k 个近邻(与 U_a最相似的 k 个用户)，定义为 $S(U_a)$且$|S(U_a)| = k$，来预测其评分的分数

$$R_{a,j} = \bar{R}_a + \frac{\sum_{U_b \in S(U_a)} Sim'(U_a, U_b) \times (R_{b,j} - \bar{R}_b)}{\sum_{U_b \in S(U_a)} Sim'(U_a, U_b)} \quad (4)$$

上述公式中，其中 $\bar{R}_a$，$\bar{R}_x$ 分别表示用户 U_a，U_x 对其它产品所有打分的均值，$R_{a,x}$表示用户 U_a对产品 I_x的预测评分分数，式(4)中，传统 kNN 的协同过滤推荐算法是采用 k 个最相似的近邻用于预测评分.

3　不确定近邻的协同过滤推荐算法

传统的协同过滤推荐算法本质上是利用了群体内(这里面的群体可能是用户群，也可能是产品群)个体与个体之间的相互作用(寻找对当前对象影响力最大的 k 个邻居)来为当前对象的属性作出预测的过程. 但是，面对复杂的、多变的不确定的预测场景，这种预测结果往往是片面的，只考虑了某一群的影响作用而忽视另一个关联群的影响作用，或者过分关注某一群体的影响作用. 例如要预测某一个用户对某一件商品的喜爱程度，除了可以通过其他对这件商品有过评价的用户与这个用户的相似性(基于用户协同过滤 User-Based Collaborative Filter，UBCF)的方法，还可以考虑这个用户评价过的其他商品与这件商品的相似程度(基于产品协同过滤 Item-Based Collaborative Filter，IBCF)的方法，但是如果这个用户对很多商品均有过评分，但却很少用户对某一件商品进行评分，那么通过基于用户来进行推荐过滤，则推荐准确性会相对较低. 在现实生活中往往这两者均不可或缺，这也促使本文的研究开展，从影响的群体之间寻找一个自适应地可随案例数据变化而变化的紧邻因子，同时为当前对象找到受其影响较大的群体，结合两个群体的近邻对象来共同为当前对象的作出预测.

3.1　动态选择目标的推荐对象群

如果仅仅依赖传统的基于用户或者基于产品，尤其是当某一个数据相当稀疏的时候，往往其结果不太理想. 在传统的协同过滤算法中，最常见的是使用 kNN 算法. 例如：传统的协同过滤依赖于在所有个体中，寻找相似性最大的 k 个邻居，但可能所在同一群体内的个数不足 k 个，那么最大 k 近邻也会产生一些不相似的个体进行协同过滤，从而导致推荐

结果准确性的降低.考虑采用自适应地选择预测目标的方法,可以有效避免人为设置过多的经验值,导致推荐系统无法根据数据的变化,自身进行适应性地调整.

因此我们提出采用动态选择目标的推荐对象方法(Dynamically Selected Neighbor, DSN),在进行近邻对象选择之前,需要界定预测目标的推荐对象应该如何选取,通过定义两个相似度的阈值 μ 和 ν,一个是用于用户间相似度计算的阈值,另一个是用于产品间相似度计算的阈值.我们定义 $S(U_a)$ 为基于用户选取的推荐集,$S(I_j)$ 为基于产品选取的推荐集.推荐集的对象都必须满足相似度大于阈值.我们只考虑选择与目标较接近的作为推荐对象,可以较好地解决传统 kNN 方法的不足.定义

$$S(U_a)=\{U_x \mid Sim'(U_a,U_x)\geqslant\mu,\ a\neq x\} \quad (5)$$

$$S(I_j)=\{I_y \mid Sim'(I_j,I_y)\geqslant\nu,\ j\neq y\} \quad (6)$$

计算这两个群的对象个数,分别计算 $|S(U_a)|=m$ 和 $|S(I_j)|=n$.

