基于稀疏数据预处理的协同过滤推荐算法

陈宗言[1,2],颜 俊[1,2]
(1. 南京邮电大学 通信与信息工程学院,江苏 南京 210003;
2. 宽带无线通信与传感网技术教育部重点实验室,江苏 南京 210003)

摘 要:随着推荐系统规模的不断扩大,用户-项目评分矩阵呈现出极端稀疏性,导致基于传统相似性度量方法的协同过滤推荐系统推荐质量的下降。针对该问题,文中提出了一种基于项目特征属性的稀疏数据集预处理方法来提高算法的推荐质量。首先,通过引入项目的特征属性信息,根据项目间特征属性相似度,初步预测用户对未评分项目的评分,可以使得用户-项目评分矩阵完全饱和。接着再对稀疏数据集的未评分项目进行混合填充预处理,避免了传统均值填充法中的用户对项目的评分不可能完全相同的问题以及众数填充法中的"多众数"和"无众数"问题。实验结果表明,文中提出的方法更能有效地提高推荐系统的推荐质量和推荐覆盖率。

关键词:推荐系统;协同过滤;特征属性;稀疏数据集;混合填充
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2016)07-0059-06
doi:10.3969/j.issn.1673-629X.2016.07.013

Collaborative Filtering Recommendation Algorithm Based on Sparse Data Pre-processing

CHEN Zong-yan[1,2], YAN Jun[1,2]
(1. College of Communication & Information Engineering, Nanjing University of Posts & Telecommunications, Nanjing 210003, China;
2. Key Lab of Broadband Wireless Communication and Sensor Network Technology of Ministry of Education, Nanjing 210003, China)

Abstract: With the continuous expansion of recommender systems, the sparsity of the user-item matrix can deteriorate the performance of the traditional similarity calculation based collaborative filtering recommendation approaches. In order to overcome this drawback, a new sparse data pre-processing algorithm based on item feature is proposed to mitigate this effect. First, considering the item characteristics information, the ratings of the unrated items are predicted through the similarities between each item. It can lead to saturated matrix and overcome the drawback of the sparsity matrix. Next, the hybrid filling method is utilized to process the unrated items in the sparse data sets, which can avoid the problem of full no consistency of different items for traditional mean-filling method and the multiple mode and no mode for the mode-filling approaches. The simulation demonstrates that the proposed algorithm can improve the recommended quality and coverage dramatically.

Key words: recommender system; collaborative filtering; item characteristics; sparse data set; hybrid filling

1 概 述

随着信息技术和互联网的发展,人们逐渐从信息匮乏的时代步入了信息过剩的时代。在这个时代里,无论是信息的生产者或信息的消费者都面临着巨大挑战:作为信息生产者,如何让自己的信息脱颖而出,受到广大用户的关注,是一件非常困难的事情;作为信息消费者,如何从海量的信息中找到自己感兴趣的信息也是一件非常复杂的事情。因此,推荐系统成为互联网技术研究的一个热点。

推荐系统的任务就是联系用户和信息,一方面帮助用户发现对自己有价值的信息,另一方面让信息能够展现在对它感兴趣的用户面前[1]。

为使推荐系统产生精确的推荐,保证推荐系统的实时性和有效性,研究人员提出了各式各样的推荐算

收稿日期:2015-10-24 **修回日期**:2016-01-27 **网络出版时间**:2016-06-22
基金项目:国家自然科学基金资助项目(61372122)
作者简介:陈宗言(1991-),男,硕士研究生,研究方向为大数据挖掘;颜 俊,副教授,博士,研究方向为数据挖掘、压缩感知技术及其应用。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160622.0844.026.html

法。文献[2-4]提出了通过对用户(项目)进行聚类分析来进行个性化推荐的方法,该方法可以有效提高推荐系统的响应速度,但其推荐精度并没有显著改善;文献[5]通过对用户购买的商品进行关联分析,发掘商品间关联度并产生推荐,但在关联过程中会产生较多的候选项目集,输入输出负载较大;文献[6]提出了分类算法,通过对用户(项目)进行分类,进而产生分类推荐,可以有效提高推荐的准确率;文献[7-8]提出了协同过滤推荐算法,通过对用户的历史兴趣进行分析,根据得到的预测结果为用户提供推荐,显著提高了推荐的精确性,并且其算法模型简单、数据采集方便。

