DOI：10.7652/xjtuxb201506013

邻域线性最小二乘拟合的推荐支持度模型

孙忱[1]，奚宏生[1]，高荣[1,2]
（1.中国科学技术大学自动化系，230027，合肥；2.广西财经学院信息与统计学院，530003，南宁）

摘要：针对协同过滤推荐系统在稀疏数据集条件下推荐准确度低的问题，提出了推荐支持度模型以及用于该模型计算的邻域线性最小二乘拟合的推荐支持度评分算法（linear least squares fitting，LLSF）。该模型描述用户对被推荐项目更感兴趣的可能性，通过用高支持度的评分估计取代传统的期望估计法来找出用户更喜欢的项目，从而提高推荐的准确度，并从理论上论述了该算法在稀疏数据集条件下相对其他算法具有更强的抗干扰能力。该模型还易于与其他推荐模型融合，具有很好的可拓展性。实验结果表明：LLSF 算法显著提升了推荐的准确性，在 MovieLens 数据集上，F_1 分数可达到传统的 kNN 算法的 3 倍多，对于越是稀疏的数据集，准确率提升幅度越大，在 Book-Crossing 数据集上，当稀疏度由 91% 增加到 99% 时，F_1 分数的改进由 22% 提高到 125%。同时该方法不会牺牲推荐覆盖率，可以保证长尾项目的挖掘效果。

关键词：协同过滤；推荐系统；邻域线性最小二乘拟合；推荐支持度
中图分类号：TP391；TP274 **文献标志码**：A **文章编号**：0253-987X(2015)06-0077-07

A Recommendation-Support Model Using Neighborhood-Based Linear Least Squares Fitting

SUN Chen[1], XI Hongsheng[1], GAO Rong[1,2]
(1. Department of Automation, University of Science and Technology of China, Hefei 230027, China;
2. School of Information and Statistics, Guangxi University of Finance and Economics, Nanning 530003, China)

Abstract: A recommendation-support model and a neighborhood-based linear least squares fitting (LLSF) algorithm for the calculation of recommendation-support rating are proposed to solve the low accuracy problem of collaborative filtering based recommender systems on sparse data sets. The model focuses on the probability of users' more interests on the recommended items, and uses the estimation with high recommendation-support rating to replace the traditional expectation estimation so that users' preferred items are found and the accuracy of recommendation is improved. A theoretical analysis shows that the anti-interference ability of the LLSF algorithm is better than those of other algorithms under the condition of sparse data sets. The model is also expansible by integrating other models. Experimental results show that the LLSF algorithm improves the recommendation accuracy remarkably. The F_1 score is 3 times of that of the traditional kNN algorithm on the MovieLens data set. The more sparse the data set is, the more the improvement on accuracy obtains. When the sparsity grows from 91% to 99% on the Book-crossing data set, the improvement on F_1 scores increases from 22% to 125%. Moreover, the

收稿日期：2014-08-18。 作者简介：孙忱（1981—），女，博士生；奚宏生（通信作者），男，教授，博士生导师。 基金项目：国家自然科学基金重点资助项目（61233003）；国家自然科学基金资助项目（61262002）；广西省自然科学基金资助项目（2013GXNSFBA019274）；广西省社科规划研究资助项目（13BXW007）。
网络出版时间：2015-03-23 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20150323.1713.002.html

algorithm can guarantee the ability of long tail mining without loss of recommendation coverage.

Keywords: collaborative filtering; recommender system; neighborhood-based linear least squares fitting; recommendation-support model

随着互联网的迅速发展,推荐系统已被越来越多地运用到各种网站和电子商务系统中,它不需要用户提供明确的需求,而是通过分析用户的历史行为给用户的兴趣建模,从而主动给用户推荐能够满足他们兴趣和需求的个性化信息[1]。协同过滤算法是最重要的推荐系统技术之一,其原理是根据用户或项目的相似性来预测并推荐当前用户没有进行过评分、购买或浏览等行为,但是很可能会感兴趣的信息[2]。基于邻域的协同过滤推荐由于计算实时性好、可扩展性高、意义清晰易于解释等特点,应用最为广泛[3]。数据稀疏性问题是绝大部分电子商务推荐系统面临的最大挑战,这是因为相似度的计算是基于用户或项目的共同历史行为的,当数据非常稀疏时,就会使得相似度计算不可靠,从而影响基于邻域的协同过滤推荐的准确性[4]。

