doi:10.11959/j.issn.1000-436x.2016024

基于用户模糊相似度的协同过滤算法

吴毅涛，张兴明，王兴茂，李晗
(国家数字交换系统工程技术研究中心，河南 郑州 450002)

摘 要：针对离散评分不能合理表达用户观点和传统协同过滤算法存在稀疏性等问题，借鉴年龄模糊模型，提出了梯形模糊评分模型。该模型将离散评分模糊化为梯形模糊数，考虑了评分模糊性和信息量，通过梯形模糊数来计算用户相似度，据此设计了协同过滤算法，并证明了该算法是传统协同过滤算法在模糊域的扩展。实验表明，该算法在数据稀疏且用户数远多于项目数时性能突出，并且算法运行时间远小于传统协同过滤算法。
关键词：协同过滤；梯形模糊评分模型；模糊距离；模糊相似度
中图分类号：TP393 **文献标识码**：A

User fuzzy similarity-based collaborative filtering recommendation algorithm

WU Yi-tao, ZHANG Xing-ming, WANG Xing-mao, LI Han
(National Digital Switching System Engineering and Technological R&D Center, Zhengzhou 450002, China)

Abstract: In order to reflect the actual case of human decisions and solve the data sparseness problem of traditional collaborative filtering recommendation algorithm, a trapezoid fuzzy model based on age fuzzy model was proposed. In this model, crisp point was fuzzified into trapezoid fuzzy number and the fuzziness and information of users' grade was taken into account when calculating user's similarity by trapezoid fuzzy number. Based on this model, the user fuzzy similarity-based collaborative filtering recommendation algorithm was designed. The algorithm was proved to be an extension of traditional collaborative filtering algorithm in fuzzy fields. The experimental results show that, the proposed algorithm performs better when implemented in the sparse dataset with more user than item, and its running time is much less than traditional collaborative filtering algorithm.
Key words: collaborative filtering, trapezoid fuzzy model, fuzzy distance, fuzzy similarity

1 引言

电子商务的快速发展，使用户难以处理种类繁多的信息。而推荐系统已经被证明能帮助用户过滤无用信息，做出合理选择[1~3]。推荐系统根据使用内容不同，可分为基于内容推荐系统和协同过滤推荐系统[4]。

基于内容推荐系统主要利用用户的统计信息，如年龄、收入等，根据统计信息的关系进行推荐。协同过滤推荐系统根据评分信息寻找相似用户，寻找相似性大的前 k 个邻居，根据邻居的评分进行预测。算法的关键是选取合理的相似性计算方法。传统算法大多采用余弦、Pearson 等方法来计算用户相似度。协同过滤推荐系统易于处理数据并易于实现，是最成功和流行的推荐系统。

但目前协同过滤推荐系统大都使用离散评分[5]，用户在 5 评分等级集合{1,2,3,4,5}中选择对项目的评分。但用户对项目的喜好程度是非常模糊的，没有特定的标准，离散评分不能合理表达用户的观点，例如离散评分不能表达介于评分 4 和评分 5 之

收稿日期：2014-12-08；修回日期：2015-06-15
基金项目：国家重点基础研究发展计划（“973”计划）基金资助项目（No.2012CB315901）；国家高技术研究发展计划（“863”计划）基金资助项目（No.2011AA01AA103）
Foundation Items: The National Basic Research Program of China(973 Program)(No.2012CB315901), The National High Technology Research and Development Program of China (863 Program)(No.2011AA01AA103)

间的喜好程度[6]。并且协同过滤系统没有考虑评分信息量的问题，例如用户评分为 1 携带的信息量比用户评分为 3 携带的信息量要多。当评分矩阵稀疏时，协同过滤推荐系统的性能非常差。

为了合理表述用户间的关系，Yager[7]引入模糊理论，用模糊子集表示统计信息间的关系；Shamri等[8]提出了统计信息模糊模型，建立统计信息同模糊语言的映射关系，通过模糊语言来计算其相似度；Le[9]利用统计信息模糊模型，计算其统计信息相似度，再利用 Pearson 算法计算评分相似度，加权两部分得到最终相似度。在数据稀疏时，引入模糊理论的推荐系统精确度较高[10]，但忽略了评分的模糊性，只能片面表述用户观点，并且统计信息难于获得和处理，引入模糊理论的推荐系统适用范围很小。

上述研究表明，协同过滤推荐系统不能合理表达用户的观点，没有考虑评分信息量，且存在稀疏性等问题，引入模糊理论的推荐系统只能片面表述用户的观点，且系统的适用范围很小。

针对以上问题，本文借鉴统计信息模糊模型，提出了一种梯形模糊相似度模型，用模糊子集来表示用户评分间的关系，建立离散评分值和梯形模糊评分值的映射关系，将用户评分模糊化，并且考虑了评分的信息量，能合理表达用户观点。用梯形模糊评分进行用户相似度计算，设计了基于用户模糊相似度的协同过滤推荐（Fuzzy-UBCF）算法。实验结果表明，在数据稀疏且用户数远多于项目数时，准确度高，并且算法运行时间远小于传统协同过滤算法。

2　理论基础

2.1　模糊子集

引入模糊理论的推荐系统用模糊子集来表示统计信息间的关系，模糊子集是经典子集的推广，它是具有不分明边界的集合。Zadeh[11]对模糊子集的定义是：给定论域 U 上的一个模糊子集 A，就是给定论域 U 到区间[0,1]的一个映射，如式(1)所示。

$$\mu_A : U \to [0,1]$$
$$u \mid \to \mu_A(u) \in [0,1] \tag{1}$$

映射 μ_A 叫做模糊子集的隶属函数，$\forall \mu \in U$ 对应着一个确定值，该值 $\mu_A(u)$ 叫做 $\mu \in U$ 对 A 的隶属度。