鉴于协同过滤算法在推荐系统方面体现出的巨大优势,文中重点讨论基于项目(Item-Based)的协同过滤推荐算法[9]。

协同过滤推荐主要有两大类:基于模型的方法[10]和基于记忆的方法[11]。基于模型的方法通过系统已有的用户-商品信息学习并产生一个模型,从而根据得到的模型进行预测推荐;基于记忆的方法主要通过计算用户(项目)间的相关性,根据算法所设定的邻居个数,寻找目标用户(项目)的最近邻,这些最近邻居可由目标项目与其他项目相似值从高到低排列所得,并通过最近邻为用户(项目)进行推荐。

迄今为止,协同过滤技术在电子商务发展中得到了广泛应用。以亚马逊为代表的电子商务公司通过对目标用户的历史购买记录进行分析,进而产生个性化推荐。据 VentureBeat 统计,亚马逊公司每年至少有 35% 的销售来自于推荐算法。著名视频网站 Netflix 通过对用户的历史观看记录的分析,得到用户的个人观影喜好,进而为用户推荐相似的电影。

但是,随着推荐系统规模的不断扩大,用户评分数据呈现出极端的稀疏性,比如:在大型商务系统中,用户评分的项目一般不会超过项目总和的 1%[12]。研究发现,协同过滤技术在对由用户历史信息得到的用户-项目评分矩阵进行用户(项目)相似度计算时,得到的结果不能让人满意。比如:一个用户由于是新用户或者其做出评分的项目过少,可能会导致该用户和其他用户之间的相似度无法计算,从而不能做出有效推荐,导致推荐算法精确度的下降。

针对上述问题,文献[13-14]提出了相似性度量方法的优化算法,但由于其在计算项目(用户)相似性时仍然是基于稀疏数据集,所以性能仍然提高不大。因此,在大型商务系统中,基于稀疏数据集的协同过滤推荐算法的相似度计算已经成为制约推荐系统性能的一个关键因素,如何在计算相似度之前对稀疏的数据集进行合适的处理则成为提高推荐效率的一个关键。

为此,文中提出了一种基于项目特征属性的数据预处理技术。主要贡献有以下两点:

(1)通过引入项目的特征属性信息,根据项目间特征属性相似度,初步预测用户对未评分项目的评分,再对稀疏数据集的未评分项目进行混合填充预处理,可以使得用户-项目评分矩阵完全饱和。

(2)有效避免了文献[15]所提出的均值填充法中的用户对不同项目评分不可能完全一致以及众数填充法中“多众数”和“无众数”的问题。

实验结果表明,相对于对稀疏数据集的缺失项不进行填充和使用“众数填充”的方法,文中所提方法更能有效地提高推荐系统的推荐质量和推荐的覆盖率。

2 相关工作

目前,学术界对推荐算法的相似性度量和稀疏数据集的预处理已经开展了很多研究工作。

相似性的度量方法主要有三种:Pearson 相关相似性、余弦相似性以及修正的余弦相似性。鉴于修正的余弦相似性原理与余弦相似性相同,只是公式上的略微变化,这里不做详细介绍。前两种方法的计算公式如下:

Pearson 相关相似性:设 U_{ij} 是项目 i , j 共同的用户集合,则项目 i 和项目 j 之间的相似性 $\mathrm{sim}(i,j)$ 通过 Pearson 相关系数度量:

$$\mathrm{sim}(i,j)=\frac{\sum_{c\in U_{ij}}(R_{ic}-\overline{R_i})(R_{jc}-\overline{R_j})}{\sqrt{\sum_{c\in U_{ij}}(R_{ic}-\overline{R_i})^2}\sqrt{\sum_{c\in U_{ij}}(R_{jc}-\overline{R_j})^2}} \tag{1}$$

