软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(8):1868−1884 [doi: 10.3724/SP.J.1001.2013.04350]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

融合奇异性和扩散过程的协同过滤模型*

杨兴耀[1], 于 炯[1,2], 吐尔根·依布拉音[1], 廖 彬[1], 钱育蓉[2]

[1](新疆大学 信息科学与工程学院,新疆 乌鲁木齐 830046)

[2](新疆大学 软件学院,新疆 乌鲁木齐 830008)

通讯作者: 杨兴耀, E-mail: yangxy@xju.edu.cn

摘 要: 作为解决信息过载问题的有效方式,推荐系统能够根据用户偏好对海量信息进行过滤,为用户提供个性化的推荐.但在推荐过程中,性能表现优异的协同过滤模型并没有充分利用上下文信息,这在一定程度上使系统面临性能瓶颈.为了进一步提高系统性能,从评分上下文信息着手,通过对项目评分进行分类统计获得评分奇异性,同时借鉴多渠道扩散相似性模型将推荐系统作为用户-项目二分网络的思想,提出了融合奇异性和扩散过程的协同过滤模型(collaborative filtering model fusing singularity and diffusion process,简称 CFSDP).为了表明模型的优越性,比较实验基于 MovieLens,NetFlix 和 Jester 这 3 个不同的数据集展开.实验结果表明,该模型不仅具有良好的扩展性,而且在合理的时间开销下,可以显著提高系统的预测和推荐质量.

关键词: 推荐系统;协同过滤;相似性度量;奇异性;扩散过程

中图法分类号: TP311 **文献标识码**: A

中文引用格式: 杨兴耀,于炯,吐尔根·依布拉音,廖彬,钱育蓉.融合奇异性和扩散过程的协同过滤模型.软件学报,2013,24(8):1868−1884. http://www.jos.org.cn/1000-9825/4350.htm

英文引用格式: Yang XY,Yu J, Ibrahim T, Liao B, Qian YR. Collaborative filtering model fusing singularity and diffusion process. Ruan Jian Xue Bao/Journal of Software, 2013,24(8):1868−1884 (in Chinese). http://www.jos.org.cn/1000-9825/4350.htm

Collaborative Filtering Model Fusing Singularity and Diffusion Process

YANG Xing-Yao[1], YU Jiong[1,2], Turgun IBRAHIM[1], LIAO Bin[1], QIAN Yu-Rong[2]

[1](College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China)

[2](School of Software, Xinjiang University, Urumqi 830008, China)

Corresponding author: YANG Xing-Yao, E-mail: yangxy@xju.edu.cn

Abstract: As a key solution to the problem of information overload, the recommender system can filter a large deal of information according to user's preference and provide personalized recommendations for the user. However, traditional collaborative filtering models with excellent performance haven't made full use of the contextual information in the process of recommendation, which to some extent confronts the system with the performance bottleneck. In order to improve the system performance further, this paper starts with the contextual information on ratings, and proposes a collaborative filtering model fusing singularity and diffusion process (CFSDP) by taking advantage of ratings' singularities obtained from the classified statistics of ratings and referring to the similarity model of multi-channel diffusion which regards recommender system as a user-item bipartite network. To demonstrate the superiority of the proposed model, the study provides comparative experimental results based on the MovieLens, NetFlix and Jester data sets. Finally, the results show that the model not only has better extensibility, but also can observably improve the prediction and recommendation quality of system with a reasonable time cost.

Key words: recommender system; collaborative filtering; similarity measure; singularity; diffusion process

* 基金项目: 国家自然科学基金(61063042, 61262088, 61063026); 新疆自治区自然科学基金(2011211A011); 新疆高校重大科研项目(XJEDU2012I10); 新疆大学博士创新项目(XJUBSCX-2011007); 新疆大学博士科研启动基金(BS100128)

收稿时间: 2012-06-26; 修改时间: 2012-08-20; 定稿时间: 2012-11-15

在推荐系统中,传统的协同过滤模型在推荐过程中并没有考虑项目评分的上下文信息,仅仅是对评分值本身进行一般性运算,这在一定程度上限制了系统性能的进一步提高.事实上,通过对大量的评分进行统计分析,可以获得很多上下文信息,利用这些信息对系统性能做进一步的改善,显然具有重要的意义.

