计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
50(7): 1444-1451, 2013

基于核方法的 User-Based 协同过滤推荐算法

王 鹏 王晶晶 俞能海
(中国科学技术大学电子工程与信息科学系 合肥 230027)
(pengwang@mail.ustc.edu.cn)

A Kernel and User-Based Collaborative Filtering Recommendation Algorithm

Wang Peng, Wang Jingjing, and Yu Nenghai
(*Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei* 230027)

Abstract With the development of information technology, people can get more and more information nowadays. To help users find the information that meets their needs or interest among large amount of data, personalized recommendation technology has emerged and flourished. As a most widely used and successful recommendation technique, collaborative filtering algorithm has widely spread and concerned many researchers. Traditional collaborative filtering algorithms face data sparseness and cold start problems. As traditional algorithms only consider the limited data, it is difficult to estimate the accurate similarity between users, as well as the final recommendation results. This paper presents a kernel-density-estimation-based user interest model, and based on this model, a user-based collaborative recommendation algorithm based on kernel method is proposed. Through mining users' latent interest suggested by the limited ratings, the algorithm can well estimate the distribution of users' interest in the item space, and provide a better user similarity calculation method. A distance measurement based on classification similarity is proposed for the kernel methods, and two kernel functions are investigated to estimate the distribution of user interest. KL divergence is utilized to measure the similarity of users' interest distribution. Experiments show that the algorithm can effectively improve the performance of the recommendation system, especially in the case of sparse data.

Key words collaborative filtering; personalized recommendation; kernel method; data sparseness; similarity measurement

摘 要 作为在实际系统中运用最为广泛和成功的推荐技术,协同过滤算法得到了研究者们的广泛关注.传统的协同过滤算法面临着数据稀疏和冷启动等问题的挑战,在计算用户之间相似度时只能考虑有限的数据,因此难以对用户之间的相似度进行准确的估计.提出了一种基于核密度估计的用户兴趣估计模型,并基于此模型,提出了一种基于核方法的 user-based 协同过滤推荐算法.通过挖掘用户在有限的评分数据上表现出来的潜在兴趣,该算法能更好地描述用户兴趣在项目空间上的分布,进而可以更好地估计用户之间的兴趣相似度.实验表明,该算法可以有效地提高推荐系统的性能,尤其在数据稀疏的情况下能显著地提高推荐结果的质量.

收稿日期:2011-12-17;**修回日期**:2012-06-08
基金项目:**国家自然科学基金重点项目**(60933013);**国家科技重大专项基金项目**(2010ZX03004-003);**中央高校基本科研业务费专项基金项目**(WK2100230002)
通信作者:**俞能海**(ynh@ustc.edu.cn)

关键词　协同过滤;个性化推荐;核方法;数据稀疏;相似性度量

中图法分类号　TP391

随着信息科技以及 WEB 2.0 技术的迅速发展,互联网信息日趋庞大且保持高速增长.对于互联网用户而言,要解决的问题是如何高效快速地从海量信息中挖掘对自己有价值的信息,而对于一些社交网站、电子商务网站等站点,要考虑如何有效地将网站内容呈现给用户,进而提高服务质量,使得用户粘性增加并吸引更多的用户,最后得以盈利.个性化推荐技术[1-2]正是在这样的背景下逐步发展起来的.

作为传统推荐技术的一种,协同过滤算法由于其简单高效的特点,在实践中获得了更多的青睐,同时也受到了大量研究者的关注.协同过滤推荐算法可以分为基于邻居集(neighborhood-based)[3-4]和基于模型(model-based)[5-6]的两种.其中,基于邻居集的推荐算法又可以分为基于用户(user-based)[4]和基于项目(item-based)[3]的算法.协同过滤推荐算法的核心思想是利用近似用户或者用户喜欢的项目的近似项目来过滤大量信息,从而为用户筛选出其可能感兴趣的项目.具体地,user-based 推荐算法将与特定用户兴趣相似的用户所喜欢的项目推荐给该用户;而 item-based 推荐算法是筛选出那些与用户喜欢的项目相似的项目作为推荐结果.model-based 算法是利用统计和机器学习技术得到一个推荐模型,进而用该模型产生推荐结果.在算法选择方面,根据 Wolpert 等人提出的 NFL 定理(no free lunch theorem)[7],评判一个算法的优劣必须与特定的语境与应用联系起来.user-based 推荐算法能够挖掘出用户的潜在兴趣,并且也能够针对推荐结果进行合理地解释,因此得到了广泛地应用.