研究人员提出了各种方法来提高数据稀疏条件下协同过滤推荐的准确性。Sarwar 等研究了不同相似度计算方法及不同数据稀疏度对准确性的影响[5]。黄创光等通过自适应选择近邻数目的方法来缓解数据稀疏带来的问题[6]。罗辛等把共同评分的数目转化为相似度支持度的概念来引入计算[7]。这些方法通过提升预测的准确性来提高推荐的准确性,但是提升效果有限。Adamopoulos 另辟蹊径,通过高百分比的加权计算法,改变了通常的平均加权评分预测方式,大大提高了推荐准确性[8]。然而,Adamopoulos 所使用的线性插值方法容易受到数据扰动的影响,当数据稀疏时推荐效果急剧降低。

注意到虽然推荐以预测为基础,但侧重点并不相同。本文专注于解决稀疏数据条件下提高推荐效果的问题,在现有研究的基础上,提出推荐支持度的概念,选择合适的推荐支持度实现更有效的推荐,同时设计一种新的邻域线性最小二乘拟合的方法来进行计算。实验结果表明,本文提出的方法能大幅度提高推荐的准确性,同时还可略微提升或至少不会牺牲推荐的覆盖率。当数据越是稀疏时,本文方法所能提供的改进就越明显。

1 相关工作

目前广泛研究和应用的推荐系统技术绝大多数都是由基于邻域的协同过滤方法拓展或融合而来,而 Adamopoulos 进一步提出了加权百分比的方法来提高推荐准确度。

1.1 基于邻域的协同过滤推荐

基于邻域的协同过滤算法[1-5,9-10]分为基于用户的算法和基于项目的算法,两者计算原理相同,只是考察维度相互对换,下面以基于用户的算法为例进行介绍。

一般地,把用户 u 对项目 i 的评分 r_{ui} 作为该用户对该项目感兴趣的程度,基于用户的算法通过找出与某用户 u 最相似的 k 个用户(称为用户 u 的 k-近邻)来估计用户 u 对其没有做过评分项目的可能评分。

不同用户 u 与 v 之间的相似性是一个测度,一般可以选用皮尔逊相似度

$$w_{uv}=\frac{\sum\limits_{i\in I_u\cap I_v}(r_{ui}-\bar{r}_u)(r_{vi}-\bar{r}_v)}{(\sum\limits_{i\in I_u}(r_{ui}-\bar{r}_u)^2)^{1/2}(\sum\limits_{i\in I_v}(r_{vi}-\bar{r}_v)^2)^{1/2}} \tag{1}$$

式中:I_u 表示用户 u 作过评分的项目集合;$\bar{r}_u$ 表示用户 u 所有评分的均值。u 的 k-近邻就是与 u 相似度最高的 k 个用户的集合,记作 $N_{k,u}$,并用 U_i 表示对项目 i 作过评分的用户集合,则用户 u 对项目 i 的评分预测为

$$\hat{r}_{ui}=\bar{r}_u+\frac{\sum\limits_{v\in N_{k,u}\cap U_i}w_{uv}(r_{vi}-\bar{r}_v)}{\sum\limits_{v\in N_{k,u}\cap U_i}w_{uv}} \tag{2}$$

这就是通常所说的 kNN 估计。以用户和项目为两个维度的评分矩阵也称为效用矩阵,评分预测问题可以看作是填充效用矩阵中的空白元素。取得评分预测后,便可将用户最可能感兴趣的项目推荐给用户,一般采用 Top-N 推荐[11-12],即把用户 u 评分估计最高的 N 个项目推荐给该用户。

推荐系统可以通过离线测试来评价推荐效果,将收集到的用户行为数据集分割为正交的训练集和测试集,使用训练集的数据计算填充效用矩阵中的空白元素,然后用测试集的数据作为真实用户行为来检测推荐的效果。推荐的评价指标通常有准确性指标和覆盖率指标。准确性指标有准确率和召回率,准确率反映了推荐的项目是用户真正感兴趣的项目的比率,记作 $p_{\mathrm{P}}=(\sum\limits_u|R_u\cap T_u|)/\sum\limits_u|R_u|$,召

回率则反映了用户感兴趣的项目被推荐出来的比率，记作 $p_{\mathrm{R}}=(\sum_u|R_u\cap T_u|)/\sum_u|T_u|$。

上面式子中的累加运算表示对用户全集中的所有用户 u 进行计算，R_u 表示推荐给用户 u 的项目集合，T_u 表示测试集中用户 u 做出评分的项目集合，$|A|$ 表示集合 A 中元素的个数。

此外，统计学中还使用 F_1 分数（记作 F_1）来兼顾准确率和召回率，作为综合准确性指标，$F_1=2p_{\mathrm{P}}p_{\mathrm{R}}/(p_{\mathrm{P}}+p_{\mathrm{R}})$。