Chen[12]定义梯形模糊数为 $\tilde{A} = (a,b,c,d;W)$，a、b、c、d 分别表示梯形的 4 个顶点，并且是实数；W 表示 $x \in X$ 对模糊数 $\tilde{A}$ 的最大隶属度，$0<W\leq 1$。梯形模糊数可以描述用户对项目的喜好程度，梯形模糊数 $\tilde{A}$ 如图 1 所示。

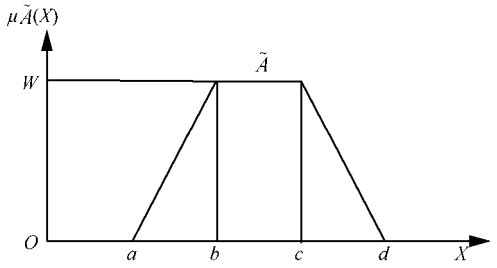

图 1　梯形模糊数 $\tilde{A}$

2.2　年龄模糊模型

Shamri[8]提出的年龄模糊模型描述了年龄同模糊语言的映射关系，如图 2 所示。

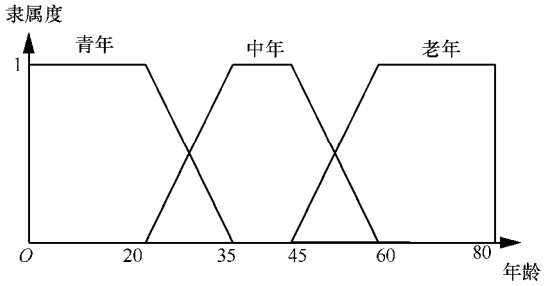

图 2　年龄模糊模型

将年龄作为给定论域，以梯形隶属函数映射到模糊语言集中。相应的模糊语言集为{青年，中年，老年}，这种模型有以下优点。

1) 用模糊语言表示没有特定标准的统计信息，一个年龄可能映射到 2 个不同的模糊语言集，能合理地表述统计信息间的关系。

2) 在实际中，统计信息和模糊语言的隶属函数近似于正态分布，用梯形函数近似隶属函数比较合理。

3) 模型左右对称，用模糊梯形数表示模糊语言，计算简单。

年龄模糊模型是统计信息模糊模型的一种，但统计信息模糊模型只考虑了用户的部分信息，精确度较低，适用范围小。

3　基于用户模糊相似度的协同过滤推荐算法

本文用模糊子集来表示用户评分间的关系，建立了梯形模糊评分模型，将模糊理论引入协同过滤推荐系统中。

3.1 梯形模糊评分模型

本文在年龄模糊模型优点的基础上，对于一个 5 评分等级的集合，提出一种梯形模糊评分模型，如图 3 所示。

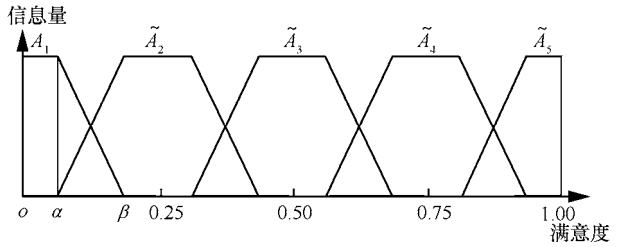

图 3 梯形模糊评分模型

梯形模糊评分模型将满意度作为给定论域，用等腰梯形隶属函数将满意度映射到离散评分集中，满意度表示用户对项目的满意程度，满意度值越大，用户对项目越满意。梯形隶属函数的隶属度用信息量 W_i 来表示，信息量表示模糊数携带的信息的多少，信息量越大，模糊数携带的信息越多，信息量同模糊数出现的概率成反比，也就是同等腰梯形隶属函数的面积成反比，梯形面积越大，信息量越小。

本文定义了 2 个参数：α 和 β。α 和 β 可以表示用户对一个项目的喜好程度，线段 $o\alpha$ 表示满意度区间对离散评分为 1 的确定度，在此范围内，离散评分和满意度是一一映射；线段 $\alpha\beta$ 表示满意度区间对离散评分为 1 的模糊度，在此范围内，离散评分和满意度不是一一映射。根据模型关系可得：$\beta \le 0.25$, $\alpha \le \beta$。当 β 不变，α 变大时，模糊评分数的确定度增加，适用于用户和项目关系较紧密的数据集，也就是稀疏度低的数据集；当 β 不变，α 变小时，适用于稀疏度高的数据集。同理，当 α 不变，β 变大时，适用于稀疏度高的数据集；当 α 不变，β 变小时，适用于稀疏度低的数据集。

在模糊理论中，梯形模糊数 $\tilde{A}_i$ 用($a_{i,1}$, $a_{i,2}$, $a_{i,3}$, $a_{i,4}$;W_i)表示，$a_{i,1}$, $a_{i,2}$, $a_{i,3}$, $a_{i,4}$ 分别表示梯形的 4 个顶点，并且是确定数，W_i 代表信息量，计算式如式(2)所示。

$$W_i = -\mathrm{lb}\left(\frac{a_{i,4}+a_{i,3}-a_{i,2}-a_{i,1}}{2}\right) \tag{2}$$