关于评分的上下文信息的种类很多,比如用户评分的波动情况、各个评分级别内的评分个数等,这些都可以为推荐提供参考,如何发现和利用这些上下文信息就成了不同推荐模型的关键.例如,文献[1]在推荐过程中,利用基于项目的 slop-one 协同过滤方法获得个性化的上下文信息来对未评分项目进行评分预测,实验结果表明,文中模型在预测精度上比先前的模型有较大的提高;社会网络在为用户访问资源带来方便的同时,在推荐过程中却出现了上下文问题,为此,文献[2]提出了融合上下文信息的协同过滤推荐模型,该模型不仅较好地解决了问题,而且具有更好的预测准确度;文献[3]为了增强基于项目的协同过滤模型的推荐效果,在项目相似性计算中引入项目类别因素等上下文信息,并在此基础上提出了集成语境信息的多维推荐模型,对比实验结果表明,该模型在数据比较稀疏的情况下能够获得较好的推荐效果.

另外,值得关注的是,文献[4]对系统中的评分级别进行划分,并统计各个级别内的评分个数获得评分奇异值,提出了基于奇异性的协同过滤相似性度量模型(collaborative filtering similarity measure based on singularity,简称 SM).与传统的性能优异的皮尔逊相关系数相比,该模型在预测质量和推荐质量方面均有较大的改善,而且在系统存在多个评分级别时,具有良好的扩展性.

上述模型在利用上下文信息方面均有自己的特色之处,性能上也比传统的协同过滤模型有很大的提高,这充分体现了上下文信息所起的关键作用.另一方面,换个角度采用新的用户相似性度量模型,在性能改进方面也可以收到别样的效果.典型的如文献[5],其将推荐系统看成一个关于用户-项目的二分网络,用户之间的关系通过系统评分级别来连接,当且仅当两个用户的评分相同时,用户之间才是连通的,具有相似性.基于此提出了多渠道扩散的协同过滤模型(collaborative filtering based on multi-channel diffusion,简称 Diffusion).该模型与皮尔逊相关系数相比,在预测质量方面性能更优,但却使数据稀疏性问题恶化.

对不同相似性模型进行分析,尤其是文献[4,5]中的两个模型,它们没有以传统的度量模型为基础,却均获得了比传统模型更好的性能.但如果将两者结合起来,相互弥补各自的不足,性能或许会更好.本文出于这样的想法提出了新模型,然而在具体实现过程中需要迎接如下的挑战:

(1) 必须重新确定项目评分的划分标准,以适应不同的评分范围和不同的评分类型;

(2) 不应该进一步恶化数据稀疏性问题;

(3) 不应该显著增加模型的时间复杂性.

在对模型 SM 分析的基础上,本文首先对项目评分按照区间范围而非评分级别进行划分,分别进行统计以获得评分的奇异值,用来参与计算用户的相似性值,从而克服了评分值为连续值时模型的不适应性.然后,对模型 Diffusion 进行适应性改进,将原有的多渠道扩散模型改为单渠道模型,在避免恶化数据稀疏性问题的同时,扩大了模型的适用范围.最后综合上述两个模型,提出了融合奇异性和扩散过程的协同过滤模型(collaborative filtering model fusing singularity and diffusion process,简称 CFSDP).验证实验基于不同的数据集和不同的质量评估标准进行.实验结果表明,本文模型相对于原来的参照模型,不仅拥有更好的预测和推荐性能,而且时间花费较合理,具有较好的扩展性.