覆盖率指标反映推荐系统对长尾项目的挖掘能力，也就是考察推荐物品的分布，这个分布越平均，则长尾挖掘能力越好，覆盖率越高；反之，若分布越陡峭，则推荐集中于部分物品，长尾挖掘能力差，覆盖率低。覆盖率指标可以用比较粗略的覆盖率来描述，记作 $c_{\mathrm{C}}=|\bigcup R_u|/|I|$。

另一方面，由于信息熵（记作 c_{E}）能够很好地反映信息的混乱程度，即推荐的物品是否足够平均，故也可作为评估覆盖率的良好指标。记 p_i 表示项目 i 的流行度，即项目 i 被推荐的次数除以所有项目被推荐次数之和，则有 $c_{\mathrm{E}}=-\sum_i p_i\log p_i$，这里的累加是对项目全集中的所有项目进行。

1.2　基于线性插值法的加权百分比的推荐方法

由式(2)可见，通常的基于邻域的协同过滤方法在计算用户 u 对项目 i 评分估计时，实际上是把近邻集合 $N_{k,u}$ 中的每一个用户 v 对 i 的评分，按照该用户与 u 的相似度进行加权平均。也就是说，若把每个近邻 v 与 u 的相似度占所有近邻相似度总和的比值当成 u 对 i 的评分可能等于的概率，则 u 对 i 的评分估计取值为所有近邻对 i 的评分期望。

从另一方面看，推荐的基本原理是将用户最可能感兴趣的项目推荐给用户，基于这种考虑，加权百分比的推荐方法[8]不采用上述的期望计算法来评估用户对项目的兴趣，而是提出了一种高概率百分比的推荐，通过评估用户会以高概率（大于 50%）对项目感兴趣的程度来实现推荐。

算法 1　加权百分比估计。

输入　近邻用户的已知评分 $r_1,r_2,\cdots,r_k$ 和对应的近邻与用户 u 之间的相似度权重 $w_1,w_2,\cdots,w_k$，且有 $\sum_{j=1}^{k}w_j=1$，百分比为 p。

输出　p-加权百分比估计值 $\hat{r}^p$。

步骤 1　将 $r_1,r_2,\cdots,r_k$ 按从小到大排序，并对应变化 $w_1,w_2,\cdots,w_k$ 的顺序，仍然记为 $w_1,w_2,\cdots,w_k$；

步骤 2　找出区间序号 n_p，使得 $a=\sum_{j=1}^{n_p-1}w_j\leqslant p\leqslant\sum_{j=1}^{n_p}w_j=b$；

步骤 3　使用线性插值法计算 $\hat{r}^p=(r_{n_p}-r_{n_p-1})(p-a)/(b-a)+r_{n_p-1}$。

最后，p-加权百分比估计值 $\hat{r}^p$ 将作为评估用户对项目的兴趣度指标来计算推荐项目。

2　邻域线性拟合的推荐置信度模型

2.1　推荐支持度模型

在加权百分比推荐方法的基础上，本文完整地提出了推荐支持度模型来推荐更趋向于给出用户最感兴趣的项目。

定义 1　关于用户 u、项目 i 和评分 r 的推荐支持度函数 $R(u,i,r)$，定义其值为 u 对 i 的评分最高可达到 r 的概率。对于给定概率 p，若有 $\hat{r}_{ui}^p$ 使得

$$R(u,i,\hat{r}_{ui}^p)=p \tag{3}$$

则称 $\hat{r}_{ui}^p$ 为满足推荐支持度为 p 的条件下，用户 u 对项目 i 的 p-支持度评分。

推荐支持度模型下的推荐计算问题，由传统的式(2)的评分估计计算转化为计算给定用户 u 对所有项目的 p-支持度评分 $\hat{r}_{ui}^p$，然后选取 $\hat{r}_{ui}^p$ 最高的 N 个项目加入推荐列表，从而实现 Top-N 推荐。

推荐支持度模型中，一般选取 p 为大概率数值，从而描述了用户可能更趋向于喜欢某项目的程度，大大提高了推荐的准确性指标。

2.2　邻域线性拟合算法

由前述可知，推荐支持度模型的计算关键在于如何求解 p-支持度评分 $\hat{r}_{ui}^p$。考察 u 的邻域 $N_{k,u}$，把所有邻域用户的评分及其与 u 的相似度组成的元组 (r_{vi},w_{uv})，$v\in N_{k,u}$，按 r_{vi} 从小到大排列，记为 $(r_1,w'_1),(r_2,w'_2),\cdots,(r_k,w'_k)$，为便于计算，我们再将 w'_j 进行归一化处理，即令 $w_j=w'_j/\sum_{l=1}^{k}w'_l$，$j=1,2,\cdots,k$，则得到序列 $(r_1,w_1),\cdots,(r_k,w_k)$，且有 $r_1\leqslant r_2\leqslant\cdots\leqslant r_k$，$\sum_{j=1}^{k}w_j=1$。