由模型左右对称，可知 $W_1=W_5$，$W_2=W_4$。为了使 W_i 的值域为[0,1]，对 W_i 进行归一化处理，如式(3)所示。

$$W_i = \frac{W_i}{\max(W_1,W_2,W_3)} \tag{3}$$

其中，$\max(W_1,W_2,W_3)$表示 W_1,W_2,W_3 中的最大值。

根据模型对称关系，得到 5 评分等级集合的梯形模糊评分值如式(4)所示。

$$\begin{aligned}
\tilde{A}_1 &= (0,0,\alpha,\beta;W_1)\\
\tilde{A}_2 &= (\alpha,\beta,0.5-\beta,0.5-\alpha;W_2)\\
\tilde{A}_3 &= (0.5-\beta,0.5-\alpha,0.5+\alpha,0.5+\beta;W_3)\\
\tilde{A}_4 &= (0.5+\alpha,0.5+\beta,1-\beta,1-\alpha;W_4)\\
\tilde{A}_5 &= (1-\beta,1-\alpha,1,1;W_5)
\end{aligned} \tag{4}$$

通过本模型，用梯形模糊数 $\tilde{A}_i$ 来代替离散评分 i 进行后续计算，将评分模糊化后会比离散评分携带更多的信息，适用于处理信息不足的情况，也就是稀疏度高的数据集。

3.2 模糊相似度计算

本文首次将 Chen 等[13]提出的梯形模糊数相似度计算方法引入推荐系统中，以此设计了用户模糊相似度计算方法，并证明模糊相似度是余弦相似度在模糊域的扩展。梯形模糊数相似度计算方法考虑了梯形模糊数的常规距离和重心距离，如式(5)所示。

$$S(\tilde{A}_i,\tilde{A}_j) = \left[1-\frac{\sum_{k=1}^{4}|a_{i,k}-a_{j,k}|}{4}\right]\cdot (1-|x^*_{\tilde{A}_i}-x^*_{\tilde{A}_j}|)^{B(S_{\tilde{A}_i},S_{\tilde{A}_j})}\frac{\min(y^*_{\tilde{A}_i},y^*_{\tilde{A}_j})}{\max(y^*_{\tilde{A}_i},y^*_{\tilde{A}_j})} \tag{5}$$

其中，$a_{i,k}$ 为梯形 $\tilde{A}_i$ 的第 k 个顶点，$(x^*_{\tilde{A}_i}, y^*_{\tilde{A}_i})$为梯形 $\tilde{A}_i$ 的重心，如式(6)所示。

$$y^*_{\tilde{A}_i} = \begin{cases} W_i\left(\dfrac{a_{i,3}-a_{i,2}}{a_{i,4}-a_{i,1}}+2\right)\Big/6, & a_{i,1}\neq a_{i,4},\ 0<W_i<1\\ W_i/2, & a_{i,1}=a_{i,4},\ 0<W_i<1 \end{cases}$$

$$x^*_{\tilde{A}_i} = \frac{y^*_{\tilde{A}_i}(a_{i,3}+a_{i,2})+(a_{i,4}+a_{i,1})(W_i-y^*_{\tilde{A}_i})}{2W_i} \tag{6}$$

W_i 为梯形 $\tilde{A}_i$ 的信息量，$B(S_{\tilde{A}_i},S_{\tilde{A}_j})$用来决定是否运用重心距离，定义如式(7)所示。

$$B(S_{\tilde{A}_i},S_{\tilde{A}_j})=\begin{cases}1, & a_{i,4}-a_{i,1}+a_{j,4}-a_{j,1}>0\\ 0, & a_{i,4}-a_{i,1}+a_{j,4}-a_{j,1}=0\end{cases} \tag{7}$$

模糊相似度由常规模糊距离 (i.e, $\left[1-\frac{\sum_{k=1}^{4}|a_{i,k}-a_{j,k}|}{4}\right]$) 和重心距离 (i.e., $(1-|x^*_{\tilde{A}_i}-x^*_{\tilde{A}_j}|)^{B(S_{\tilde{A}_i},S_{\tilde{A}_j})}$) 组成。常规距离体现的是模糊数整体上的差异，而重心距离体现的是模糊数信息量的差别，信息量差别越大，重心距离越大。当 2 个模糊数为确定数时，则 $a_{i,4}-a_{i,1}+a_{j,4}-a_{j,1}=0$，就不再考虑重心距离。

根据式(5)可以得出2个用户关于一个项目的相似度，加权所有项目的相似度就可以得到用户模糊相似度，如式(8)所示。

$$sim(u,v)=\frac{\sum_{i\in U}S(R_{u,i},R_{v,i})}{n} \tag{8}$$

其中，U 代表用户 u 和用户 v 的共同评过分的项目的集合，$R_{u,i}$ 表示用户 u 对项目 i 的评分，n 为用户 u 评分的项目数，上式还考虑了用户共同评分项目占总评分项目的比例，更能体现出用户 u 和 v 间的差异。

可以证明模糊相似度是余弦相似度在模糊域的扩展。

在模糊评分模型中，评分值 $\tilde{A}_i$ 用($a_{i,1}$,$a_{i,2}$, $a_{i,3}$, $a_{i,4}$;W_i)表示，当 $a_{i,1}=a_{i,2}=a_{i,3}=a_{i,4}$，且 $W_1=W_2=W_3=W_4=W_5$ 时，模型就退化为离散评分模式。则 $B(S_{\tilde{A}_i},S_{\tilde{A}_j})=0$，不再考虑重心距离。模糊相似度计算式退化如式(9)所示。

$$sim(u,v)=\frac{n(U)}{n}-\frac{\sum_{i\in U}\left|R_{u,i}-R_{v,i}\right|}{n} \tag{9}$$