然而,不可避免地,user-based 推荐算法也需要考虑各种协同过滤推荐算法中普遍存在的弊端,归纳起来主要是:第一,传统相似性度量方法在计算项目或者用户间的相似性时只考虑了有共同评分的数据,导致只有拥有共同评分项目的用户有相似的可能,与实际情况不符;第二,协同推荐面临着数据稀疏和冷启动问题的挑战.针对协同过滤推荐算法的不足,研究者提出了多种解决方案,主要包括矩阵填充和矩阵降维两种技术.矩阵填充技术包括用默认的值填充缺失的数据[8-9],以及利用项目本身的内容信息预测缺失的评分[10-11].该策略在填充缺失数据的同时也引入了新的误差,并且对后者来说,常常需要特定领域的语义知识,导致其难以在不同类项目间推广,适用范围较窄.矩阵降维技术是利用 SVD 等矩阵分解技术[12],通过分解用户-项目评分矩阵或者稀疏的相似性矩阵来将高维数据投影到低维数据,进而发现项目或用户间隐含的相似性.矩阵降维技术有 2 个缺点,首先,降维导致信息丢失,某些情况下会影响推荐精度;其次,通过降维技术得到的推荐结果难以给出易于理解的解释,这将影响用户对推荐结果的接受程度.

从传统推荐算法中用户相似性度量方法的缺陷以及面临的数据稀疏与冷启动问题出发,本文提出了一种基于核方法的 user-based 协同过滤推荐算法(kernel and user-based collaborative filtering recommendation algorithm, KUCF).该算法利用核密度估计技术构建用户兴趣模型,并在此兴趣模型的基础上提出了相应的相似性度量方法.该用户兴趣模型以及对应的相似性度量方法能够更好地反映实际系统中的用户兴趣分布的情况.相对于传统的 user-based 推荐算法,KUCF 不需要引进手动调节的参数,并且能通过估计用户对于未评分项目的兴趣程度来更好地估计用户相似性.实验表明,在数据稀疏的情况下,该算法能够显著地提高推荐系统的质量.

1　相关工作

user-based 推荐算法有 3 个关键步骤.首先是用户兴趣建模,即用一定的数学模型来表示用户在整个项目空间的兴趣分布;然后,利用该兴趣模型计算用户间的兴趣相似度,产生基于该相似性度量的邻居集;最后将目标用户的邻居所感兴趣的项目通过一定的推荐策略返回给用户.

1.1　数据表示与用户兴趣模型

在协同过滤推荐系统中,我们通过用户对系统中项目评分的历史数据来预测用户对未评分项目的喜好程度.假设分别有 m 个用户和 n 个项目,定义用户集合 $U=\{user_1, user_2, \cdots, user_m\}$,以及项目集合 $I=\{item_1, item_2, \cdots, item_n\}$.假设用户 u 对项目 i 的评分为 $r_{u,i}$,则我们可以用一个用户-项目评分矩阵来表示相应的推荐系统中所处理的数据,如图 1 所示:

	$item_1$	$item_2$	…	$item_n$
$user_1$	$r_{1,2}$	$r_{1,2}$	…	$r_{1,n}$
$user_2$	$r_{2,1}$	$r_{2,2}$	…	$r_{2,n}$
⋮	⋮	⋮	⋮	⋮
$user_m$	$r_{m,1}$	$r_{m,2}$	…	$r_{m,n}$

Fig. 1 Matrix of user-item ratings.

图 1 用户-项目评分矩阵

需要注意的是,这里并非所有的评分数据 $r_{u,i}$ 都存在,我们的任务正是利用可获取的评分数据来预测用户-项目评分矩阵的缺失数据. 有了用户-项目评分矩阵,我们将很容易地得到常用的用户兴趣模型,即用该用户-项目评分矩阵的行向量来表示对应用户的兴趣. 该兴趣模型简单,便于计算,但向量中存在缺失数据,所以并不能精确地表示用户兴趣在整个项目空间的分布.

1.2 相似性度量

常用的并且具有代表性的相似性度量方法有 3 种:余弦相似性[13]、Pearson 相关系数[8]和修正的余弦相似性[14]. 虽然针对不同的数据集各种相似性度量有不同的表现,但总的来说,采用 Pearson 相关系数和修正的余弦相似性的推荐系统具有较好的推荐性能.

1) Pearson 相关系数

设 $I_u=\{i:i\in I, r_{u,i}\neq\varnothing\}$ 为用户 u 评分过的项目集合,$\bar{r}_{u,*}$ 为用户 u 产生的评分的均值. 则由 Pearson 相关系数计算用户相似性的方法如式(1)所示:

$$corr_{u,v}=\frac{\sum_{i\in I_u\cap I_v}(r_{u,i}-\bar{r}_{u,*})(r_{v,i}-\bar{r}_{v,*})}{\sqrt{\sum_{i\in I_u\cap I_v}(r_{u,i}-\bar{r}_{u,*})^2}\sqrt{\sum_{i\in I_u\cap I_v}(r_{v,i}-\bar{r}_{v,*})^2}}. \tag{1}$$