其中, R_{ic} , R_{jc} 是用户 c 对项目 i 和项目 j 的评分; $\overline{R_i}$, $\overline{R_j}$ 是项目 i , j 的平均得分。

余弦相似性:项目所得评分被看做是 n 维用户空间上的向量,如果用户对项目没有进行评分,则将该用户对项目的评分设为 0,项目间的相似性通过向量间的余弦夹角度量:

$$\mathrm{sim}(i,j)=\frac{\vec{i}*\vec{j}}{\|\vec{i}\|*\|\vec{j}\|} \tag{2}$$

其中,向量 $\vec{i}$, $\vec{j}$ 是项目 i , j 在 n 维空间上的评分。

这两种相似性方法可以有效计算用户(项目)间的相似度,但如前文所述,如果一个用户由于是新用户或者其做出评分的项目过少,使得该用户和其他用户之间没有共同的项,导致该用户和其他用户之间的相似度无法计算。因此,针对这种由用户-项目评分矩阵极度稀疏引起的问题,研究人员又提出了各式各样对稀疏数据集进行预处理的方法。

文献[16-18]提出通过奇异值分解(SVD)算法估

计未评分项目的评分来填充用户–项目矩阵,奇异值分解在用户–项目矩阵较为稠密的情况下会取得较高的准确度,在数据极度稀疏情况下,预测效果并不是很理想。文献[19]提出对用户未评分项目进行均值填充(mean filling)及众数填充(mode filling)等方法,但是这种赋予固定值的方法并不是很可信。比如采用众数法时,利用一组数据中出现频率最高的数来填充缺失值,会出现“多众数”和“无众数”的问题。采用均值法时,对用户未评分的项目赋予项目或用户的评分均值,同样也存在用户对项目的评分不可能完全相同的问题。更为重要的是,采用现有的相似度计算方法对此填充矩阵进行项目(用户)相似度计算时,所得结果的可信度不高。

用户–项目矩阵如表 1 所示。

表 1　用户–项目矩阵

	$item_1$	$item_2$	$item_3$	$item_4$
$user_1$	3	/	4	2
$user_2$	/	3	/	/
$user_3$	2	3	/	/

如果对缺失项不进行填充或者采用平均值填充的方法,在使用相关相似度计算公式计算项目间相关性时(如式(1)所示),会发现 $item_2$ 和其他项目的相关性无法计算(分子分母同时为 0)。

在使用余弦相似性计算方法时(如式(2)所示),将用户未评分的项目假设为 0,会发现 $item_3$ 与 $item_4$ 相似度恒为 1,然而 $item_3$ 与 $item_4$ 的实际所得评分差距还是很大的。

所以,基于以上两种的填充方法都不能得到令人满意的效果。

3　文中方法

3.1　方法思路

针对现有的预处理方法在处理稀疏数据集时往往只是利用已有的用户评分数据进行简单的填充,却没有利用数据自身特征属性的问题,文中所提方法充分利用特征属性隐含的潜在价值,从相似度计算和填充方法这两个角度出发,提出了基于项目特征属性的稀疏数据预处理的混合填充预处理方法,具体包括项目特征属性相似度计算和数据混合填充两个步骤。

在任何商务领域,任意项目往往可以用若干个特征属性来表示:电影可以通过战争、喜剧、文艺等属性进行分类;商品则可以通过电器、食物、书籍等属性进行分类。这些属性信息可以从项目所属的网页或者推荐系统中用来记录项目信息的数据集里抽取得到。