其中，U 为用户 u 和 v 共同评分的项目集合，$n(U)$ 为集合 U 中的项目数，n 为用户 u 评分的项目数。$\sum_{i\in U}\left|R_{u,i}-R_{v,i}\right|$ 即曼哈顿距离 d_{M}，是指 2 个点在标准坐标系上的绝对轴距离总和，也就是欧氏距离在坐标轴上投影的距离总和，欧氏距离 d_{E} 表示 2 个点的实际距离，如式(10)所示。它们的关系如图 4 所示。

$$d_{\mathrm{E}}=\sqrt{\sum_{i\in U}\left|R_{u,i}-R_{v,i}\right|^2} \tag{10}$$

其中，$\vec{\boldsymbol{u}}$ 和 $\vec{\boldsymbol{v}}$ 分别表示用户 u 和用户 v 的共同评分项目的矩阵向量，在 $\vec{\boldsymbol{u}}$ 和 $\vec{\boldsymbol{v}}$ 的长度固定的情况下，向量间的夹角为 θ, 欧氏距离 d_{E} 为 2 个向量终点的距离，曼哈顿距离 $d_{\mathrm{M}}=l_1+l_2$，结果如式(11)所示。

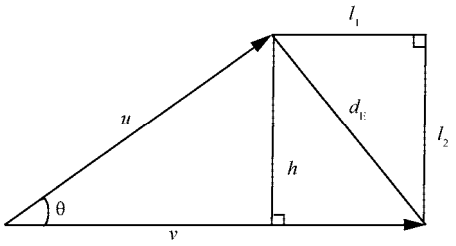

图 4　用户 u 和用户 v 相似度及距离关系

$$\begin{aligned}d_{\mathrm{M}}&=l_1+l_2\\&=(\|\vec{\boldsymbol{u}}\|\sin\theta)+(\|\vec{\boldsymbol{v}}\|-\|\vec{\boldsymbol{u}}\|\cos\theta)\\&=\|\vec{\boldsymbol{v}}\|+\|\vec{\boldsymbol{u}}\|(\sin\theta-\cos\theta)\\&=\|\vec{\boldsymbol{v}}\|-\sqrt{2}\cos(\frac{\pi}{4}+\theta)\|\vec{\boldsymbol{u}}\|\end{aligned} \tag{11}$$

而 $\cos\theta=\frac{\vec{\boldsymbol{u}}\vec{\boldsymbol{v}}}{\|\vec{\boldsymbol{u}}\|\|\vec{\boldsymbol{v}}\|}$，用户 u 和 v 的评分值都是正数，故 $0^\circ\le\theta\le90^\circ$，且 $\frac{n(U)}{n}$ 为常数，故 $sim(u,v)=\frac{n(U)}{n}-\frac{d_{\mathrm{M}}}{n}\propto\cos\theta$。

当 $\theta=0^\circ$ 时，$\cos\theta=1$，$sim(u,v)=\frac{n(U)}{n}-\frac{\|\vec{\boldsymbol{v}}\|-\|\vec{\boldsymbol{u}}\|}{n}=\frac{n(U)}{n}$，余弦相似度和模糊相似度都取到最大值；当 $\theta=90^\circ$ 时，$\cos\theta=0$，$sim(u,v)=\frac{n(U)}{n}-\frac{\|\vec{\boldsymbol{v}}\|+\|\vec{\boldsymbol{u}}\|}{n}$，余弦相似度和模糊相似度都取到最小值。故模糊相似度是余弦相似度在模糊域的扩展。

3.3　算法的流程

综合模糊相似度的计算，本节给出 Fuzzy-UBCF 对目标用户 u 未知项目的预测评分集的流程。

算法　Fuzzy-UBCF ($\boldsymbol{R}$,k)

输入：用户对项目的评分矩阵 $\boldsymbol{R}$，用户邻居数 k。

Begin：

1) 用户相似度计算。

①根据式(5)计算目标用户 u 和其他用户 v 关于一个共同评分的项目的相似度 $S(R_{u,i},R_{v,i})$。

②根据式(8)得出目标用户 u 和其他用户 v 的模糊相似度 $sim(u,v)$。

2) 产生推荐集

①挑选出相似度最高的 k 个用户，作为邻居集 K。

②对近邻采用平均加权方法进行评分预测[14]，如式(12)所示。

$$P_{u,i}=\bar{R}_u p+\frac{\sum_{v\in K} sim(u,v)(R_{v,i}-\bar{R}_v p)}{\sum sim(u,v)} \tag{12}$$

其中，$P_{u,i}$ 表示用户 u 对项目 i 的预测评分，$sim(u,v)$ 表示用户 u 和用户 v 之间的相似度，$\bar{R}_v$ 表示用户 v 项目评分的均值，通过参数 p 来减弱平均值对预测评分的影响，增加预测的模糊性，本文令 p=0.8。

输出：目标用户 u 未知项目的预测评分集。

End

4 算法分析

4.1 基于用户模糊相似度算法正确性分析

传统相似性同用户向量的夹角是正相关关系，同用户向量的方向直接相关。模糊相似度同用户间的距离是负相关关系，同个体特征的维度，即同用户向量的长度直接相关。当用户向量的方向保持不变，长度增加时，传统相似度的值不变，而向量间的距离会变大，造成模糊相似度变小。所以项目数越多，模糊相似度的精确度越低。

传统的相似性计算，是从整体上计算用户差异，对单个用户对项目的评分值不敏感，项目越多，评分矩阵越密集，越容易分析其差异，故传统相似性适用于项目数多，且评分稀疏度低的数据集。而模糊相似性，分析的是用户评分的绝对差异，对单个用户项目的评分值敏感，用户越多，项目数越小，即用户项目比越大，越容易分析其差异，故模糊相似性适用于用户项目比大的评分矩阵。