2) 修正的余弦相似性

余弦相似性将用户的评分看作 n 维项目空间上的向量,用两个用户评分向量夹角的余弦值来衡量两个用户的相似性. 但该方法忽略了用户评分尺度的差异,修正的余弦相似性通过将所有评分减去用户对项目的平均评分来改善,计算方法如式(2)所示:

$$corr_{u,v}=\frac{\sum_{i\in I_u\cap I_v}(r_{u,i}-\bar{r}_{u,*})(r_{v,i}-\bar{r}_{v,*})}{\sqrt{\sum_{i\in I_u}(r_{u,i}-\bar{r}_{u,*})^2}\sqrt{\sum_{i\in I_v}(r_{v,i}-\bar{r}_{v,*})^2}}. \tag{2}$$

如前所述,传统的相似性度量方法都只考虑了有共同评分的那些项目. 而实际上,那些缺失评分的项目可能隐含了用户其他方面的兴趣,导致相似性度量存在偏差. 在数据稀疏的情况下这种偏差更为明显.

1.3 评分预测规则

根据用户间的相似度可以划分目标用户的最近邻居集,然后用相似性作为权重可以得到目标用户邻居集中用户对目标项目的评分的加权平均,该值可作为目标用户对目标项目评分的预测. 考虑到不同的用户有不同的评分尺度,故将评分预测策略作相应的修改,得到式(3)所示的预测规则:

$$p_{u,i}=\mu+\frac{\sum_{v\in N_u}corr_{u,v}\times(r_{v,i}-\mu)}{\sum_{v\in N_u}|corr_{u,v}|}, \tag{3}$$

其中,μ 为用户 u 的评分均值,N_u 为用户 u 的邻居集.

1.4 矩阵填充技术

引言中提到,为了解决协同过滤的数据稀疏问题,研究者提供了利用矩阵填充和矩阵降维技术的解决方案. 矩阵降维技术属于 model-based 的协同过滤推荐,基于该模型的推荐技术难以对推荐结果进行解释,并将影响用户对推荐结果的接受程度,而这恰恰是设计一个推荐系统应当考虑的重要因素之一[15-16]. 矩阵填充技术是解决 user-based 协同过滤算法面临的数据稀疏与冷启动问题的一种重要手段. 通用的矩阵填充技术包括采用默认值填充和均值填充. 前者将未评分的数据用一个预设的初值填充,后者是用对应用户的平均打分或者对应项目的平均得分来填充缺失的数据. 在通用的矩阵填充技术基础之上,研究者们还提出了利用额外的内容信息等对缺失的评分项进行初步的预测,并用预测分数对矩阵进行填充,但该方法常常需要特定领域的语义知识以及难以在不同类项目间推广,适用范围较窄. 无论是怎样的填充策略,该技术在填充缺失数据的同时也引入了新的误差.

2 算法设计

传统的协同过滤推荐算法在估计用户间相似度时仅仅考虑有共同评分的项目,在数据稀疏的情况下会对推荐结果造成较大的误差. 本文提出的 KUCF 算法能够通过挖掘用户在已知数据上的兴趣度在缺失数据上的扩散来估计用户的兴趣分布,

因而可以更好地对用户兴趣建模. 下面介绍 KUCF 算法的 3 个关键步骤.

2.1　分类相似度

推荐系统中所涉及到的项目往往会有相应的类别信息,例如电影评价网站将电影分为喜剧片、冒险片、战争片、动作片等. 一个项目可能同时属于几个类别,例如可以同属于动作片和冒险片. 定义项目类别集合 $C=\{c_1,c_2,\cdots,c_k\}$,其中 c_k 为某一个类别,定义 $C_i\subseteq C$ 为项目 i 的所属类别集合. 为了接下来的用户兴趣估计,这里我们定义 Item 间的分类相似度,如式(4)所示:

$$sim_c(i,j)=\frac{|C_i\cap C_j|^2}{|C|\times|C_i\cup C_j|}. \tag{4}$$

该相似度定义一方面考虑到 2 个项目所占类别中重合的比例,另一方面也考虑了重合的类别在整个类别集合中所占的比例. 在后面我们用到 2 个项目的距离度量,相似性越大,则 2 个项目在项目空间上的距离越小,因此我们定义 2 个项目 i, j 之间的距离度量为

$$d_{i,j}=1-sim_c(i,j). \tag{5}$$

2.2　用户兴趣估计

在传统的用户相似性算法中,仅仅考虑有共同评分的那些项目. 而实际上,用户对于尚未评分的那些项目也有自己的喜好,所以,如果能估计用户在整个项目空间上的兴趣密度分布,然后再计算两个用户兴趣密度分布的相似性更加符合实际情况. 在数据稀疏的情况下,由于绝大多数数据都未知,这种估计方法有着更为重要的意义. 本文将核密度估计方法的思想应用到用户兴趣估计上来. 核密度估计是统计学中非参数估计方法之一[17],用户兴趣密度的多模性(即有多个局部极大值,如图 2 所示)使其更适于采用非参数的方式进行估计.