3.2 在推荐对象中选择信任子群

针对目标进行推荐对象选择过程中,相似度计算成了主要的衡量指标,但是,在实际的推荐系统中,往往用户的相似度计算,可能仅仅来源于对少数几个产品的打分,甚至可能只有一个共同评分的产品,这样的相似度计算,存在较大的偶然因素,不能代表用户之间的相似性度量.除了要考虑相似度的同时,也需要考虑两者之间共同评价产品的个数.于是,我们根据式(7),来计算共同打分个数大于设定的阈值 ε 的用户推荐群,定义为 $S'(U_a)$.同样,通过式(8),来计算目标项目推荐准确率较高的信任子群,定义为 $S'(I_j)$.

$$S'(U_a)=\{U_x \mid Sim'(U_a,U_x)\geqslant\mu\ \&\ |I_{U_a}\cap I_{U_x}|>\varepsilon,\ a\neq x\} \quad (7)$$

$$S'(I_j)=\{I_y \mid Sim'(I_j,I_y)\geqslant\nu\ \&\ |U_{I_j}\cap U_{I_y}|>\gamma,\ j\neq y\} \quad (8)$$

计算这两个信任子群的对象个数,分别为 $|S'(U_a)|=m'$ 和 $|S'(I_j)|=n'$.

3.3 引入不确定近邻因子的概念

结合两个群体的推荐结果,往往采用基于用户和基于产品的预测平均值作为结果,或者设定某一经验值,但这些均很难产生较好的推荐结果.在本文中提出的不确定近邻因子的协同过滤框架,改变了原来基于用户或者基于产品来寻找 k 个邻居的方法,并对两者的预测结果进行一个可调整因子的加权,产生了新的推荐结果.我们提出的不确定近邻的协同过滤推荐算法 UNCF(Uncertain Neighbors' Collaborative Filtering),是根据不确定的场景,结合用户以及产品的相似性计算,产生一个近邻因子,通过近邻因子去计算基于用户和产品的预测评分并产生推荐.

不失一般性,假设用户 U_a 对产品 I_j 的预测评分,按照式(7)和(8),假设有 $|S'(U_a)|=m'$ 个,分别为 $\{U_{a_1},U_{a_2},\cdots,U_{a_{m'}}\}$,而另外寻找与产品 I_j 近邻的产品集且用户 U_a 对它有评分,这个产品集定义为 $S(I_j)$,同样这个产品集数目 $|S(I_j)|$ 不固定,假设这个 $|S'(I_j)|=n'$,分别为 $\{I_{j_1},I_{j_2},\cdots,I_{j_{n'}}\}$.我们需要预先知道 m' 和 n' 的值,也需要了解预测目标准确率较高的推荐方法.

定义不确定近邻因子 λ 和 $1-\lambda$,分别作为用户群和产品群推荐结果的平衡因子,这两个因子的和为 1,两者结合基于用户以及基于产品的推荐来计算推荐的最终结果,我们通过以下过程来计算 λ 和 $1-\lambda$:

$$\begin{aligned}&\text{If } (m'+n')\geqslant 0,\\&\quad\lambda=\frac{\phi\times m'}{\phi\times m'+n'},\quad 1-\lambda=\frac{n'}{\phi\times m'+n'}\\&\text{Else if } (m+n)\geqslant 0\\&\quad\lambda=\frac{\phi\times m}{\phi\times m+n},\quad 1-\lambda=\frac{n}{\phi\times m+n}\\&\text{Else}\\&\quad\lambda=1-\lambda=0.5\end{aligned} \quad (9)$$

其中 ϕ 作为一个调和参数,以第 1 种情况为例,当 $(m'+n')\geqslant 0$ 存在以下 4 种可能:

(1)若 $m'=0$ 同时 $n'>0$,则与 ϕ 值无关,明显可见 $\lambda=0$,表示完全由基于产品的方法进行推荐.

(2)当满足 $\phi=\frac{n'}{m'}$ 时,则 $\lambda=1-\lambda=0.5$,也就是用户群的推荐和产品群的推荐权重相同,都是 0.5.

(3)当 $\phi\in\left(\frac{n'}{m'},\infty\right)$ 时,如果 ϕ 趋向于无穷大时,则 λ 的值将从 0.5 逐渐变化为 1,而 $1-\lambda$ 从 0.5 逐渐变化为 0,表示在用户群和产品群两者之间,更加趋向于采用用户群推荐,当 $\lambda=1$,则表示完全由用户群进行推荐.

(4)当 $\phi\in\left[0,\frac{n'}{m'}\right)$ 时,如果 ϕ 趋向于 0 时,则 λ 的值将从 0.5 逐渐变化为 0,而 $1-\lambda$ 从 0.5 逐渐变化为 1,表示在用户群和产品群两者之间,更加趋向于采用产品群推荐,当 $\lambda=0$,则表示完全由产品群进行推荐.