虽然模糊相似度只考虑用户间的共同评分项目，一般来说共同评分项目很少，但本文将评分模糊化后，可以从项目较少的集合中获取更多信息，在评分矩阵很稀疏时也有很好的效果。故模糊相似度算法适用评分稀疏且用户项目比大的数据集。

4.2 算法运行时间分析

协同过滤算法的时间开销主要在相似度计算中，本文只考虑相似性计算的运行时间，若用户的数量为 n，项目的数量为 m，余弦相似度算法的运行时间分析如表 1 所示。

表 1 余弦相似度算法的运行时间分析

程序段	开销	次数
for(int i=0;i<n; i++)	c_1	n
{		
a = *rateMatrix.getRow*(i);	c_2	n
for(int i=0;i<n; i++)	c_3	n^2
{		
b = *rateMatrix.getRow*(i);	c_4	n^2
sim(a,b)=(a·b)/(||a||*||b||);	c_5	n^2
}		
}		

由于 c_5 远远大于 c_1 、c_2、 c_3 和 c_4，故 $T(n)=c_5n^2=O(n^2)$ ，算法时间复杂度为 $O(n^2)$。

因为用户模糊相似性需要先计算 $S(R_{u,i}, R_{v,i})$，即用户对单个项目的相似度，但只有 5 种评分，$S(R_{u,i}, R_{v,i})$只有 25 种可能值，可以提前计算出，这部分计算开销可以忽略不计。模糊相似度算法的运行时间分析如表 2 所示。

表 2 模糊相似度算法的运行时间分析

程序段	开销	次数
for(int i=0;i<n; i++)	c_1	n
{		
a = *rateMatrix.getRow*(i);	c_2	n
for(int i=0;i<n; i++)	c_3	n^2
{		
b = *rateMatrix.getRow*(i);	c_4	n^2
sim(a,b)=*S*(a,b);	c_6	n^2
}		
}		

由于 c_6 远远大于 c_1、c_2、c_3 和 c_4，故 $T(n)=c_6n^2=O(n^2)$，算法时间复杂度为 $O(n^2)$。

虽然模糊相似度和传统相似度的算法复杂度都为 $O(n^2)$，但开销 c_5 需要进行 $3m$+1 次乘法，2(m−1)加法，2 次开方和 1 次除法，开销 c_6 只需进行 m−1 次加法和一次除法，故 $c_6 \ll c_5$。Pearson 相似度算法的运行时间远高于余弦相似度算法的运行时间，所以模糊相似性算法的运行时间远小于传统相似性算法。

5 实验与分析

5.1 数据集及实验环境

本文使用的是 Netflix 电影评分数据集(评分值为 1~5 的整数)，用于 Netflix Prize 比赛中。Netflix 有 2 个不同大小的数据集，具体参数如表 3 所示。

表 3 数据集的具体参数

名称	用户数	项目数	用户项目比	评分总数	评分密度	稀疏度
Netflix_3m1k	4 427	1 000	4.427	56 136	1.27%	98.73%
Netflix_5m3k	8 662	3 000	2.954	293 299	1.13%	98.87%

本文实验环境为：win 7 操作系统, 8 GB 内存, Inter(R) Core(TM) i7-2600 CPU 3.40 GHz, 实验程序使用 java 1.5 语言开发。

5.2 评价指标

本文采用平均绝对误差 *MAE* 作为算法性能的评价指标，如式(13)所示[15]。

$$MAE=\frac{1}{N}\sum_{i=1}^{N}\left|p_i-r_i\right| \tag{13}$$

其中，p_i 为算法的预测评分，r_i 为测试数据中的实际评分，N 为测试集中项目数目。*MAE* 越小，推荐精度越高。

5.3 比较算法及参数确定

本文采用以下 2 种算法作为对比算法。

余弦相似性的协同过滤算法(Cosine-CF)是协同过滤原始的经典算法。Pearson 相似性的协同过滤算法(Pearson-CF)，在 Cosine-CF 的基础上进行改进，是目前应用广泛的算法，并且是基于用户的共同评分项目进行计算相似度，和本文提出的算法相同。

为了选取合理的 α 和 β，本实验通过 netflix_3m1k_split.txt 文件，将 Netflix_3m1k 数据集中的 95%作为训练集, 5%作为测试集。将各组合的 *MAE* 减去基准 *MAE*（0.735 0），在把差值扩大 500 倍，对比在不同 α 和 β 组合下 *MAE* 大小，实验结果如图 5 所示。可得，当 $0.36\leq\alpha+\beta\leq0.38$ 时，*MAE* 值比较小，经过多次实验，本文选取 α=0.13, β=0.23 进行后续实验。