本文第 1 节介绍相关工作以及本文研究的意义.第 2 节详细阐述本文提出的模型和预测推荐过程.第 3 节介绍相关的质量评估标准.第 4 节详细介绍模型的仿真实验与结果分析.第 5 节对本文的工作进行总结与展望.

1 相关工作

互联网技术与应用的快速发展在带来方便的同时,也产生了信息过载问题.如何从海量信息中准确而快速地过滤出用户所需要的信息,推荐系统应运而生[6].与搜索不同,推荐系统能够获取隐含信息,将用户潜在模糊的需求现实化、具体化,快速给出个性化的服务推荐,减轻用户盲目搜寻的痛苦,提高用户满意度.

目前,推荐系统的应用领域日益广泛,尤其是 Web 2.0 技术的成熟,推荐被作为一个独立的概念提出来并得

到系统研究,已成为电子商务、社交网络、视频点播等 Web 2.0 主要服务的核心技术[7−9].其中,协同过滤作为推荐系统普遍采用的重要技术之一,其原理是根据相似用户的兴趣对当前用户可能感兴趣的信息进行推荐.另外,根据推荐信息产生方式的不同,推荐技术除协同过滤之外,通常还有以下几类:

- 基于内容的过滤[10,11]:首先分析项目或者资源的内容属性,然后根据用户评价过的项目建立用户的兴趣模型,利用项目与模型之间的相似性来对信息进行过滤.其不足在于,只能发现与用户已有兴趣相似的资源,不能发现新的潜在的用户感兴趣的资源.
- 基于规则的过滤[12,13]:关注用户的行为模式,找出用户项目集合中的项目关联性,建立关联规则.例如,购买鱼缸的用户通常会购买金鱼,因此可以在鱼缸和金鱼之间建立关联规则,通过这种规则向用户推荐其他项目.其不足同上.
- 基于人口统计信息的过滤[14,15]:首先获得用户的基本信息,例如年龄、性别、职业、受教育程度等,然后根据基本信息衡量用户之间的相关程度,将相似用户喜爱的其他项目推荐给当前用户.其不足在于,用户的人口统计基本信息难以获得.
- 基于网络结构的过滤[16,17]:不考虑用户项目的属性特征,仅仅将它们看作抽象的点,利用用户-项目二分图来建立关联关系,这种关系包含了推荐所需要的各种信息,以此向用户进行推荐.这种模型的不足在于会进一步恶化数据稀疏性问题.
- 混合过滤:单一的过滤模型均有各自的不足,目前常用的是将它们结合起来,即混合过滤,这样可以弥补彼此的缺陷.典型的成果有:文献[18,19]融合基于用户和基于产品的协同过滤算法产生推荐,较好地解决了数据稀疏性问题,提高了预测结果的准确性;文献[20]在协同过滤系统中加入内容,同样取得了较好的推荐效果;文献[21,22]利用基于社会网络信息对已有的协同过滤算法进行改进,结论表明,该算法显著提高了性能;同时,在混合过滤模型中也出现了新的类型的组合,例如,文献[23]将概率融入到语义分析中并与基于产品的协同过滤算法结合,改善了原有算法的可扩展性,提高了推荐质量.

为了比较不同模型的性能,文献[24,25]提出了一些评估标准,分为两类:(1) 预测质量方面,如预测准确率、预测覆盖率、较好预测百分数、较差预测百分数等;(2) 推荐质量方面,如推荐精度、推荐召回率、新颖度、信任度等.关于采用何种衡量标准以及标准的具体含义,本文后面会详细介绍.另外,为了更好地评估模型的性能,还需要考虑实验条件等其他因素,比如项目类型、数据集种类等,以及需要忽略某些无关的因素,比如文献[26].