因此我们引入 ϕ 这个调和参数,用来协调基于用户以及基于产品的两者之间不同的影响,对两个不同维度的目标对象数进行调和,避免两者影响度不同而造成推荐质量的下降.

3.4　不确定近邻的协同过滤推荐算法

基于上述的讨论，对于目标的在线用户 U_a 以及其他未浏览过的产品 I_j，不确定近邻的协同过滤推荐算法 UNCF，同时结合用户的最近邻集和产品的最近邻集对用户在产品上的评分进行预测，我们给推荐定义以下推荐公式：

$$R_{a,j}=\lambda\times\left(\bar{R}_a+\frac{\sum_{U_x\in S(U_a)} Sim'(U_a,U_x)\times(R_{x,j}-\bar{R}_x)}{\sum_{U_x\in S(U_a)} Sim'(U_a,U_x)}\right)+(1-\lambda)\times\left(\bar{R}_j+\frac{\sum_{I_y\in S(I_j)} Sim'(I_j,I_y)\times(R_{a,y}-\bar{R}_y)}{\sum_{I_y\in S(I_j)} Sim'(I_j,I_y)}\right)\tag{10}$$

其中 $\bar{R}_a$，$\bar{R}_x$ 分别表示用户 U_a，U_x 对其它产品所有打分的均值，$\bar{R}_j$，$\bar{R}_y$ 分别表示产品 I_j，I_y 已知所有用户打分的均值. 公式中根据用户 U_a 和产品 I_j 的不确定近邻群进行推荐，假如用户 U_a 的近邻群为空，即 $\lambda=0$，则完全按照产品 I_j 的近邻群来进行协同过滤. 反之，$\lambda=1$，则是按选择用户近邻的方式来进行预测.

3.5　算法过程及性能分析

算法 **1**.　寻找目标产品的近邻对象算法 DSN.

输入：用户-产品评分矩阵 $\boldsymbol{R}(s\times t)$，目标用户 U_a，产品 I_j，阈值 μ，ν，ε，γ

输出：可以为未知评分进行协同过滤的近邻用户集和近邻产品集

1. 在矩阵 $\boldsymbol{R}(s\times t)$ 中分别计算用户的相似度矩阵和产品的相似度矩阵，并分别保存这两个矩阵 ***Arr_UserSim***$(s\times s)$ 和 ***Arr_ItemSim***$(t\times t)$.

2. 根据输入用户 U_a，在步 1 中保存的用户相似度矩阵中，选取与用户 U_a 相似度大于 μ 的近邻对象，作为用户推荐集 $S(U_a)$. 根据共同打分对象个数大于 ε 的推荐对象 $|I_{U_a}\cap I_{U_x}|>\varepsilon$，得出 $S'(U_a)$ 子群.

3. 在步 1 中保存的产品相似度矩阵中，寻找与产品 I_j 相似度大于 ν 的近邻对象，作为产品推荐集 $S(I_j)$. 并根据共同打分用户个数大于 γ 的推荐对象 $|U_{I_j}\cap U_{I_x}|>\gamma$，得出 $S'(I_j)$ 子群.

在 DSN 算法中，步 1 是其它协同过滤推荐算法中常见的一个步骤，而步 2、3 则是根据近邻阈值的设定，来寻找适合为未知评分进行推荐的协同对象，分别寻找大于阈值 μ，ν 的近邻，而不像传统的协同过滤寻找 k 个. 由于在步 1 中，需要遍历所有的用户及客户，因此算法最坏的时间复杂度为 $O(s^2+t^2)$，在实际的推荐系统中，往往是通过离线的方式来对系统中的用户及产品来计算相似度，并只保留满足大于阈值的近邻对象，那么只需要进行定期的更新即可，可以节省相似度计算的时间. 在只考虑寻找满足大于阈值的近邻对象中，由于 m 和 n 均为常量，那么时间复杂度的计算为 $O(m+n+m'+n')=4\times O(1)$，可以有效避免由于用户数的逐步增多，而导致算法运算数据量的急剧增加.

算法 **2**.　不确定近邻的协同过滤推荐算法 UNCF.

输入：目标用户 U_a，待评分的产品 I_j，调和参数 ϕ

输出：用户 U_a 对产品 I_j 的预计评分 $R_{a,j}$

1. 根据式(5)和(7)，在用户 U_a 的近邻用户集中，分别计算近邻群及信任子群(分别表示为 $S(U_a)$ 和 $S'(U_a)$)，并分别计算它们的数量 m 和 m'.