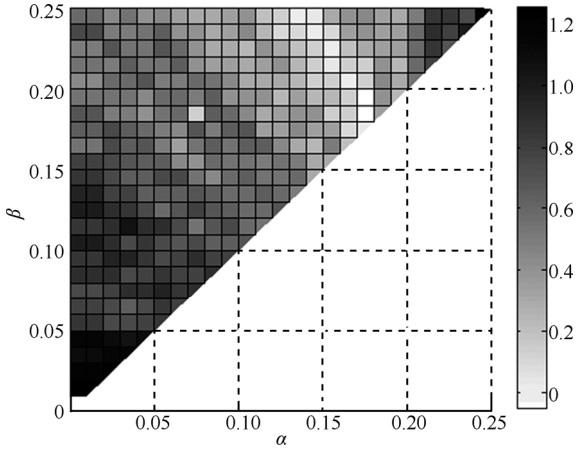

图 5　不同 α 和 β 组合算法精确度比较

5.4 实验结果与分析

本实验中随机将 80%的数据集作为训练集，20%的数据集作为测试集。为了减少随机分割数据集带来的误差，所有实验都进行 10 次，取平均值作为最终结果。

实验 1　近邻数对算法精度的影响

当近邻数 k 从 5~50 变化时，比较 3 种算法的 *MAE* 大小。实验结果如图 6 和图 7 所示。

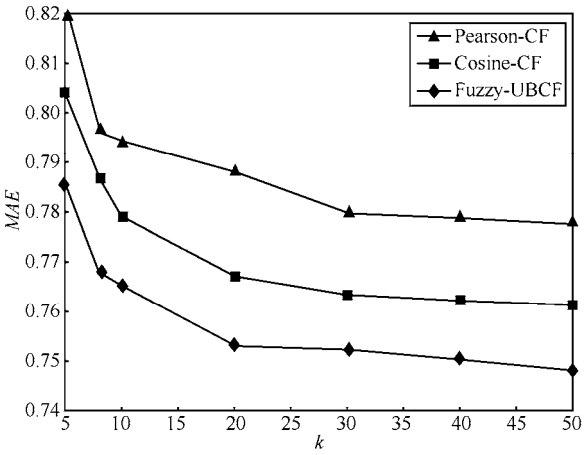

图 6　Netflix_3m1k 中 3 种算法精确度比较

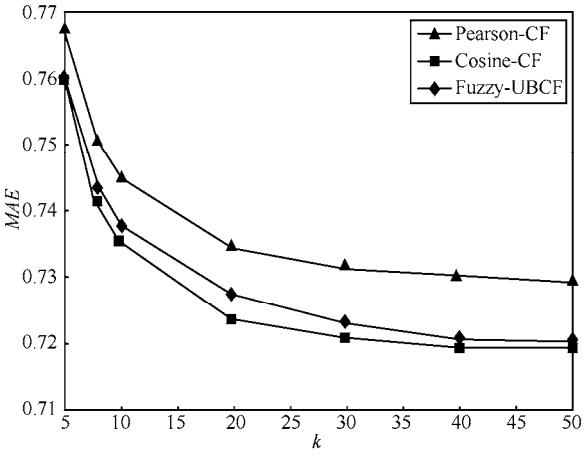

图 7　Netflix_5m3k 中 3 种算法精确度比较

从实验结果可以得出以下结论。

1) 随着 k 的增大，3 种算法的精度都会提高，但算法复杂度也会增加。当 k>20 时，算法精度趋于平稳，故本文选取 k=20 进行后续实验。

2) 在 Netflix_3m1k 数据集中，随着邻居数的变化，Fuzzy-UBCF 的精确度始终高于 Cosine-CF，Cosine-CF 的精确度始终高于 Pearson-CF。

3) 在 Netflix_5m3k 数据集中，用户项目比减小，随着邻居数的变化，Fuzzy-UBCF 的精确度略低于 Cosine-CF，Fuzzy-UBCF 的精确度始终高于 Pearson-CF。

实验结果表明：Fuzzy-UBCF 的算法在邻居数较少时，有较高的精度，因为将评分模糊化后，一个邻居所携带的信息更多。当用户项目比减少时，

Fuzzy-UBCF 的精确度就会下降, 因为 Fuzzy-UBCF 考虑用户向量的距离，用户项目比越小，用户向量长度越长，精确度越低。当用户项目比减小时，Fuzzy-UBCF 性能变差。

Pearson-CF 的效果很差，是因为本数据集稀疏度很高，用户的共同评分项目很少，Pearson-CF 没有发挥出自己的优势，在稀疏度低的数据集中，Pearson-CF 的精度优于 Cosine-CF。但 Fuzzy-UBCF 也是通过用户的共同评分项目进行计算，说明了 Fuzzy-UBCF 的优点。

实验 2 稀疏度对算法精度的影响

在 4.1 节中，分析了 Fuzzy-UBCF 适用于评分矩阵稀疏的数据集，为了比较稀疏度对算法精度的影响，本实验在 Netflix_5m3k 数据集中，保证用户数和项目数不变, 减少评分矩阵的稀疏度, 当 k=20, 比较 3 种算法的精度，实验结果如图 8 所示。

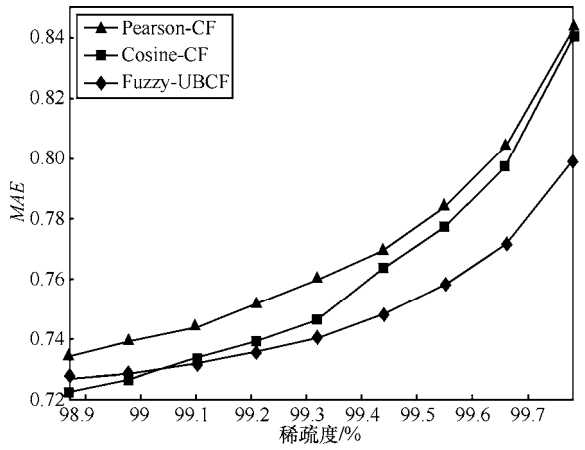

图 8 稀疏度对 3 种算法精确度比较

根据结果可以得出以下结论。

1) 随着稀疏度增加，可用信息减少，3 种算法的精确度都会下降。

2) Pearson-CF 的精度很差，不适用于稀疏度高的数据集。

3) 在稀疏度低于 99.1%时，Fuzzy-UBCF 比 Cosine-CF 稍差, 但随着稀疏度的增高 Fuzzy-UBCF 的精确度高于 Cosine-CF, 而 Cosine-CF 随稀疏度变大，性能恶化很严重。