通过特征属性计算项目间的相似度,选出和目标用户已评价或者感兴趣的项目的相似度较高的未评分项目进行预测推荐等。通过特征属性得到的项目间相似度往往比通过由评分矩阵得到的可信度更高:假如有 $item_1$ 和 $item_2$ 两个评分及属性集合,$item_1$ 的评分集合为 {1,2,1,2,1},特征属性集合为 {1,0,1,0,1};$item_2$ 的评分集合为 {4,5,4,5,4},特征属性集合为 {1,0,1,1,1}。其中,评分范围为 1 到 5 分,特征属性集合中的 1 代表 item 具有某个属性,0 代表没有,一共 5 个不同属性。在使用式(2)进行相似度计算时可以看到,由 $item_1$ 和 $item_2$ 的评分计算而来的 item 相似度为 1,由特征属性计算得到的 item 相似度为 0.86,根据 $item_1$ 和 $item_2$ 的评分集合可以看到两者所得评分差距较大,$item_1$ 的评分都低于 3 分,$item_2$ 的评分都是 3 分以上,两者的相似度不可能完全一样。相对而言,由特征属性计算得来的相似度值可信度较为可信,从而得到的未评分项目的预测值更为准确。

对此,为充分利用项目的特征属性以及避免均值填充法和众数填充法所带来的问题,文中提出了一种新的数据填充方法,即结合项目自身的特征属性,利用从这些属性计算中所获得的未评分项目的预测值结合项目均值对稀疏矩阵中未评分项进行混合填充(hybrid filling)。实验结果表明,该混合填充方法无论是在预测精度或者推荐项目的覆盖率方面,相对上述两种方法,都有明显的改善。

3.2　方法描述及流程

方法步骤如下所示:

(1)根据提取到的项目特征属性信息,建立如表 2 的项目–特征属性矩阵 I–F。

(2)对不同的特征属性赋予不同的权值,利用式(2)计算项目间特征属性相似度,得到项目属性相似度矩阵 sim_f。

假设项目特征属性矩阵如表 2 所示。其中,0 表示项目不具有某项属性,1 表示项目具有某项属性。

表 2　项目特征属性矩阵

	f_1	f_2	f_3	f_4	f_5
$item_1$	0	1	1	1	1
$item_2$	1	1	1	0	0
$item_3$	0	1	1	0	0
$item_4$	0	0	0	1	1
$item_5$	0	1	1	0	0

从表中可以看到,$item_4$ 与 $item_1$ 含有两个共同属性,为 f_4,f_5;$item_5$ 与 $item_1$ 也包含两个共同属性,分别是 f_2,f_3。通过式(2)计算 $item_4$、$item_5$ 与 $item_1$ 的相似度,发现是相等的。但是从表 2 中可以看到,项目标签属性的频率是不一样的,f_2,f_3 的标签个数完全多于其他

的标签。可以认为大部分项目都带有 f_2,f_3 的属性特征,所以为这两个属性分别赋予较高的权值。此时,将每个标签出现的次数相加,可以得到项目 $item_1$ 的标签向量为 {0,4,4,2,2},$item_4$ 和 $item_5$ 的标签向量分别为 {0,0,0,2,2},{0,4,4,0,0}。这样通过式(2)计算得到 $item_4$、$item_5$ 与 $item_1$ 的相似度分别为 0.447 和 0.632,更能区分项目间的相似度。

(3)选取一定的阈值。得到满足阈值的目标项目 i 的相似项目,进而得到用户 u 在用户-项目评分矩阵中已评分的数据集合 R 以及对应的特征属性相似度值集合 W。

由于每个项目是多标签的,几乎项目间都具有较强的相关性,所以阈值的选定对于稀疏数据集的填充很关键。文中阈值选取 0.9,只有相似度值大于 0.9 的项目才能作为目标项目 i 的相似项目。

(4)计算未评分项目的预测评分。

由第三步可以得到所有未评分项目在用户-项目矩阵中的评分数据集,通过式(3)计算出每个未评分项目的预测评分。其中,$r_{u,j} \in R, w_{i,j} \in W$。

$$\mathrm{pre}_{u,i} = \frac{\sum_{i=1}^{n} r_{u,j} * w_{i,j}}{\sum_{i=1}^{n} w_{i,j}} \tag{3}$$