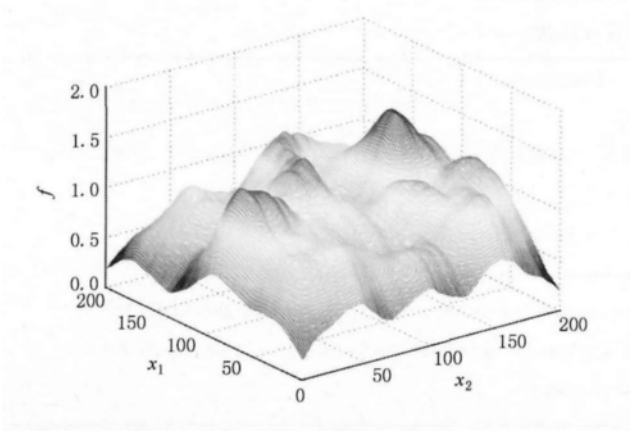


Fig. 2　Distribution of user interest.

图 2　用户兴趣分布示意图

设 $X_1,X_2,\cdots,X_n$ 为总体分布 X 的独立同分布样本,X 的密度函数 $f(X)$的核密度估计定义如下:

$$\hat{f}(X)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{\|X-X_i\|}{h}\right), \tag{6}$$

其中,$K\left(\frac{\|X-X_i\|}{h}\right)$为核函数,$h$ 通常称为核函数的窗宽,为固定值. 常用的核函数有均匀核函数、三角核函数、高斯核函数等. 文献[17]指出,核的形状对结果的影响比窗宽要小得多. 根据中心极限定理,高斯函数是复杂总和的有限机率分布,而用户兴趣恰恰可以看作是包含多种不确定因素的一个有限机率分布. 因此,这里我们首先考量高斯核:

$$K_g(z)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{z^2}{2h^2}\right\}. \tag{7}$$

用高斯核估计用户 u 兴趣分布 P_u 的公式如式(8)所示:

$$\hat{f}_{P_u}(j)=\frac{1}{|I_u|\times\sqrt{2\pi}h}\sum_{i\in I_u}r_{u,i}\times\exp\left\{\frac{d_{i,j}^2}{2h^2}\right\}. \tag{8}$$

除高斯核外,本文还考量了三角核,如式(9)所示,需要注意的是,为了与高斯核的窗宽定义对应,这里我们定义三角核的窗宽 h 也为半功率窗宽.

$$K_t(z)=\begin{cases}0, & \text{if } |z|>\sqrt{2}h,\\ \dfrac{\sqrt{2}h-|z|}{2h}, & \text{otherwise.}\end{cases} \tag{9}$$

相应的兴趣分布计算方法如式(10)所示:

$$\hat{f}_{P_u}(j)=\sum_{i\in I_u}\frac{r_{u,i}\times(\sqrt{2}h-d_{i,j})\times u(\sqrt{2}h-d_{i,j})}{|I_u|\times 2h^2}, \tag{10}$$

其中,$u(\cdot)$为阶跃函数:

$$u(z)=\begin{cases}0, & \text{if } z<0,\\ 1, & \text{if } z\geqslant 0.\end{cases} \tag{11}$$

2.3　用户相似性计算

2.2 节中我们通过核密度估计得到了用户的兴趣分布,本节说明在已知兴趣分布的情况下如何计算 2 个用户的相似度. 在信息论中,相对熵(又称为 KL 散度)是 2 个概率分布之间差别的非对称性度量. 类似于文献[18]中计算分布相似度的策略,在本文的算法中,我们也采用 KL 散度计算用户相似度. 假设 P_u 为核密度估计方法得到的用户 u 的兴趣密度函数,则和 P_v 的 KL 散度定义为

$$D_{\mathrm{KL}}(P_u\|P_v)=\sum_{i=1}^{K}P_u(i)\log\frac{P_u(i)}{P_v(i)}. \tag{12}$$

由于 KL 散度不具有对称性,所以不能成为一

个真正意义上的度量标准,本文采用式(13)计算用户间的相似性:

$$corr_{u,v}=\frac{1}{2}(D_{KL}(P_u \parallel P_v)+D_{KL}(P_v \parallel P_u)). \tag{13}$$

2.4 算法描述

综合 2.1～2.3 节中的计算步骤和 1.3 节中评分预测规则,得到了如下描述的 KUCF 算法:

算法 1. 基于核方法的 user-based 协同推荐.

输入:用户-项目评分矩阵;

输出:用户 u 对项目 i 的预测评分.

① 利用式(5)计算项目 i 与其他项目的距离;

② 由式(8)或者式(10)计算用户的兴趣在项目空间上的分布;

③ 重复步骤①②,计算所有用户的兴趣分布;

④ 由式(13)计算 2 个用户间的相似性;

⑤ 利用式(3)作出最终预测.