Fuzzy-UBCF 将评分模糊化后，适用于用户和项目关系不明显的数据集，也就是适用于稀疏度高的数据集。

实验 3 用户项目比对算法精度的影响

在 4.1 节分析中，本文得出 Fuzzy-UBCF 适用于用户项目比大的数据集，本实验在 Netflix_5m3k 数据集中，k=20，保证用户和稀疏度不变，减少项目数来提高用户项目比，比较 3 种算法的精度，实验结果如图 9 所示。

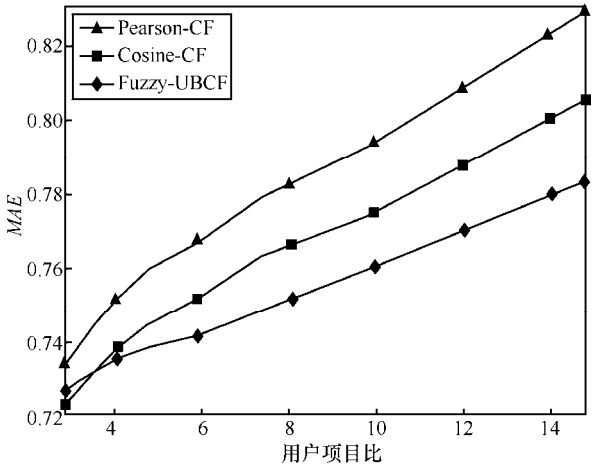

图 9 用户项目比对 3 种算法精确度比较

从结果可以得出以下结论。

1) 随着用户项目比增加，3 种算法的精确度都会下降。这是因为项目数减少，可用信息减少，精度会降低。

2) 随着用户项目比的增加，Fuzzy-UBCF 的精度会优于传统的相似性算法。

可见，Fuzzy-UBCF 适用于用户项目比高的数据集。在实际系统中, 用户数是远远大于项目数的, 并且数据集的稀疏度很高，所以 Fuzzy-UBCF 有很强的实用性。

实验 4 算法运行时间

在 4.2 节中分析了 Fuzzy-UBCF 的运行时间, 本实验来比较 3 种算法的运行时间, 实验结果如图 10 所示。

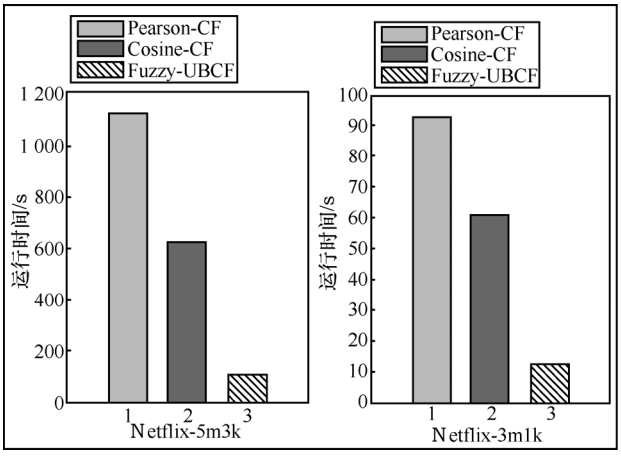

图 10 3 种算法运行时间比较

可以得出，算法的运行时间关系为 Pearson-CF>Cosine-CF>Fuzzy-UBCF。Fuzzy-UBCF 的运行时间远小于传统的相似性计算方法。

实验 5　α 和 β 参数对算法精度的影响

在 3.1 节中，分析了 α 和 β 组合的适用范围，本实验对此进行验证。在 Netflix_5m3k 数据集中，保证用户数和项目数不变，减少评分矩阵的稀疏度，为了避免出现大的误差，本实验只对参数进行微调。

当 β=0.23 时，α 分别为 0.125、0.13 和 0.135 时，比较 Fuzzy-UBCF 的精度。由于 3 组参数的实验值很接近，为了便于比较，本文以 α=0.13、β=0.23 组合为基准，比较 3 种组合与 α=0.13、β=0.23 组合的 *MAE* 差值。实验结果如图 11 所示。

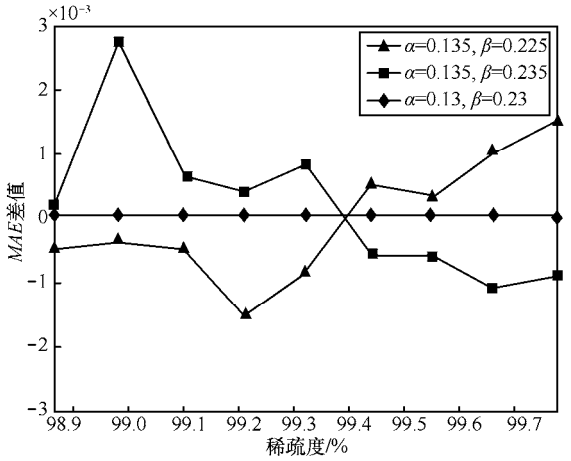

图 11　参数 α 对 Fuzzy-UBCF 的精度的影响

从结果可以得出：当 β 不变，α 变大时，适用稀疏度低的数据集；而 β 不变，当 α 变小时，适用于稀疏度高的数据集。

当 α=0.13，β 分别为 0.225、0.23 和 0.235 时，比较 Fuzzy-UBCF 的精度。和上实验一样，比较 3 种组合与 α=0.13，β=0.23 组合的 *MAE* 差值。实验结果如图 12 所示。