由式(3)易知 $R(u,i,r_j)=\sum_{l=1}^{j}w_l$，因此算法 1 可以通过分段线性插值法来计算 $\hat{r}_{ui}^p$。但是，由于用户的评分是主观行为，实际上扰动很大，而且当数据稀疏时，对相似度的影响则更大，分段线性插值法易受数据扰动影响。

本文提出采用邻域线性最小二乘拟合的方法实现数据滤波,减小扰动影响。

算法 2 邻域线性最小二乘拟合估计。

输入 $(r_{vi},w_{uv}),v\in N_{k,u}$,推荐支持度 p。

输出 p-支持度评分 $\hat{r}_{ui}^{p}$。

步骤 1 对输入元组集合排序及归一化处理后得到序列$(r_1,w_1),(r_2,w_2)\cdots,(r_k,w_k)$;

步骤 2 记 $x_j=\sum_{l=1}^{j}w_l$,$y_l=r_j$,$j=1,2,\cdots,k$,向量 $\boldsymbol{x}=(x_1,\cdots,x_k)$,$\boldsymbol{y}=(y_1,\cdots,y_k)$,利用最小二乘法求解方程 $\boldsymbol{y}=a\boldsymbol{x}+b$ 的参数 a 和 b;

步骤 3 计算 $\hat{r}_{ui}^{p}=ap+b$。

下面,通过一个例子来解释 p-支持度评分估计的方法。设用户 u 的邻域大小为 5,邻居们与 u 的相似度分别为 0.05,0.075,0.1,0.2,0.075,且已知他们对项目 A 的评分为 4,4,5,6,10,按评分排序并对相似度归一化处理后得到元组序列为{(4,0.1),(4,0.15),(5,0.2),(6,0.4),(10,0.15)},把这些元组标识的点绘制在坐标轴上,如图 1 所示,可以得到线性插值法(Interpolation,IP)的折线和邻域最小二乘线性拟合法(linear least squares fitting,LLSF)的拟合直线。由图可见,$p=0.8$ 时的 p-支持度评分估计是:线性插值法为 5.875,而邻域最小二乘线性拟合法计算得到 7.339。

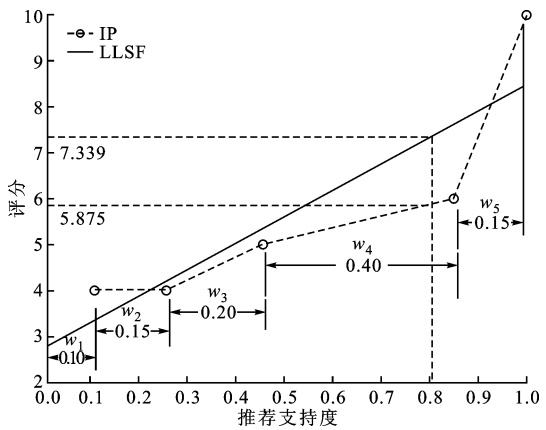


图 1 p-支持度评分估计示例图

结合图 2 进一步举例分析推荐支持度模型下推荐问题的解决方法。假设前述 u 的邻居们对项目 B 的评分为 1,8,3,6,9,则排序后的元组序列为{(1,0.1),(3,0.2),(6,0.4),(8,0.15),(0,0.15)}。若以常规 kNN 估计的期望估计法,可计算得 u 对 A 的评分估计为 $\hat{r}_{uA}=5.9$,而 u 对 B 的评分估计为 $\hat{r}_{uB}=5.65$,故系统应将项目 A 推荐给用户 u,但若以 p-支持度评分估计来计算,使用线性插值法和最小二乘法会得到各自不同的推荐结果。

由图 2 可见,A 评分始终高于 B 评分。分段插值法下,p_1 和 p_3 为 A 评分与 B 评分的交点,推荐支持度选在区间(p_1,p_3)内时,B 评分大于 A 评分,将会推荐 B,其他推荐支持度时则推荐 A。邻域最小二乘法计算 p_2 为 A 评分与 B 评分的交点,当推荐支持度选择大于 p_2 的值时,B 评分大于 A 评分,系统应推荐 B,否则推荐 A。