2. 根据式(6)和(8)，在产品 I_j 的近邻用户集中，分别计算近邻群及信任子群(分别表示为 $S(I_j)$ 和 $S'(I_j)$)，并分别计算它们的数量 n 和 n'.

3. 选择合适的调和参数 ϕ.

4. 通过 m、m'、n 和 n' 作为输入参数代入到式(9)，来计算 λ 和 $1-\lambda$ 的值.

5. 将 λ 和 $1-\lambda$ 的值代入到推荐产生式(10)，计算用户 U_a 对产品 I_j 的预计评分 $R_{a,j}$.

在 UNCF 算法中，通过步 1、步 2 来计算用户和产品的近邻群和信任子集. 我们通过在 DSN 算法中设置的阈值，可以有效控制两个近邻集合的个数 m 和 n 为一个常量，因此它的时间复杂度均为 $O(1)$，步 5 中的推荐产生式同样是进行 m 和 n 次的求和运算，时间复杂度为 $2\times O(1)$，因此，UNCF 算法的时间复杂度为 $4\times O(1)$.

4　实验结果及分析

本节通过实验，来检验我们提出算法的推荐质量，并且回答以下几个问题：(1) 自适应地选择推荐目标的近邻对象与 kNN 方法选择 k 个近邻对象的比较；在不同规模数据集上适应情况的比较；(2) 关于调和参数的使用，对 UNCF 算法推荐结果的影响，是否可以取得一个更好的预测结果；(3) 结合 UNCF 协同过滤算法推荐的结果，与其它协同过滤推荐算法的比较.

4.1　数据集

实验使用的测试数据集是 GroupLens 研究产品组(http://www.grouplens.org)提供的一个著名电影评分数据 MovieLens，它是有 10 万条记录的数据集，记录了 943 个用户对 1682 部电影的评分，

每个用户至少对 20 部电影进行了评分，评分值范围从 1 ~ 5. 5 表示“perfect”（非常好），而“1”表示“poor”（差），用户通过对不同电影上的不同评分表达了自己的兴趣. 用户和产品的评分矩阵密度为

$$\frac{100000}{943\times 1682}=6.3\%.$$

说明此数据集的评分矩阵是相当稀疏的. 我们实验分为 3 组用户进行，首先将数据集随机抽取 100、200 和 300 个用户，并且将实验数据的评分矩阵进一步划分为训练集和测试集，我们引入变量 x，表示训练集占整个数据集的百分比. 例如，$x=0.8$ 表示随机地将数据集中的 80%都将用作训练集，剩下的 20%作为测试集. 在本文的实验中，均采用 $x=0.8$.

4.2 推荐质量的度量标准

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两类. 统计精度度量方法中的平均绝对偏差 *MAE*（Mean Absolute Error）易于理解，可以直观地对推荐质量进行度量，是最常用的一种推荐质量度量方法，因此本文也采用平均绝对偏差 *MAE* 作为度量标准. 平均绝对偏差 *MAE* 通过计算预测的用户评分与实际的用户评分之间的偏差度量预测的准确性，*MAE* 越小，推荐质量越高.

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 $\{r_1, r_2, \cdots, r_N\}$，则平均绝对偏差 *MAE* 定义为

$$MAE=\frac{\sum_{i=1}^{N}|p_i-r_i|}{N}.$$

4.3 动态选择预测目标的推荐对象方法的比较

本实验对本文 3.1 节中提出的动态选择目标的推荐对象 DSN 方法（Dynamically Selected Neighbor）与传统的 *k*NN 方法做了实验进行比较，目的是选择较佳的近邻对象，来进行基于用户的协同过滤，为下一步实验打下基础. 我们以选择的 k 作为横坐标，从 1 个近邻开始，逐步增加到 2, 4, 8, 10, 20, 30, 40, 50, 60 个最近邻，实验结果如图 1 ~ 图 3 所示.