3 实验结果及分析

3.1 数据集

本文采用 GroupLens 研究组[19]提供的 MovieLens 数据集 ML-100K 对算法进行评估,该数据集包含 943 个用户对 1682 部电影的评分记录,每个用户至少对 20 部电影进行了评分. 评分范围为 1～5,1 表示最不喜欢,5 表示最喜欢. 对于每个用户,抽出 10 个评分数据划分到测试集中,剩余的为训练集. 对于 user-based 个性化推荐算法,该数据集中用户数要小于项目数,因此是比较稀疏的. 为了进一步衡量本文算法在数据稀疏情况下的性能,本文在此训练集基础上,随机筛选出更为稀疏的子训练集,包含历史评分数量为原数据集的 10%～90%. 除此之外,该数据集还包含了 5 组由原始数据的 80%和 20%划分的训练集和测试集用于交叉检验. 最后,值得说明的是,该数据集中包含的电影项目分别属于 18 个类别中的 1 个或几个,分类数据可用于计算项目间的分类相似性.

3.2 评估标准

本文采用平均绝对偏差(mean absolute error, *MAE*)和精确度(*Precision*)两个度量来评估实验结果. *MAE* 通过计算预测值与实际值的平均误差大小来度量系统的推荐质量,*MAE* 越小推荐质量越高. *MAE* 的计算方法如式(14)所示,其中,T 为测试集,$r_{u,i}$ 为 T 中用户 u 对项目 i 的评分.

$$MAE=\frac{\sum_{r_{u,i}\in T}|p_{u,i}-r_{u,i}|}{r_{u,i}}. \tag{14}$$

精确度是计算预测评分与实际评分相等的项在整个测试集中所占的比率,计算公式如式(15)所示,其中 $|R|=|\{r_{u,i}:r_{u,i}\in T, p_{u,i}=r_{u,i}\}|$ 为测试集中预测评分与实际评分相等的数目,$|T|$ 为测试集中评分总数. 精确度越大推荐质量越高.

$$Precision=\frac{|R|}{|T|}. \tag{15}$$

3.3 实验结果及分析

3.3.1 实验设计

本文设计了 3 组实验,分别从核函数选取、项目空间度量函数选取以及与传统的推荐算法比较 3 个方面来探索基于核方法的推荐算法的性能. 为便于描述实验结果,本文采用表 1 中的缩写来表示对应的算法.