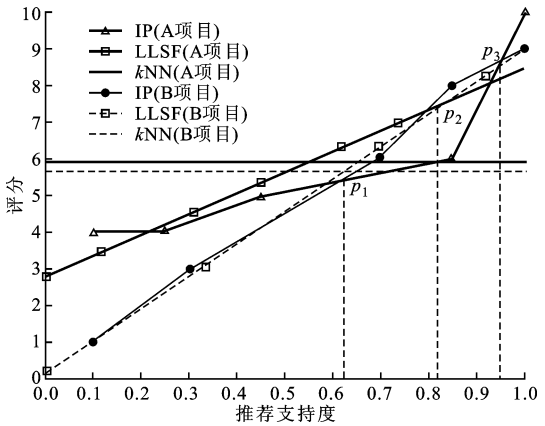


图 2 推荐支持度模型示例图

2.3 针对稀疏数据集的分析

下面,分析邻域线性拟合算法对于处理稀疏数据集的优势。数据的稀疏度对推荐准确率有很大影响的主要原因在于,稀疏的数据集使得两用户之间的共同评分项目变得很少。由式(1)可知,两用户的相似度 w_{uv} 是通过其共同评分的项目 $I_u\bigcap I_v$ 计算的,所以当共同评分的项目越少时,相似度计算受扰动的影响就越大。我们称共同评分项目很少时计算出来的相似度为不可信相似度。例如,若两个用户只有一部共同评分的电影,而他们在这部电影上恰好评分相同(若使用皮尔逊相似度,确切地说,应该是相对平均值的评分差相同),则两个用户具有很大的相似度。实际上,很可能这两个用户兴趣根本不同(从他们很少评价同一部电影就可以看出)。不可信相似度用户的评分可能会对推荐结果造成很大的误导,形成错误的推荐。

从以上论述可知,处理稀疏数据集时抗扰动能力强的推荐算法,应该降低推荐评分 $\hat{r}$ 受到单个不可信相似度用户评分的影响。为便于计算,下面对 k 近邻按归一化后的相似度 $w_1,w_2,\cdots,w_k$ 进行分析,并设某个 $w_m,m\in[1,k]$为不可信相似度。

首先,分析线性插值法的稀疏集抗扰动能力。根据算法 1,当百分比 p 落在 w_m 所在区间内时,易知推荐评分 $\hat{r}^{p}=c(r_m-r_{m-1})+r_{m-1}$,这里 c 为[0,1]

区间上的常数，即此时 $\hat{r}^p$ 完全由 r_m 和 r_{m-1} 决定，受不可信相似度用户的评分 r_m 影响很大，故易受到稀疏数据集影响。

其次，分析 kNN 算法的稀疏集表现。为简化计算，以余弦相似度[1-5]作分析，并注意到 $\sum w_j=1$，则有 $\hat{r}=\sum_{j=1}^{k}w_jr_j=\sum_{j\neq m}^{k}w_jr_j+w_mr_m$，显然，推荐评分 $\hat{r}$ 受 r_m 的影响，取决于 w_m 的大小，当 w_m 很大时，kNN 算法受扰动的影响就大。

最后，分析邻域线性拟合法的抗扰动能力。根据算法 2 步骤 2，可以计算出

$$a=(k\sum x_jr_j-\sum x_j\sum r_j)\Big/(k\sum x_j^2-(\sum x_j)^2) \tag{4}$$

$$b=(\sum x_j^2\sum r_j-\sum x_j\sum x_jr_j)\Big/(k\sum x_j^2-(\sum x_j)^2) \tag{5}$$

因此有

$$\hat{r}=(kp\sum x_jr_j-p\sum x_j\sum r_j+\sum x_j^2\sum r_j-\sum x_j\sum x_jr_j)\Big/(k\sum x_j^2-(\sum x_j)^2) \tag{6}$$

其中与 r_m 有关的因式是

$$r_m(kpx_m-p\sum x_j+\sum x_j^2-x_m\sum x_j)\Big/(k\sum x_j^2-(\sum x_j)^2) \tag{7}$$

其系数受到更多因素的制约，故抗干扰能力更强。

2.4　模型的融合与拓展

当前推荐系统已发展出大量的模型与算法，除了本文应用到的协同过滤推荐，还有基于内容的推荐、基于模型的推荐[2]等，每一大类又有很多的优化方法。每种模型各有其优缺点，在实际应用中，通常是融合多种模型来提高推荐的性能。