(5)结合项目平均值,对稀疏数据集中的未评分项目进行混合填充。

由于用户-项目数据集的极度稀疏性,可能获取的目标项目 i 的相似项目的已评分值集合 R 是空集。对此,采用评分均值来填充,得到用户 u 对项目 i 的评分,如式(4)所示:

$$\mathrm{rating} = \begin{cases} r_{u,i} \text{ if } u \text{ rated } i \\ \mathrm{pre}_{u,i} \text{ if } u \text{ not rated } i \text{ and } R \notin \varnothing \\ \mathrm{mean}_i \text{ if } u \text{ not rated } i \text{ and } R \in \varnothing \end{cases} \tag{4}$$

最终,可以得到一个完整的用户-项目评分矩阵 U-I,再对矩阵 U-I 使用式(1)计算项目间的相关性,得到项目评分相似度矩阵 sim_r,再将 sim_r 与 sim_f 组合后的相似性作为项目 i,j 的最终相似性。组合如公式(5)所示:

$$\mathrm{sim}(i,j) = w * \mathrm{sim}_r(i,j) + (1-w) * \mathrm{sim}_f(i,j) \tag{5}$$

式中,w 为设定用于调节基于两种来源的项目相似性平衡参数且 $w \in (0,1)$。

由 $\mathrm{sim}(i,j)$ 可以得到目标项目 i 的 k 个最近邻集合 $N = \{i_1, i_2, \cdots, i_k\}$,并且 i 不属于 N,且目标项目 i 和集合 N 中所有元素的相似值 $\mathrm{sim}(i,i_j)$,$j = 1,2,\cdots,k$ 依次递减,则用户 u 对项目 i 的最终预测评分计算公式如下所示:

$$R_{u,i} = \frac{\sum_{j \in N} \mathrm{sim}(i,j) * \mathrm{rating}}{\sum_{j \in N} |\mathrm{sim}(i,j)|} \tag{6}$$

最终,可以得到对用户 u 的所有项目的预测评分,并对预测评分进行排序,选择分值较高且用户实际未评分的项目向用户进行推荐。

4 实 验

4.1 数据集及实验环境

文中实验数据采用的是 Minnesota 大学(美国)提供的 Movielens 数据集[20],其中包含了 943 名用户对 1 682部电影的 100 000 条评分数据,其数据稀疏度为 0.94。该数据集已经在推荐系统中得到了广泛的使用和测评。

Movielens 数据集一共有三个数据集合:电影的属性描述集合、评分用户的个人信息集合以及用户对电影的评分集合。文中用到了第一和第三个集合。电影的属性描述集合记录了电影的名称、发行日期、电影所属类别(喜剧、战争等)。文中所用到的特征属性可以从这个集合中提取。用户对电影的评分集合记录了用户对电影的评分和时间戳。评分分值是从 1 到 5 的整数并且每个用户都评价了至少 20 部电影,每部电影至少都有一次以上的评价。整个数据集按 8∶2 的比例分成训练集和测试集,训练集数据作为算法输入,而测试集用于测试改进后的算法性能。

算法的实验环境:R 语言编程软件 Rgui。

4.2 衡量指标

文中主要采用推荐的质量即精度(MAE)和推荐的范围即覆盖率(coverage)作为主要衡量指标,并分析了在邻居个数固定情况下推荐数目对覆盖率的影响。

MAE 通过计算由训练集得到的用户预测评分和测试集中用户的实际评分的变差来度量预测的准确性。MAE 越小表明推荐质量越好。假定预测的用户评分集合为 $\{p_1, p_2, \cdots, p_N\}$,对应用户的实际评分集合为 $\{q_1, q_2, \cdots, q_N\}$,则 MAE 的定义式如下:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N} \tag{7}$$

覆盖率用于描述一个推荐系统对物品长尾的发掘能力。具体可以定义为推荐系统能够推荐出来的物品占总物品集合的比例。假定系统的用户集合为 U,推荐系统给每一个用户推荐一个长度为 L 的物品列表 $R(u)$。那么推荐系统的覆盖率可以通过式(8)计算得到:

$$\mathrm{coverage} = \frac{\left| \bigcup_{u \in U} R(u) \right|}{|I|} \tag{8}$$