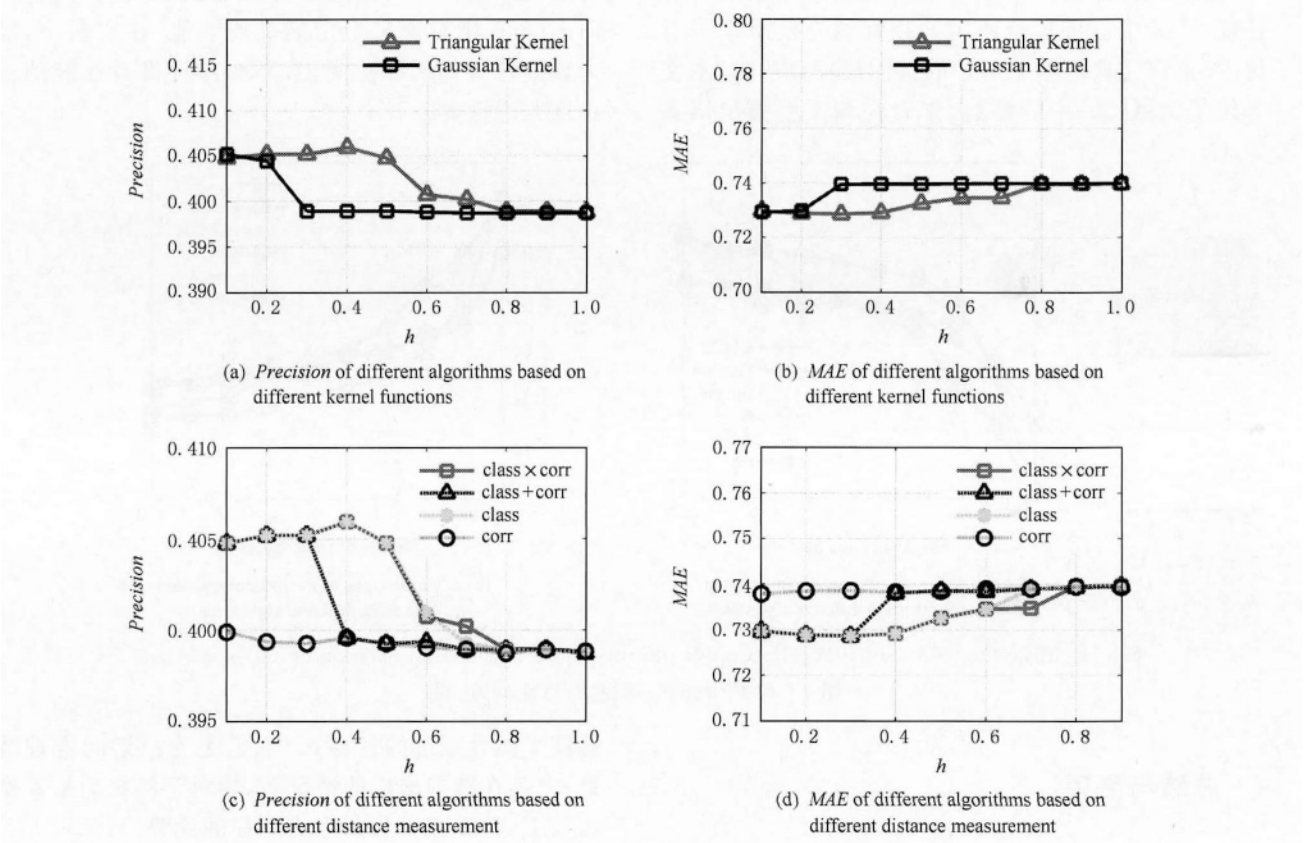
Table 1 Algorithm abbreviations

表 1 算法缩写术语表

Abbreviation	Description
class	Distance measurement descripted in Formula (5)
corr	Distance measurement based on pearson correlation coefficient
class×corr	Distance measurement based on the product of class similarity and pearson correlation coefficient
class+corr	Distance measurement based on the sum of class similarity and pearson correlation coefficient
Cosine	User-based collaborative recommendation algorithm based on ajusted cosine, with neighor set of 50
Pearson	User-based collaborative recommendation algorithm based on pearson correlation, with neighor set of 50
KUCF	User-based collaborative recommendation algorithm based on kernel method, using triangular kernel, $h=0.4$
m_1	Matrix completion using the mean ratings of certain item
m_2	Matrix completion using the mean ratings of certain user
m_3	Matrix completion using the given value 3

3.3.2　不同核函数在不同窗宽上的效果

本组实验考量在用核密度估计方法估计用户兴趣分布时,核函数以及窗宽的选取对结果的影响.在这里,我们比较了三角核以及高斯核对推荐结果的影响.图 3(a)(b)分别是基于不同核函数的推荐算法的精确率和 MAE,横坐标为核函数的窗宽.从图 3 我们可以看出,在窗宽充分大(比如 $h=1$)或者窗宽充分小(比如 $h=0.1$)时,核函数的选取对结果影响并不大;而当窗宽从 0.1 逐渐增大时,基于三角核的推荐算法显现了其相对于基于高斯核的推荐算法的优势.从总体上来看,两种算法的性能都有随着窗宽的增大而逐渐降低的趋势,这是因为窗宽的增大会在估计兴趣时将用户对某项目的兴趣扩散到与该项目相似性不大的项目上,因而引入了相应的误差.这也正解释了三角核函数优于高斯核函数的现象,因为三角核函数的定义域为一个闭区间,因而在估计用户兴趣时,将特定兴趣的影响限定在一个有界的空间中,排除了微小项的误差.虽然核函数的选取并没有显著地影响推荐算法的性能,但基于三角核的推荐算法在更多的窗宽下展现了其性能优势,因此,后面的实验都将基于三角核估计用户兴趣分布,并选取窗宽 $h=0.4$,交叉检验的结果也支持了这种选择.



(a) *Precision* of different algorithms based on different kernel functions

(b) *MAE* of different algorithms based on different kernel functions

(c) *Precision* of different algorithms based on different distance measurement

(d) *MAE* of different algorithms based on different distance measurement

Fig. 3　Efficiency of recommender systems using different kernel functions and distance measurements.