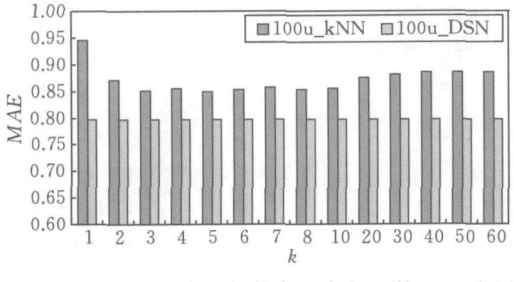

图 1 kNN 方法与动态选择推荐对象的比较（100 个用户）

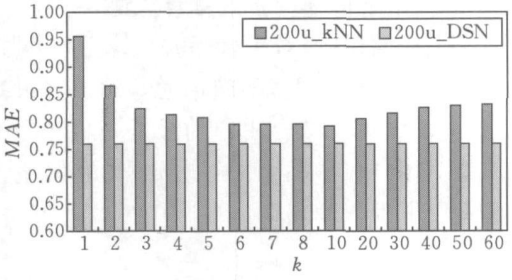

图 2 kNN 方法与动态选择推荐对象的比较（200 个用户）

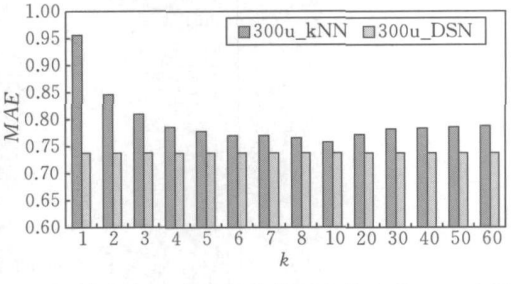

图 3 kNN 方法与动态选择推荐对象的比较（300 个用户）

从图中可以看出，在 100 个用户的实验数据中，当选择 $k=5$ 的时候，100u_kNN 方法可以取得最优的结果，但在同样条件下的 100_DSN 方法，比 100u_kNN 的结果还要更好. 同样，在 200 和 300 个用户的数据集上的实验，都取得类似的结果. 另外，对比 100、200 和 300 个用户之间的实验结果，我们可以看到训练的用户数目越多，越容易找到目标用户合适的推荐对象，通过越多用户的训练结果，可以得到更好的推荐结果. 通过实验，我们可以看出，动态选择的推荐对象的方法同比相应的 *k*NN 方法都要取得更小的 *MAE*，说明动态选择的推荐对象方法可以更加有效地给用户推荐. 下面的实验中，基于用户和基于产品的协同过滤推荐，均采用动态选择推荐对象的方法进行实验.

4.4 近邻因子及调和参数 ϕ 的比率实验

这个实验的目的是为了比较引入近邻因子和调和参数 ϕ 的实验情况，基于用户的协同过滤和基于产品的协同过滤作为参考算法进行比较. 目的是通过比较，考察应该如何选择适当的调和参数 ϕ 进行协同过滤推荐，平衡基于用户与基于产品之间的推荐结果.

实验中，横坐标表示调和参数 ϕ 的取值. 因为通过 UBCF 的预测结果优于 IBCF 的结果，将调和参数 ϕ 从 1 开始进行取值，并逐步增加，通过 100、200 和 300 个用户的实验，可以看到当参数 ϕ 在较小的时候，其结果可能还不如 UBCF 方法，但在从 1 ~20 递增的时候，3 个实验的结果的 *MAE* 值逐步减小，均取得比 UBCF 和 IBCF 更低的值，说明推荐结果

更优. 另外, 如果当调和参数 ϕ 继续逐步递增的时候, 从 20~200 取值时, 也正如我们在 3.2 节中分析的, $\lambda=\frac{\phi\times m'}{\phi\times m'+n}$ 逐步趋向于 1, 结合调和参数 ϕ 的推荐结果逐步趋向于基于用户群协同过滤的 UBCF 方法. 但推荐结果还是优于任意一种 UBCF 或者 IBCF 的推荐方法. 通过实验, 我们得出一个结论, 当调和参数在满足一定条件的时候(以 100、200 和 300 个用户为例, 调和参数 ϕ 在 20 附近), *MAE* 降到最低. 由于调和因子与数据集的关联性非常紧密, 与数据集的构成特征以及稀疏性有关, 甚至可以理解就是数据集的一个属性, 一个可以较好把握数据集基于用户和基于产品对推荐结果产生的影响的一个重要特征. 获取这个取值的方法, 可以通过对数据集进行训练和计算来获取.