式中,I 代表电影的总数。

4.3　实验结果及分析

为了检验文中算法的有效性,将混合填充方法和传统的对未评分项不进行填充(non filling)以及对未评分项进行众数填充(mode filling)的方法作比较。由于式(5)中的 w 为设定的用于调节项目相似性的平衡参数,所以 w 的取值对系统的推荐精度有影响。先计算在相同邻居个数下不同 w 值对两种算法的 MAE 值的影响,这里,设定邻居个数 k 为 10。实验中 w 的取值从 0.1 到 1,每次增加 0.1,观察 w 的变化对推荐系统效率的影响。实验结果如图 1 所示。

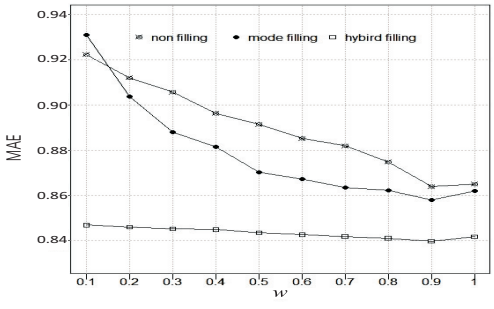

图 1　参数 w 对 MAE 的影响

从图中可以看到,当 w 的取值从 0.1 到 1 时,文中提出的 hybrid filling 所获得的 MAE 值远低于 non filing 以及 mode filling。当 w 取 0.9 时,三种方法的 MAE 都取得了最小值。再分析当 w 取 0.9,推荐个数取 10 时,邻居个数对推荐系统质量的影响。图 2 为邻居个数 k 从 3 增加到 18,间隔为 3 时,三种方法的 MAE 值。

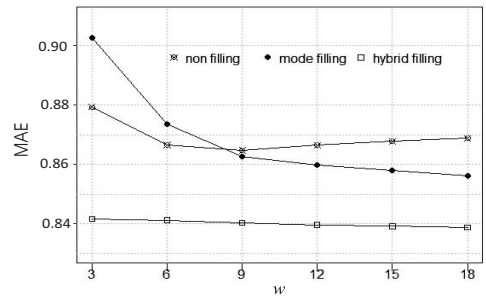

图 2　邻居个数 k 对 MAE 的影响

由图 2 可知,当推荐个数和 w 取值固定时,邻居个数 k 从 3 增加到 18 的过程中,文中提出的 hybrid filling 相比于 non filling 和 mode filling 均具有最小的 MAE 值。再分析当 w 取值固定(0.9)和推荐个数为 10 时,三种方法的推荐覆盖率(coverage),如图 3 所示。

由图 3 可知,文中方法的推荐覆盖率相比于其他两种方法有了明显的提高。

传统的对稀疏数据集缺失项不进行填充以及赋予固定缺省值的方法,在数据矩阵极度稀疏的情况下,会使得大部分项目的相似项目无法计算或者都是相同的,故而会影响推荐的覆盖率,而混合填充方法则主要利用由项目的特征属性得到的预测值来填充,这会使得为同类型项目推荐更多相似的同类型项目,故而提高了推荐的覆盖率。

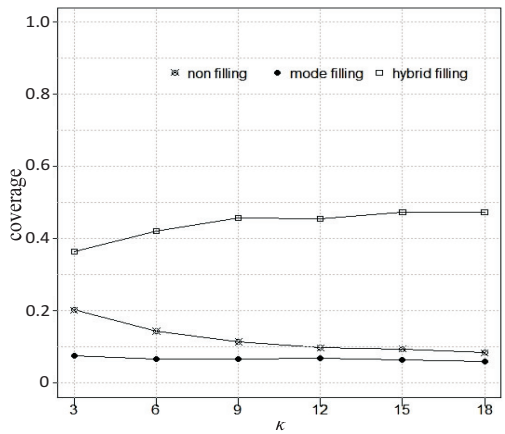

图 3　邻居个数 k 对推荐覆盖率的影响

再分析 w 、k 的取值固定时,推荐个数对推荐覆盖率的影响(w 取 0.9, k 取 10),结果如表 3 所示。

表 3　推荐个数对推荐覆盖率的影响　　%

推荐个数	non filling	mode filling	hybird-filling
10	10.46	6.78	45.42
20	14.15	9.75	60.94
30	17.90	13.50	69.78