事实上,推荐系统中无论采用何种模型,其目标通常是最大程度地减小预测偏差,提高推荐质量.其中,基于协同过滤的推荐模型由于具有优秀的性能表现而得到广泛应用,这主要是由于协同过滤模型拥有较好的度量用户相似性的模型.可以通过如下的步骤为需要获得推荐的用户(即目标用户)提供推荐:

(1) 利用相似性度量模型获得一定数目的近邻作为目标用户的推荐近邻;

(2) 利用预测函数对推荐近邻的评分进行处理,获得目标用户关于未评分项目的评分预测值;

(3) 选择评分预测值最高的 N 个项目向目标用户进行推荐.

其中的关键是相似性度量模型[27,28],常用的有:(1) 余弦法 Cosine,简称 COS;(2) 皮尔逊相关系数法 Pearson correlation,简称 PC;(3) 改进的皮尔逊相关系数法 Constrained Pearson's correlation,简称 CPC;(4) 斯皮尔曼等级相关系数 Spearman rank correlation,简称 SPR.

这些模型虽然都有良好的性能表现,但相似性值的获取均是对评分的直接操作,未能充分利用背后隐含的上下文信息,这在一定程度上限制了模型性能的进一步提升.在协同过滤过程中如何处理上下文信息,也是本文需要重点考虑的问题之一.目前通用的做法是基于分布式集群[29]和统计模型,比如概率式潜藏语意分析,来发掘利用关于评分的附加信息,以提高推荐系统的预测和推荐质量.

2 CFSDP 阐述

推荐系统中,协同过滤的关键是利用相似性度量模型为目标用户选择近邻.本节将给出融合奇异性和扩散过程的协同过滤模型 CFSDP:首先介绍奇异性和扩散过程的概念,然后给出相似性度量的基本模型及扩展模型,

最后说明整个预测和推荐过程.

2.1 奇异性概念

推荐系统针对用户评价,通常会提供一个数值的评分级别范围,可以是离散的,也可以是连续的,用户从中选取一个数值表达自己对项目的评价,通常数值越大,评价越高.这时,如果系统中某些用户对某个项目的评分与其余的用户评分不同,则这部分用户应该得到更多的关注,因为它们之间除了评分还具有另外一种意义上的相似性.以离散的评分级别集合{1,2,3,4,5}为例,可以将其简单地划分为两类:R^+={4,5},R^-={1,2,3},分别称为积极、非积极评价级别集合.当然,也可以将其划分为 3 类:R^1={1,2},R^2={3},R^3={4,5}.具体如何划分应根据实际需要而定,这样,评分属于某一集合的用户就应该具有另外一种意义上的相似性,而且这种相似性与各个集合中的用户数量成反比.特殊的情况是,仅有两个用户的评分与众不同,则这两个用户之间除了评分上的相似性之外,在其他用户看来显然还具有另外的相似性,而且程度比较高.

本文将上述的相似性概念称为奇异性,相应的值称为奇异值.这样,当在计算传统的相似性时,如果考虑奇异性因素,则某个项目对用户相似性值的影响将不再是绝对的,而与项目的其他评分也有关系.这在通常情况下是合理的,即相似性度量不能仅关注单个评分,还必须考虑整体评分的差异程度.差异程度越大,对相似性的影响越大.

2.2 多渠道扩散过程

基于多渠道扩散的协同过滤是一种基于网络结构的资源分配模型.该模型的核心思想是:首先将推荐系统中的每个项目按照评分级别划分成不同的渠道,每个渠道对应一个评分,然后利用用户对项目的评分信息构造用户-渠道二分网络,再经过两步的扩散过程完成资源分配,最终获得用户之间的相似性.

在具体的资源分配过程中,假定每个用户均拥有一定的资源,这里的资源可以理解为推荐能力等.

第 1 步,每个用户首先将自己拥有的资源按照项目评分均等的分配给对应的渠道,则项目 i 上的渠道 c 从用户 u 获得的资源量为

$$R_{uic}=a_{uic}/d(u),$$

其中,a_{uic} 为渠道连通因子 0 或 1.当用户关于项目的评分与渠道对应时,a_{uic}=1;否则,a_{uic}=0.$d(u)$为用户 u 的度数,即评价项目的个数.