图 3　不同核函数及距离函数的系统性能

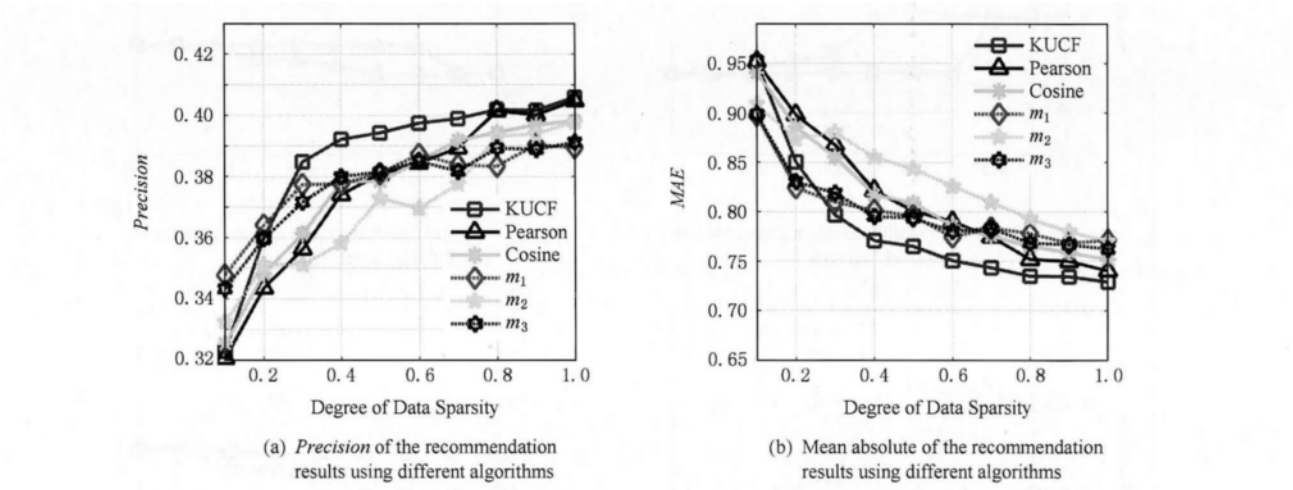
3.3.3　项目空间上的距离函数对推荐结果的影响

在估计用户兴趣分布时,项目空间上的距离度量是一个关键的参数.我们用式(5)来计算项目空间上的距离,除了基于分类相似性的距离度量外,考察在项目空间上是否有更好的度量函数是本组实验的目的.直观地,除了分类相似性外,还有传统的 Pearson 相关系数等方法,因此,这里我们考察了基于分类相似性的距离度量、基于 Pearson 相关系数的距离度量以及两者的结合对推荐结果的影响.图 3(c)(d)显示了本组实验的结果.由图 3 可以看出,采用基于分类相似度的距离度量的算法在推荐质量上要优于基于 Pearson 相关系数的距离度量的算法.这是因为后者在估计用户兴趣的扩散时反而忽略了项目间的相关性,而实际的项目空间中,起于原点并终于每个项目所在位置的向量并不一定正交.实践经验告诉我们,人们的兴趣更多地会分布在特定几个类别的项目上,因此,在这里基于类别的 18 维项目空间更符合实际情况.同时,值得注意的是,在实验过程中我们发现,基于两者乘积的距离度量会比单独采用基于分类相似性的度量产生更好的推荐结果,

这是因为前者在考虑项目类别属性的同时,也考虑了项目的个体属性对距离度量的影响.采用两者之和的距离度量效果较差是因为分类相似性和Pearson相关系数并不具有一致的量纲,因而其和产生的推荐结果不可预料.

3.3.4 推荐算法在不同数据稀疏程度下的表现

在3.1节中,我们提到不同稀疏程度的数据集,本组实验的目的是考察不同算法在数据稀疏时的表现.为了比较本文提出的兴趣模型以及对应的相似度计算方法与传统的兴趣模型对推荐系统性能的影响,我们重点考察了KUCF与经典算法的比较.同时,为了进一步检验KUCF在数据稀疏的情况下的性能,实验中也将其与通用的数据填充技术作了对比.在实现经典的基于用户的协同推荐算法时,本文采用了文献[20]中的参数,作者在该文中指出该参数环境改善了推荐结果.如图4所示,KUCF在数据稀疏的情况下相对于经典的基于用户的协同过滤推荐有明显的优势,鉴于KUCF能够充分地从已有数据中挖掘用户对缺失数据的评分,该结果是可预料的.此外,从图4可以看出,传统的数据填充技术虽然在个别情况下会对推荐结果有所改进,但是难以提供稳定的推荐性能,比如到数据完整度为原始数据的60%以后,推荐性能反而弱于经典的协同过滤算法.除了比较极端的情况,比如数据极度稀疏,KUCF展现了其在不同数据集上可以产生相对稳定的推荐结果的优势.在我们的实验中,传统的推荐算法在数据集不同时其推荐性能会有较大的波动,而KUCF则有比较稳定的推荐性能.在实践中,因为数据每时每刻都在变化,因此这种推荐性能的稳定性尤为重要.



(a) *Precision* of the recommendation results using different algorithms

(b) Mean absolute of the recommendation results using different algorithms

Fig. 4 Efficiency of user-based collaborative filtering algorithms in the condition of data sparseness.

图4 数据稀疏下不同推荐算法性能比较

4 总结与展望

作为在实际系统中运用最为广泛和成功的推荐技术,协同过滤算法得到了研究者们的广泛关注.然而传统的协同过滤算法面临着数据稀疏和冷启动问题等的挑战,在计算用户之间相似度时只考虑有限的数据,因此难以对用户之间的相似度进行准确地估计.针对传统的推荐算法的不足,本文提出了一种基于核密度估计的用户兴趣估计模型;基于此模型,提出了一种基于核方法的user-base协同推荐算法KUCF.KUCF算法充分挖掘了已有数据与缺失数据的潜在联系,通过将用户兴趣从已知的项目点上扩散到整个项目空间,能更好地估计用户兴趣的分布,进而可以更准确得估计用户间的相似性,最终产生更符合实际情况的推荐结果.实验表明,KUCF提高了推荐系统的性能,产生了更为稳定的推荐结果;尤其在数据稀疏的情况下,相对于传统的推荐算法,KUCF在推荐质量上有明显的优势.