本文提出的邻域线性拟合算法同样易于通过与其他模型及优化算法进行融合来提升性能。例如，可以应用以下一些模型融合方法。①相似度优化。首先利用各种相似度优化方法对相似度计算进行优化，再利用算法 2 计算推荐项目，通过提高相似度的准确性来提升推荐效果。②模型级联。可以首先使用基于内容的模型或者隐语义模型等其他的推荐模型进行效用矩阵填充，减少其稀疏度，再在此基础上应用邻域线性拟合算法，计算最终的推荐项目。③模型组合。可以使用线性加权模型组合，例如有另一种模型计算出的推荐评分 $\hat{r}'_{ui}$，以及邻域线性拟合模型计算的 p-支持度评分 $\hat{r}^p_{ui}$，并记 $n(r_{ui})=r_{ui}/\sum_j r_{uj}$ 表示评分归一化函数，则组合后的评分为 $\hat{r}_{ui}=R(cn(\hat{r}'_{ui})+(1-c)n(\hat{r}^p_{ui}))$，这里 c 为(0,1)上的常数，R 为评分区间的上限。最后，使用 $\hat{r}_{ui}$ 来计算推荐项目。

3　实验评估

3.1　实验方案

分别使用 MovieLens 数据集和 Book-Crossing 数据集对本文提出的邻域最小二乘线性拟合推荐支持度模型进行 Top-N 推荐离线测试，评估其对推荐准确度的提升效果，并同时考查推荐覆盖率指标的满足情况。

为防止过拟合，实验采用以下步骤来进行。

步骤 1　将数据集随机分成 M 份，第 1 份作为测试集，另外 M-1 份作为训练集；

步骤 2　使用训练集来训练模型，使用测试集来检测得到待评估的指标；

步骤 3　更换随机数种子，返回步骤 1 再次开始，共重复 M 次；

步骤 4　把 M 次实验计算的指标值进行平均，得到最后的指标评估。

本文中的实验选取 $M=5$，并且使用传统 kNN 算法及前面介绍的线性插值算法进行数据对比研究。

3.2　MovieLens 数据集实验

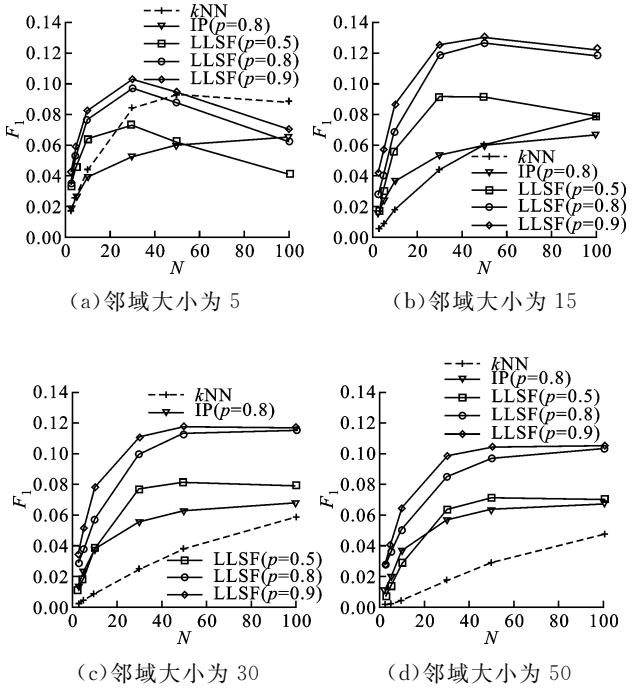
MovieLens 数据集[13]（简写为 ML）是由明尼苏达大学 GroupLens 研究小组提供的电影评分数据集，本文使用其大小为 10 万条记录的数据集进行离线测试，数据稀疏度为 93.7%。

首先测试了不同算法在不同邻域大小时对推荐准确率的影响，并使用 F_1 分数来评估准确率。实验结果如图 3 所示，图 3 的 4 个子图分别给出了邻域 k 选取 5、15、30、50 时的推荐结果。推荐集大小 N 作为坐标横轴，实验计算了 N 取 3、5、10、30、50、100 时的推荐 F_1 分数，F_1 分数作为坐标纵轴，实验对比了 kNN 算法、分段线性插值(IP)的加权百分比算法(选取 $p=0.8$)、邻域最小二乘线性拟合(LLSF)的推荐支持度算法(分别选取 $p=0.5,0.8,0.9$)。

由图 3 可知，邻域 k 为 15、30、50 时，IP 算法与 LLSF 算法的推荐准确率都优于 kNN 算法，LLSF 算法又显著优于 IP 算法。对于 LLSF 算法来说，$p=0.9$ 时效果最佳，特别地，$k=15$ 且 $N=50$ 时，取得最高推荐准确率。当 $k=5$ 时，IP 算法准确率反而低于 kNN 算法，当 N 较小时 LLSF 算法仍然优于 kNN 算法，只有当 N 较大时推荐效果才比 kNN