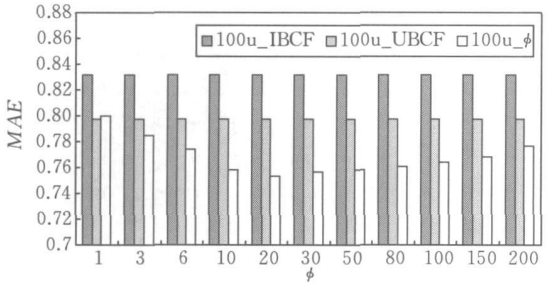

图 4　不同调和参数 ϕ 的结果与 IBCF、UBCF 的比较(100 个用户)

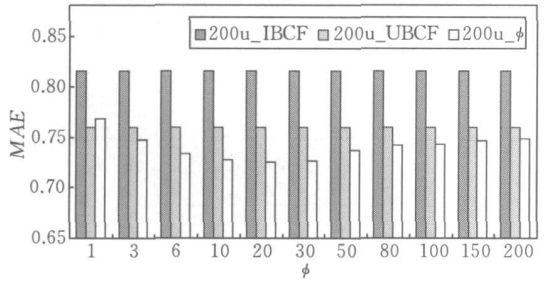

图 5　不同调和参数 ϕ 的结果与 IBCF、UBCF 的比较(200 个用户)

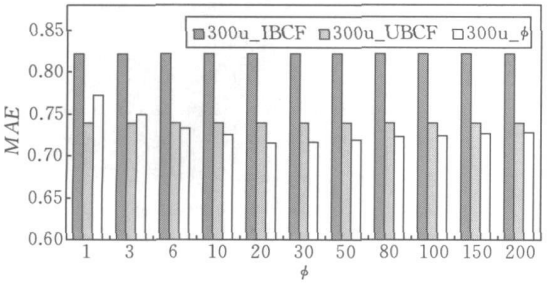

图 6　不同调和参数 ϕ 的结果与 IBCF、UBCF 的比较(300 个用户)

4.5　推荐目标近邻群与信任子群推荐结果的比较

这个实验的目的是为了比较推荐目标近邻群以及信任子群的不同推荐结果, 通过比较来验证采用信任子群的推荐结果. 我们实验数据集分别为 100、200 和 300 个用户, 实验中纵坐标采用 *MAE* 作为度量标准.

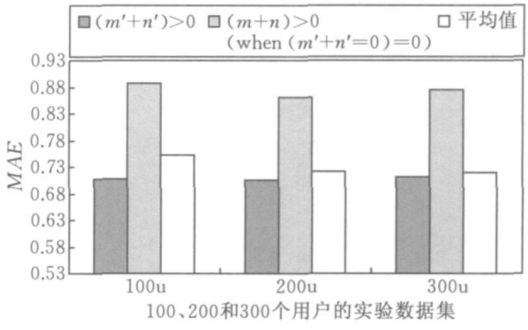

图 7　近邻群与信任子群推荐结果的比较

在实验图 7 中, 左边柱子表示通过信任子群来进行推荐, 即当推荐目标在 $(m'+n')>0$ 的情况. 中间柱子表示满足 $(m+n)>0$, 通过近邻群推荐结果, 右边柱子表示所有测试数据的平均结果. 我们可以看出通过信任子群的推荐结果, *MAE* 值均低于在同样数据集条件下近邻群的结果, 表示通过信任子群可以获得更高的预测准确性.

4.6　**UNCF** 算法与其它协同过滤算法的比较

本实验的目的, 是比较不确定近邻的协同过滤推荐算法 UNCF 与传统的协同过滤算法以及近期业界比较领先的研究进行比较. 选择传统的协同过滤算法 UBCF 和 IBCF, 并与文献[19]中提出的 EMDP(Effective Missing Data Prediction)方法, 采用同样的实验数据进行比较. 我们的实验采用 3.3 节的推荐方法来产生推荐, 实验中的横坐标表示所预测目标产品的近邻数目, 纵坐标采用 *MAE* 作为度量标准.