第 2 步,每个项目渠道将获得的资源量均等地分配给相应的用户,这样可以将用户 u 从用户 v 获得的资源量 S_{uv} 定义为用户 u 和 v 间的相似度,S_{uv} 值不由评分确定的性质可以减少评价信息丢失,计算过程如下:

$$S_{uv}=\sum_{c=1}^{n}\frac{a_{uic}R_{vic}}{d(c)}=\frac{1}{d(v)}\sum_{c=1}^{n}\frac{a_{uic}a_{vic}}{d(c)},\ S_{uv}\in(0,1] \tag{1}$$

其中,a_{vic} 为渠道连通因子;$d(c)$为渠道 c 的度数,即对渠道 c 评价过的用户个数;n 为渠道总数目.

这样,根据 S_{uv} 逐一计算系统中用户之间的相似度,就可以获得用户相似性矩阵 $S=\{S_{uv}\}$.可以看出,S 是一个非对称的矩阵,即 $S_{uv}\neq S_{vu}$.另外,由于每个用户的初始资源量是一定的,在资源分配过程中,用户的资源支出量和其他用户的资源获得量相等,所以,矩阵 $S=\{S_{uv}\}$还是一个列正则化的矩阵,即 $\sum_u S_{uv}=1$, 并且在整个资源分配过程中,资源总量守恒.

2.3 基本模型

2.3.1 奇异性模型

文中定义 U 为推荐系统中的用户集合,I 为项目集合,$r_{u,i}$ 为用户 $u\in U$ 对项目 $i\in I$ 的评分.推荐系统中,用户对项目的所有评分可以看成一个二维矩阵,整个评分级别为[m,M]之间的数,通常是整数,也可能是小数,另外还包含一个空值 *null*,表示用户未对项目评分.为方便说明,本文特选取 9 个用户对 5 个项目的评分构成一个 9×5 的二维矩阵,评分级别范围为[1,5]之间的整数,级别数为 5,其中,m=1 表示评分最低,M=5 评分最高,*null* 表示未评分,见表 1.

此时,由于 m=1,M=5 评分级别范围较小,文中将其分为两个区间,重新定义 R^+=(4,5],R^-=[1,3]分别为积极、

非积极评分级别区间,这样有评分级别集合$\{4,5\}\subseteq R^+$,$\{1,2,3\}\subseteq R^-$成立.

Table 1 An example of rating matrix

表 1 评分矩阵样例

	i_1	i_2	i_3	i_4	i_5
u_1	1	5	2	1	1
u_2	1	4	2	2	
u_3	1	5	1	1	1
u_4	5	4	5	3	5
u_5	5	4	4	3	3
u_6	–	–	1	–	–
u_7	–	–	2	–	1
u_8	–	5	4	1	1
u_9	–	4	4	–	1

同时定义$P_i=\{u\in U|r_{u,i}\in R^+\}$,$N_i=\{u\in U|r_{u,i}\in R^-\}$分别为关于项目$i$的积极、非积极评分用户集合.定义$s_p^i$,$s_n^i$分别为关于项目$i$的积极、非积极评分的奇异值,值的大小在一定程度上反映了两者评分与整体评分之间的差异性,计算表达式分别为$s_p^i=1-num(P_i)/num(P_i+N_i)$,$s_n^i=1-num(N_i)/num(P_i+N_i)$,$num$表示集合中元素的数目.可以看出,相应奇异性用户的数目越多,奇异值就越小,反之亦然.在表 1 中,以项目 1 和项目 3 为例,有:

- $P_1=\{4,5\}$, $s_p^1=1-2/5=0.6$, $P_3=\{4,5,8,9\}$, $s_p^3=1-4/9=0.556$;
- $N_1=\{1,2,3\}$, $s_n^1=1-3/5=0.4$, $N_3=\{1,2,3,6,7\}$, $s_n^3=1-5/9=0.444$.