从表 3 可以看出,推荐个数越多,推荐的范围就越广,推荐的覆盖率就越高。故而三种方法的推荐覆盖率随着推荐个数的增加而增大。

由上可知,文中所提出的混合填充方法,相比于对缺失项不作填充以及采用“众数填充”的方法,有效提高了推荐系统的推荐质量和推荐的覆盖率。

5　结束语

针对协同过滤算法的稀疏性问题,文中给出了一种新的填充方法,即充分利用由项目自身的特征属性得到的预测值并结合项目均值对稀疏数据集进行混合填充处理,使用户-项目评分矩阵得以饱和。通过实验验证了该方法相比于传统的对缺失值不作处理以及对缺失值赋予固定缺省值的方法,无论是在推荐精度亦或是推荐的覆盖率方面都有明显的改善。下一步的工作将侧重于对协同过滤算法进行改进,以提高推荐算法的质量、效率。

参考文献:

[1]　巩　亮. 推荐系统实践[M]. 北京:人民邮电出版社,2012.

[2]　Conner M, Herlocker J. Clustering items for collaborative filtering[C]//Proceeding of the ACM SIGIR workshop on recommender systems. [s. l.]:ACM,2001.

[3]　邓爱林,左子叶,朱扬勇. 基于项目聚类的协同过滤推荐算

法[J]. 小型微型计算机系统,2004,25(9):1665-1670.
[4] 邓爱林,朱扬勇,施伯乐. 基于项目评分预测的协同过滤推荐算法[J]. 软件学报,2003,14(9):1621-1628.
[5] Sarwar B,Karypis G,Konstan J,et al. Analysis of recommendation algorithms for E-commerce[C]//Proc of the second ACM conference on electronic commerce. New York: ACM Press,2000:158-167.
[6] Lee J S,Jun C H,Lee J,et al. Classification-based collaborative filtering using market basket data[J]. Expert System with Applications,2005,29(3):700-704.
[7] Resnick P,Iacovou N,Suchak M,et al. Grouplens:an open architecture for collaborative filtering of netnews[C]//Proceeding of the 1994 ACM Conference on computer supported cooperative work. New York,USA:Chapel Hill Press,1994:175-186.
[8] Bresse J,Hecherman D,Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[C]//Proceeding of the 14th conference on uncertainty in artificial intelligence. [s. l.]:[s. n.],1998:43-52.
[9] Linden G,Smith B,York J. Amazon. com recommendations item-to-item collaborative filtering[J]. IEEE Internet Computing,2003,7(1):76-80.
[10] Marlin B. Modeling user rating profiles for collaborative filtering[C]//Advances in neural information processing systems. Toronto,Canada:MIT Press,2003.
[11] Delgado J,Ishii N. Memory-based weighted-majority prediction for recommender systems [C]//Proceeding of ACM SIGIR'99 workshop on recommender systems. UC,USA:Berkeley Press,1999:251-257.
[12] Sarwar B,Karypis G,Konstan J,et al. Item-based collaborative filtering recommendation algorithms[C]//Proc of the tenth international conference on world wide web. New York: ACM Press,2001:285-295.
[13] 张忠平,郭献丽. 一种优化的基于项目评分预测的协同过滤推荐算法[J]. 计算机应用研究,2008,25(9):2658-2660.
[14] 徐 翔,王煦法. 协同过滤算法中的相似度优化方法[J]. 计算机工程,2010,36(6):52-54.
[15] Dasu T,Johnson T. Exploratory data mining and data cleaning [M]. [s. l.]:Wiley Press,2003.
[16] 余 刚,王知衍,邵 璐,等. 基于奇异值分解的个性化评论推荐[J]. 电子科技大学学报,2015,44(4):605-610.
[17] 霍淑华,杜永萍,黄 亮,等. 融入用户与物品特征信息的动态 RSVD 算法[J]. 山西大学学报:自然科学版,2015,38(1):24-30.
[18] 孙小华,陈 洪,孔繁胜. 在协同过滤中结合奇异值分解与最近邻方法[J]. 计算机应用研究,2006,23(9):206-208.
[19] 夏建勋,吴 非,谢长生. 应用数据填充缓解稀疏问题实现个性化推荐[J]. 计算机工程与科学,2013,35(5):15-19.
[20] Lee H,Kwon J. Similar User clustering based on MovieLens data set[J]. Advanced Science and Technology Letters,2014,51(8):32-35.