在实验中, 我们可以看出不确定近邻的协同过滤推荐算法 UNCF, 比较基于产品和基于用户的协同过滤推荐算法以及 EMDP 算法, 本文提出的

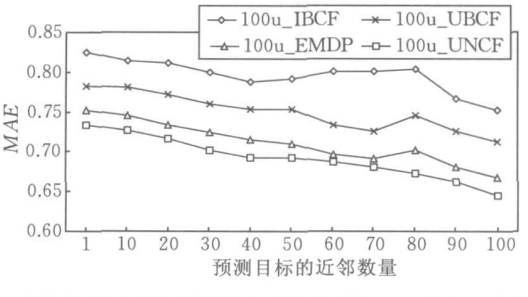

图 8　相关协同过滤推荐算法与 UNCF 算法的比较(100 个用户)

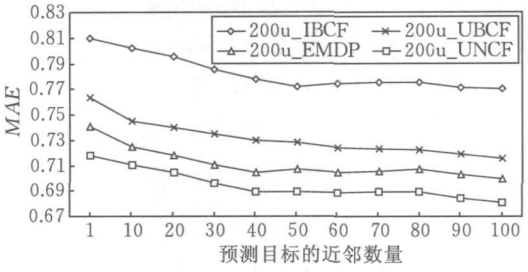

图 9 相关协同过滤算法与 UNCF 算法的比较(200 个用户)

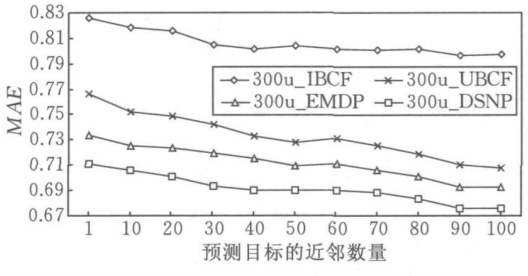

图 10 相关协同过滤算法与 UNCF 算法的比较(300 个用户)

UNCF 算法都可以获得更低的 *MAE* 值,因此推荐的效果更好.当随着所预测产品的近邻数目增加的时候,预测的质量也随之提升,在用户较多的情况下,在 200 和 300 个用户的实验结果提升尤为明显.在不同的实验中,UNCF 算法均比 EMDP 算法得到 3%~7%的提升.

5 结论与未来工作

协同过滤推荐算法,是推荐系统应用中的一个热门研究,应用的领域也非常广泛,针对预测场景的不确定性,本文从自适应地寻找目标的推荐对象出发,优先选择相似度高且推荐准确率较高的信任近邻群,同时深入分析基于用户和基于产品预测的不同影响,提出了一种不确定近邻的协同过滤推荐算法(UNCF),在不确定的场景中,结合用户以及产品的推荐结果,通过不确定近邻因子及调和参数去计算基于用户和产品的预测评分并产生推荐.实验结果表明,不确定近邻的协同过滤推荐算法对系统的正确推荐起了积极作用.另外 UNCF 算法在推荐上,由于结合基于用户和基于产品的推荐结果,使得推荐结果可以在更广泛的数据源中获得,而不仅仅在于用户群或者产品群,因此可以较好地解决数据稀疏性的问题,并提高推荐的质量.

在本文开头,提到推荐系统研究领域,目前有许多不同的研究思路,如何结合各种不同的研究方法各自的优势,例如双聚类[20] 和信任传递[14] 等,是当今一个很重要的研究方向.另外,推荐系统中的数据往往是带有一定的时间属性,而且用户和产品本身也是具有一些可以获取的属性,通过结合时间性因素以及产品属性,建立带有这些参数的相关模型,例如本体模型[7] 等,对目标产品进行更准确的预测,来提高推荐系统的准确性,也是一个有意义的方向.

参 考 文 献

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

[2] Resnick P, Iacovou N, Suchak M, Bergstorm P, Riedl J. GroupLens: An open architecture for collaborative filtering of netnews//Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work. Chapel Hill, North Carolina, United States, 1994: 175-186

[3] Shardanand U, Maes P. Social information filtering: Algorithms for automating "word of mouth"//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Denver, Colorado, United States, 1995: 210-217

[4] Hill M, Stead L, Furnas G. Recommending and evaluating choices in a virtual community of use//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Denver, Colorado, United States, 1995: 194-201

[5] Sarwar B M, Karypis G, Konstan J A, Riedl J. Application of dimensionality reduction in recommender system — A case study//Proceedings of the ACM WebKDD Web Mining for E-Commerce Workshop. Boston, MA, United States, 2000: 82-90

[6] Massa P, Avesani P. Trust-aware collaborative filtering for recommender systems. Lecture Notes in Computer Science, 2004, 3290: 492-508

[7] Vincent S-Z, Boi Faltings. Using hierarchical clustering for learning the ontologies used in recommendation systems//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California, United States, 2007: 599-608