算法差。

取 $k=30$、$N=50$ 为考察维度，不同算法的准确率指标和覆盖率指标结果如表 1 所示。IP 算法虽然准确度(准确率和召回率)高于 kNN 算法，但是覆盖率指标(覆盖率和信息熵)偏低，而 LLSF 算法不但准确率明显优于 kNN 和线性插值算法，覆盖率也没有很大损失，可以保证对长尾项目的挖掘能力。



(a)邻域大小为 5 (b)邻域大小为 15

(c)邻域大小为 30 (d)邻域大小为 50

图 3 ML 数据集不同邻域大小时各算法的 F_1 分数

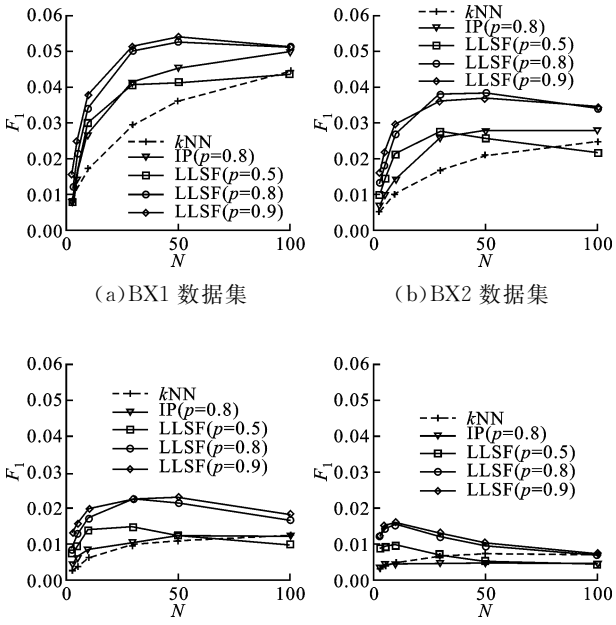
表 1 ML 数据集各算法推荐指标分析

算法	准确率	召回率	F_1 分数	覆盖率	信息熵
kNN	0.027	0.063	0.038	0.714	9.47
IP ($p=0.8$)	0.044	0.104	0.062	0.307	6.68
LLSF ($p=0.5$)	0.057	0.136	0.081	0.620	8.80
LLSF ($p=0.8$)	0.080	0.189	0.113	0.539	8.50
LLSF ($p=0.9$)	0.083	0.197	0.117	0.527	8.49

3.3 Book-Crossing 数据集实验

Book-Crossing 数据集[14](简写为 BX)是由 Ziegler 等爬取 www.bookcrossing.com 社区获得的书籍评分数据集。本文对原始数据集进行处理，剔除评分数过少的用户后，形成不同稀疏度的 4 个数据集，数据稀疏度分别是 BX1 为 90.9%，BX2 为 95.2%，BX3 为 98.2%，BX4 为 99.3%。

图 4 给出了对不同稀疏度的 BX 数据集进行离线实验，几种算法在选取邻域 $k=30$ 时所得到的推荐准确度指标 F_1 分数与推荐数量 N 的关系。由图可见，在不同稀疏度条件下，大支持度($p=0.8$，0.9)的 LLSF 算法推荐准确度都优于 kNN 和 IP 算法。IP 算法虽然在稀疏度较低时推荐准确度较好，但当数据稀疏度上升时，准确度迅速下降，甚至低于基本的 kNN 算法。



(a)BX1 数据集 (b)BX2 数据集

(c)BX3 数据集 (d)BX4 数据集

图 4 BX 数据集不同稀疏度时各算法的 F_1 分数

不同稀疏度 BX 数据集实验中各算法的推荐效果比较见表 2。对于准确率指标，N_0 为使得 F_1 分数达到最大值的推荐数量，$F_1(N_0)$为当前算法的最

表 2 不同稀疏度 BX 数据集各算法推荐指标分析

数据集	算法	N_0	$F_1(N_0)$	F_1 优化率	信息熵
BX1	kNN	100	0.044 4	—	8.46
	IP($p=0.8$)	100	0.050 0	12.6%	5.54
	LLSF($p=0.9$)	50	0.054 0	21.6%	8.02
BX2	kNN	100	0.024 7	—	9.04
	IP($p=0.8$)	100	0.027 9	13.0%	5.39
	LLSF($p=0.9$)	50	0.036 9	49.4%	8.78
BX3	kNN	100	0.012 2	—	9.58
	IP($p=0.8$)	50	0.012 1	−0.8%	5.42
	LLSF($p=0.9$)	50	0.022 6	85.2%	9.10
BX4	kNN	50	0.007 1	—	9.80
	IP($p=0.8$)	50	0.004 6	−35.2%	5.72
	LLSF($p=0.9$)	10	0.016 0	125.4%	9.07