由此可以看出,对于$\forall i\in I$, $s_p^i,s_n^i\in[0,1]$,其中,值为 0 表示奇异性最低,值为 1 表示奇异性最高,且$s_p^i+s_n^i=1$.

这样,当两个用户u,v对项目i的评分不为$null$,即$r_{u,i}\neq null\wedge r_{v,i}\neq null$时,奇异性可以有 4 种组合,对每种组合的奇异值计算进行整理,见表 2.

Table 2 Four singularity combinations

表 2 奇异性的 4 种组合

组合	项目集合表示	奇异值计算
$r_{u,i}\in R^+\wedge r_{v,i}\in R^+$	I^{++}	$s_p^i s_p^i$
$r_{u,i}\in R^+\wedge r_{v,i}\in R^-$	I^{+-}	$s_p^i s_n^i$
$r_{u,i}\in R^-\wedge r_{v,i}\in R^+$	I^{-+}	$s_n^i s_p^i$
$r_{u,i}\in R^-\wedge r_{v,i}\in R^-$	I^{--}	$s_n^i s_n^i$

2.3.2 扩散过程模型

在考虑奇异性时,本文同时借鉴多渠道扩散过程模型,对目前用户之间的相似性度量进行改进.多渠道扩散过程模型与传统的 PC 相比虽然具有更加优越的性能,但也存在不足之处,表现在以下 3 个方面:

(1) 只有当两个用户之间存在对同一个项目具有相同的评分时,用户间的渠道才是连通的,才具有相似性.由于评分矩阵具有稀疏性的特点,这样做会恶化数据稀疏性问题.

(2) 由于渠道连通因子的限制,项目的渠道数目不能太多,所以该模型只能处理评分级别个数有限的情况,而无力处理评分级别为连续值的情况,否则会使数据稀疏性问题进一步恶化.

(3) 在两步的资源扩散过程中,用户和渠道之间的资源都是均等分配的,这种简单的分配方式容易导致某些信息丢失,比如用户的积极评价信息等,降低了模型的性能.

面对上述缺陷,本文首先需要对现有的多渠道扩散过程模型进行适应性改进,使其在不恶化数据稀疏性问题的基础上,不仅可以应对评分级别为连续值时的情况,而且可以让资源分配更好地反映用户评价信息以优化模型性能.具体改进的方法如下:

(1) 放宽相似性存在的标准,将用户间渠道连通的条件设定为用户之间对于某一项目存在评分,而非相同评分,即将原来的多渠道改为现在的单渠道.这样,在弥补上述缺陷的同时,还可以与奇异性模型相

结合,同时具备可扩展性.这一点会在扩展模型中介绍.

(2) 渠道连通因子的取值不再为简单的 0 或 1,而是直接为用户关于项目的评分,用户渠道的度数也进行相应的调整,因为这样就可以让资源在扩散过程中根据评价信息按比例进行合理的分配,从而更好地量化用户之间的相似性,改善模型各方面的性能.

改进后的多渠道扩散过程模型在应用过程中首先需要完成两步扩散过程中的第 1 步,即用户将自己的资源按评分比例分配给相应的项目或渠道,于是,项目 i 上从用户 u 获得的资源量为

$$R_{ui}=r_{u,i}/d(u),$$

其中, $d(u)=\sum_{i\in I}r_{u,i}$ 为用户 u 的度数.