KUCF采用项目间的分类相似性作为项目空间上的距离度量,这实际上引入了额外的分类信息.并且,基于类别信息张成的项目空间也仅仅是对实际的项目空间的一个估计.在实验中我们发现,基于传统相似性的兴趣模型与基于分类相似性的兴趣模型对推荐结果的影响差别较小,两者都优于传统的推荐算法.但项目空间上是否存在更好的距离度量仍值得更进一步的研究,因此通过对已有数据的进一步挖掘,去发现项目空间更本质的特征是进一步研究所需要关注的重点.此外,本算法引入了大量的计算,虽然兴趣估计以及相似度计算可离线进行,但如何在有新的数据加入时及时对离线数据进行更新以提高系统的可扩展性也是本研究下一步的工作方向.

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 the 1994 ACM Conf on Computer Supported Cooperative Work. New York: ACM, 1994: 175-186

[2] Hill W, Stead L, Rosenstein M, et al. Recommending and evaluating choices in a virtual community of use [C] //Proc of the SIGCHI Conf on Human Factors in Computing Systems. New York: ACM, 1995: 194-201

[3] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms [C] //Proc of the 10th Int Conf on World Wide Web. New York: ACM, 2001: 285-295

[4] Shi Y, Larson M, Hanjalic A. Exploiting user similarity based on rated-item pools for improved user-based collaborative filtering [C] //Proc of the 3rd ACM Conf on Recommender Systems. New York: ACM, 2009: 125-132

[5] Kamishima T, Akaho S. Nantonac collaborative filtering: a model-based approach [C] //Proc of the 4th ACM Conf on Recommender Systems. New York: ACM, 2010: 273-276

[6] Zhou Ke, Yang Shuanghong, Zha Hongyuan. Functional matrix factorizations for cold-start recommendation [C] // Proc of the 34th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2011: 315-324

[7] Wolpert D H, Macready W G. No free lunch theorems for optimization [J]. IEEE Trans on Evolutionary Computation, 1997, 1(1): 67-82

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

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

[10] Degemmis M, Lops P, Semeraro G. A content-collaborative recommender that exploits wordnet-based user profiles for neighborhood formation [J]. Journal of User Modeling and User-Adapted Interaction. 2007, 17(3): 217-255

[11] Melville P, Mooney R J, Nagarajan R. Content-boosted collaborative filtering for improved recommendations [C] // Proc of the 18th National Conf on Artificial Intelligence. Menlo Park: American Association for Artificial Intelligence, 2002: 187-192

[12] Zhao Liang, Hu Naijing, Zhang Shouzhi. Algorithm design for personalization recommendation system [J]. Journal of Computer Research and Development, 2002, 39(8): 986-991 (in Chinese)
(赵亮, 胡乃静, 张守志. 个性化推荐算法设计[J]. 计算机研究与发展, 2002, 39(8): 986-991)

[13] Billsus D, Pazzani M J. Learning collaborative information filters [C] //Proc of the 15th Int Conf on Machine Learning. San Francisco: Morgan Kaufmann, 1998: 46-54

[14] Last. fm. Music recommendation service [EB/OL]. [2012-05-30]. http://www.last.fm

[15] Pu P, Chen L. Trust building with explanation interfaces [C] //Proc of the 11th Int Conf on Intelligent User Interfaces. New York: ACM, 2006: 93-100

[16] Pu P, Chen L. Trust-inspiring explanation interfaces for recommender systems [J]. Journal of Knowledge Based Systems, 2007, 20(6): 543-556

[17] Givens G H, Hoeting J A. Computational Statistics [M]. New York: Wiley-Interscience, 2005

[18] Jiang Kai, Wang Peng, Yu Nenghai. ContextRank: Personalized tourism recommendation by exploiting context information of geotagged Web photos [C] //Proc of the 6th Int Conf on Image and Graphics. Los Alamitos, CA: IEEE Computer Society, 2011: 931-937

[19] GroupLens. MovieLens Data Sets [EB/OL]. [2012-05-30]. http://www.grouplens.org

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

Wang Peng, born in 1986. Master candidate in signal and information processing from the University of Science and Technology of China since 2009. His main research interests include data mining, information retrieval and personalized recommendation.

Wang Jingjing, born in 1987. Master candidate in signal and information processing from the University of Science and Technology of China since 2010. His main research interests include information retrieval, data mining, and personalized recommendation (kkwang@mail.ustc.edu.cn).

Yu Nenghai, born in 1964. Professor and PhD supervisor of the University of Science and Technology of China. His main research interests include multimedia data processing and analysis, information retrieval, and security of digital content (cloud computing and cloud computing security).