[8] Park S-T, Pennock D M. Applying collaborative filtering techniques to movie search for better ranking and browsing//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California, United States, 2007: 550-559

[9] Tomoharu I, Kazumi S, Takeshi Y. Modeling user behavior in recommender systems based on maximum entropy//Proceedings of the 16th International Conference on World Wide Web. Banff, Alberta, Canada, 2007: 1281-1282

[10] Yang J-M, Li K-F. An inference-based collaborative filtering approach//Proceedings of the 3rd IEEE International Symposium on Dependable, Autonomic and Secure Computing (DASC). Columbia, MD, United States, 2007: 84-94

[11] Chen M-C, Chen L-S, Hsu F-H, Hsu Y, Chou H-Y. HPRS: A profitability based recommender system//Proceed-

ings of the IEEE International Conference on Industrial Engineering and Engineering Management. Singapore, 2007: 219-223

[12] Chen Jian, Yin Qian. A collaborative filtering recommendation algorithm based on influence sets. Journal of Software, 2007, 18(7): 1685-1694(in Chinese)
(陈健, 印鉴. 基于影响集的协作过滤推荐算法. 软件学报, 2007, 18(7): 1685-1694)

[13] Liu X, Datta A, Rzadca K, Lim E-P. StereoTrust: A group based personalized trust model//Proceedings of the 18th ACM conference on Information and Knowledge Management. Hong Kong, China, 2009: 7-16

[14] Jamali M, Ester M, TrustWalker: A random walk model for combining trust-based and item-based recommendation//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 397-406

[15] Sarwar B, Karypis G, Konstan J, Riedl J. Item-based collaborative filtering recommendation algorithms//Proceedings of the 10th International Conference on World Wide Web. Hong Kong, China, 2001: 285-295

[16] Lee H-C, Lee S-J, Chung Y-J. A study on the improved collaborative filtering algorithm for recommender system//Proceedings of the 5th ACIS International Conference on Software Engineering Research, Management & Applications. Toowoomba, Australia, 2007: 297-304

[17] Herlocker J, Konstan J A, Riedl J. An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. Information Retrieval, 2002, 5(4): 287-310

[18] McLaughlin M-R, Herlocker J L. A collaborative filtering algorithm and evaluation metric that accurately model the user experience//Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, United Kingdom, 2004: 329-336

[19] Ma H, King I, Lyu M R. Effective missing data prediction for collaborative filtering//Proceedings of the 30th Annual International ACM SIGIR Conference. Amsterdam, The Netherlands, 2007: 39-46

[20] Deodhar M, Ghosh J. A framework for simultaneous co-clustering and learning from complex data//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California, United States, 2007: 250-259

HUANG Chuang-Guang, born in 1978, Ph. D. candidate. His major research interests include information filtering, data mining and their applications in social behavior analysis and recommender systems.

YIN Jian, born in 1968, professor, Ph. D. supervisor. His research interests include machine learning and data mining.

WANG Jing, born in 1980, Ph. D. candidate. Her major research interests focus on personalized recommendation.

LIU Yu-Bao, born in 1975, Ph. D., associate professor. His research interests include database system, data warehouse and data mining.

WANG Jia-Hai, born in 1977, Ph. D., associate professor. His main research interests include optimization theory, algorithms, artificial intelligence and data mining.

Background

Collaborative filtering (CF) algorithm is widely used in recommender system. But in the real circumstance, especially the scene of recommendation is uncertain, the traditional CF exists some drawbacks, such as using *k*NN for neighbor objects selection and disregarding some useful information between user-based and item-based in making a prediction. In this paper, we proposed a novel algorithm Uncertain Neighbors' Collaborative Filtering recommendation algorithm (UNCF) to overcome those drawbacks. This algorithm is based on the fact that two perspectives of user-based and item-based may differently effect the result of prediction. With the optimum coordinator parameter, we made the leveraged utilization of the information to capture accurately. Experimental results show that UNCF outperforms more accuracy. This work is supported by the National Natural Science Foundation of China (60773198, 60703111), Natural Science Foundation of Guangdong Province (7300272, 8151027501000021), Research Foundation of National Science and Technology Plan Project (2008ZX10005-013), Research Foundation of Science and Technology Plan Project in Guangdong Province (2008B050100040, 2009A080207005, 2009B090300450), Program for New Century Excellent Talents in University of China (NCET-06-0727)