由实验结果可知：当 α 不变，β 变大，适用于稀疏度高的数据集；当 α 不变，β 变小时，适用于稀疏度低的数据集。

6　结束语

本文提出了一种梯形模糊评分模型，将离散的评分模糊化，考虑了评分信息量等因素，能更合理地表达用户的观点，并提出了一种基于用户模糊相似度的协同过滤算法，证明了 Fuzzy-UBCF 是传统协同过滤算法在模糊域上的扩展，通过与传统协同过滤算法比较，实验结果表明本文提出的算法有以下优点。

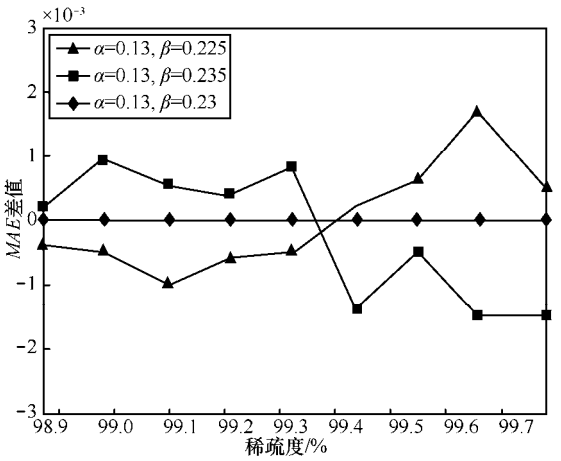

图 12　参数 β 对 Fuzzy-UBCF 的精度的影响

1) Fuzzy-UBCF 更适用于评分矩阵稀疏的数据集。

2) Fuzzy-UBCF 适用于用户项目比大的数据集，而现实的系统中用户项目比都很大，故 Fuzzy-UBCF 有很强的实用性。

3) Fuzzy-UBCF 的算法运行时间远小于传统的协同过滤算法。

本文下一步计划，考虑用户的评分尺度，对梯形模糊评分模型进行优化，并优化模糊相似度中信息量计算部分，寻找到更合理的模糊相似度的加权方法，进一步提高算法精度。

参考文献：

[1] 荣辉桂, 火生旭, 胡春华, 等. 基于用户相似度的协同过滤推荐算法[J]. 通信学报, 2014,35(2):16-24.
RONG H G, HUO S X, HU C H, et al. User similarity based couaborative fettering recommendation algorithm[J]. Journal on Communications, 2014,35(2):16-24.

[2] 李英壮, 高拓, 李先毅. 基于云计算的视频推荐系统的设计[J].通信学报, 2013,34(Z2):138-140.
LI Y Z, GAO T, LI X Y. Design of video recommender system based on cloud computing[J]. Journal on Communications, 2013, 34(Z2): 138-140.

[3] 丁欣, 马严, 吴军. 适用于校园网的视频推荐系统的设计与实现[J]. 通信学报, 2013,34(Z2):175-179.
DING X, MA Y, WU J. Design and implementation of a video recommendation system in campus network[J]. Journal on Communications, 2013,34(Z2):175-179.

[4] ZHAO Z D, SHANG M S. User-based collaborative-filtering recommendation algorithms on hadoop[C]//WKDD'10 Third International Conference on Knowledge Discovery and Data Mining. c2010: 478-481.

[5] YANG J M, LI K F. Recommendation based on rational inferences in collaborative filtering[J]. Knowledge-Based Systems, 2009, 22 (1):105-114.

[6] HUANG C K. Mining the change of customer behavior in fuzzy time-interval sequential patterns[J]. Applied Soft Computing, 2012, 12(3):1068-1086.

[7] YAGER R R. Fuzzy logic methods in recommender systems[J]. Fuzzy Sets and Systems, 2003, 136(2):133-149.

[8] SHAMRI M Y H, BHARADWAJ K K. Fuzzy-genetic approach to recommender system based on a novel hybrid user model[J]. Expert Systems with Applications, 2008, 35(3): 1386-1399 .

[9] LE H S. HU-FCF: a hybrid user-based fuzzy collaborative filtering method in recommender systems[J]. Expert Systems with Applications, 2014, 41(15):6861-6870.

[10] LUCAS J P, LUZ N, MORENO M N, et al. A hybrid recommendation approach for a tourism system[J]. Expert Systems with Applications, 2013, 40(9):3532-3550.

[11] ZADEH L A. Probability measures of fuzzy events[J]. Journal of Mathematical Analysis and Applications, 1968, 23(2):421-427.

[12] CHEN S H. Ranking generalized fuzzy number with graded mean integration[C]//The Eighth International Fuzzy Systems Association World Congress. c1999: 899-902.

[13] CHEN S J, CHEN S M. Fuzzy risk analysis based on similarity measures of generalized fuzzy numbers[J]. IEEE Transactions on Fuzzy Systems, 2003, 11(1):45-56.

[14] ZIEGLER C N, LAUSEN G. Analyzing correlation between trust and user similarity in online communities[J]. Lecture Notes in Computer Science, 2004:251-265.

[15] 朱郁筱, 吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报, 2012, 41(2): 163-175.
ZHU Y X, LYU L Y. Evaluation metrics for recommender systems[J]. Journal of University of Electronic Science and Technology of China, 2012, 41(2): 163-175.

作者简介：

吴毅涛（1991-），男，陕西西安人，国家数字交换系统工程技术研究中心硕士生，主要研究方向为数据挖掘、社会化网络、推荐算法。

张兴明（1963-），男，河南新乡人，国家数字交换系统工程技术研究中心教授，主要研究方向为通信与信息系统、宽带信息网络等。

王兴茂（1989-），男，辽宁营口人，国家数字交换系统工程技术研究中心硕士生，主要研究方向为数据挖掘、用户行为分析、推荐算法。

李晗（1987-），女，河南汤阴人，国家数字交换系统工程技术研究中心工程师，主要研究方向为嵌入式系统。