(上接第 58 页)

参考文献:

[1] 章毓晋. 图像分割[M]. 北京:科学出版社,2001.
[2] Kanungo T,Mount D,Netanyahu N,et al. An efficient k-means clustering algorithm: analysis and implementation[J]. IEEE Transactions on Pattern Analysis on Machine Intelligence,2002,24(7):881-892.
[3] Baillard C,Hellier P,Barillot C. Segmentation of brain 3D M R images using level sets and dense registration[J]. Medical Image Analysis,2001,5(3):185-194.
[4] 桂 阳,苑 云,杜 晶. 融合均值漂移和加权谱聚类的彩色图像分割[J]. 计算机应用研究,2012,29(9):3528-3530.
[5] 张 琦,卢志茂,徐 森,等. 基于相似度矩阵的谱聚类集成图像分割[J]. 传感器与微系统,2013,32(10):21-23.
[6] 朱 峰,宋余庆,朱玉全,等. 基于多阶抽样谱图聚类彩色图像分割[J]. 计算机科学,2010,37(7):264-266.
[7] 张向荣,骞晓雪,焦李成. 基于免疫谱聚类的图像分割[J]. 软件学报,2010,21(9):2196-2205.
[8] 贾建华,焦李成. 空间一致性约束谱聚类算法用于图像分割[J]. 红外与毫米波学报,2010,29(1):69-74.
[9] 李俊英,汪西莉. 一种新的大规模复杂图像分割的谱聚类方法[J]. 计算机应用研究,2011,28(5):1994-1997.
[10] 朱长明,李 晶,顾国昌,等. 谱聚类集成的淋巴结超声图像分割算法[J]. 计算机辅助设计与图形学学报,2009,21(10):1480-1486.
[11] 丁 阳,钱鹏江. 医学图像分割中基于数据浓缩的谱聚类算法[J]. 计算机工程,2012,38(12):17-21.
[12] 钟清流,蔡自兴. 用于彩图分割的自适应谱聚类算法[J]. 计算机应用研究,2008,25(12):3697-3699.
[13] Liu H Q,Jiao L C,Zhao F. Non-local spatial spectral clustering for image segmentation[J]. Neurocomputing,2010,74(3):461-471.
[14] Bai X D,Cao Z G,Wang Y,et al. Image segmentation using modified SLIC and Nyström based spectral clustering[J]. Optik-International Journal for Light and Electron Optics,2014,125(16):4302-4307.
[15] Filippone M,Camastra F,Masulli F,et al. A survey of kernel and spectral methods for clustering[J]. Pattern Recognition,2008,41(1):176-190.
[16] 蔡晓妍,戴冠中,杨黎斌. 谱聚类算法综述[J]. 计算机科学,2008,35(7):14-18.
[17] 纳跃跃,于 剑. 一种用于谱聚类图像分割的像素相似度计算方法[J]. 南京大学学报:自然科学版,2013,49(2):159-168.
[18] 谢 红,赵洪野. 基于卡方距离度量的改进 KNN 算法[J]. 应用科技,2015,42(1):10-14.
[19] 徐 瑞. 图像分割方法及性能评价综述[J]. 宁波工程学院学报,2011,23(3):76-79.