大准确率，即 Top-N_0 推荐的 F_1 分数，F_1 优化率为当前算法对比 kNN 算法的最大准确率提升的百分比；对于覆盖率指标，信息熵为当前算法对于 N 取 3、5、10、30、50、100 时的信息熵均值。

由表 2 可见，随着数据集稀疏度的增大，LLSF 算法所提供的准确率改进就越大。同时，LLSF 算法的覆盖率指标牺牲很小，大大优于 IP 算法的覆盖率，可以满足长尾项目的挖掘要求。

4　结　论

尽管推荐系统主要有预测和推荐两类应用，但 Top-N 推荐是其实际中最广泛的应用方式[1]。然而，目前绝大部分的研究还是集中在预测领域，针对提升推荐准确率的研究还较少。本文所描述的邻域最小二乘线性拟合的推荐支持度模型可以大幅度提升推荐准确率，特别是对于稀疏数据集效果更显著，同时也没有牺牲推荐的覆盖率，保证了对长尾项目的挖掘能力。

本文所描述的方法还可以与其他模型融合以进一步提升推荐效果，具有良好的算法拓展能力。下一步，我们将具体研究本文方法与其他模型融合时涉及的模型选择、参数选择和实验效果等问题，特别是融合基于内容的推荐，从而更好地处理冷启动问题，提升算法的实际应用性能。

参考文献：

[1] 项亮. 推荐系统实践 [M]. 北京：人民邮电出版社，2012：1-34.

[2] 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.

[3] LINDEN G, SMITH B, YORK J. Amazon. com recommendations: item-to-item collaborative filtering [J]. IEEE Internet Computing, 2003, 7(1): 76-80.

[4] SU Xiaoyuan, KHOSHGOFTAAR T M. A survey of collaborative filtering techniques [J]. Advances in Artificial Intelligence, 2009, 2009: 421425.

[5] SARWAR B, KARYPIS G, KONSTAN J, et al. Item-based collaborative filtering recommendation algorithms [C]// Proceedings of the 10th International Conference on World Wide Web. New York, USA: ACM, 2001: 285-295.

[6] 黄创光，印鉴，汪静，等. 不确定近邻的协同过滤推荐算法 [J]. 计算机学报，2010，33(8)：1369-1377.
HUANG Chuangguang, YIN Jian, WANG Jing, et al. Uncertain neighbors' collaborative filtering recommendation algorithm [J]. Chinese Journal of Computers, 2010, 33(8): 1369-1377.

[7] 罗辛，欧阳元新，熊璋，等. 通过相似度支持度优化基于 K 近邻的协同过滤算法 [J]. 计算机学报，2010，33(8)：1437-1445.
LUO Xin, OUYANG Yuanxin, XIONG Zhang, et al. The effect of similarity support in K-nearest-neighborhood based collaborative filtering [J]. Chinese Journal of Computers, 2010, 33(8): 1437-1445.

[8] ADAMOPOULOS P, TUZHILIN A. Recommendation opportunities: improving item prediction using weighted percentile methods in collaborative filtering systems [C]// Proceedings of the 7th ACM Conference on Recommender Systems. New York, USA: ACM, 2013: 351-354.

[9] 吕红亮，王劲林，邓峰，等. 多指标推荐的全局邻域模型 [J]. 西安交通大学学报，2012，46(11)：98-105.
LÜ Hongliang, WANG Jinlin, DENG Feng, et al. A global neighborhood-based model with multi-criteria recommendation [J]. Journal of Xi'an Jiaotong University, 2012, 46(11): 98-105.

[10] KOREN Y. Factor in the neighbors: scalable and accurate collaborative filtering [J]. ACM Transactions on Knowledge Discovery from Data, 2010, 4(1): 1-24.

[11] DESHPANDE M, KARYPIS G. Item-based top-N recommendation algorithms [J]. ACM Transactions on Information Systems, 2004, 22(1): 143-177.

[12] ADAMOPOULOS P. On discovering non-obvious recommendations: using unexpectedness and neighborhood selection methods in collaborative filtering systems [C]// Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York, USA: ACM, 2014: 655-660.

[13] GROUPLENS. MovieLens datasets [DB/OL]. (2011-03-01) [2014-06-20]. http://www.grouplens.org/datasets/movielens/.

[14] ZIEGLER C N, FREIBURG D. Book-crossing datasets [DB/OL]. (2004-06-01) [2014-06-20]. http://www2.informatik.uni-freiburg.de/~cziegler/BX/.

（编辑　武红江）