在接下来的第 2 步中,项目需要将获得的资源按评分比例重新分配给相应的用户,从公式(1)中渠道连通因子取值同时为 1 的条件可以获得启发:模型改进后,S_{uv} 值的计算无须遍历系统中所有的项目,而只需关注用户间的公共评分项目集合即可,这样可以大幅度降低模型的计算复杂度,同时也可以为模型的拓展使用打下基础.于是,定义 $A_{u,v}=\{i\in I|r_{u,i}\neq null\wedge r_{v,i}\neq null\}$ 为用户 u 和 v 的公共评分项目集合,则改进后的 S_{uv} 如下:

$$S_{uv}=\sum_{i\in A_{u,v}}\frac{r_{u,i}R_{vi}}{d(i)}=\frac{1}{d(v)}\sum_{i\in A_{u,v}}\frac{r_{u,i}r_{v,i}}{d(i)},S_{uv}\in[0,1] \tag{2}$$

其中, $d(i)=\sum_{k\in U}r_{k,i}$ 为项目 i 的度数.

这样,根据 S_{uv} 获得的相似性矩阵 $S=\{S_{uv}\}$ 仍然是一个非对称的矩阵,即 $S_{uv}\neq S_{vu}$.这种不对称性是合理的,因为每个用户的资源是一定的,如果一个用户对很多项目进行了评分,那么该用户将有更高的概率与其他用户建立连接,从而导致该用户赋予其他用户的权重降低,反之亦然.同时,矩阵 $S=\{S_{uv}\}$ 仍然是一个列正则化的矩阵.从而说明了多渠道扩散过程模型改进的合法性.

下面举一个 S_{uv} 的例子.例如,计算用户 u_1 与 u_8 之间的相似性 S_{18},此时,$A_{1,8}=\{2,3,4,5\}$,则

$$S_{18}=\frac{1}{11}\sum_{i\in A_{1,8}}\frac{r_{1,i}r_{8,i}}{d(i)}=(5\times5)/(11\times31)+(2\times4)/(11\times25)+(1\times1)/(11\times11)+(1\times1)/(11\times13)=0.118.$$

2.3.3 模型的建立

上面分别介绍了用户评分的奇异性和扩散过程模型,事实上,可以基于公共评分项目集合 $A_{u,v}$ 将两者结合起来.例如,对于 $\forall i\in A_{u,v}$,不仅考虑 $r_{u,i},r_{v,i}$ 本身的奇异性,还考虑扩散过程 S_{uv}.这样,由于奇异性分两类,此时有 4 种组合情况,见表 3.

Table 3 Four singularity combinations with diffusion processes

表 3 考虑扩散过程的 4 种奇异性组合

组合	项目集合表示	值计算
$r_{u,i}\in R^+\wedge r_{v,i}\in R^+$	$A_{u,v}^{++}$	$(r_{u,i}r_{v,i}s_p^is_p^i)/(d(v)\times d(i))$
$r_{u,i}\in R^+\wedge r_{v,i}\in R^-$	$A_{u,v}^{+-}$	$(r_{u,i}r_{v,i}s_p^is_n^i)/(d(v)\times d(i))$
$r_{u,i}\in R^-\wedge r_{v,i}\in R^+$	$A_{u,v}^{-+}$	$(r_{u,i}r_{v,i}s_n^is_p^i)/(d(v)\times d(i))$
$r_{u,i}\in R^-\wedge r_{v,i}\in R^-$	$A_{u,v}^{--}$	$(r_{u,i}r_{v,i}s_n^is_n^i)/(d(v)\times d(i))$

两方面信息的结合,为了更好地衡量用户 $u,v\in U$ 之间基于公共评分项目集合的相似性 $sim(u,v)$ 提供了基础.在给出 $sim(u,v)$ 的表达式之前,需要先对 $r_{u,i}$ 和 $r_{v,i}$ 进行处理,通常的方式主要有平均绝对偏差(mean absolute difference,简称 MAD)、平均平方偏差(mean square difference,简称 MSD).其中,$num(A_{u,v})$ 表示集合 $A_{u,v}$ 中的项目数目,形式如下:

$$MAD=\frac{1}{num(A_{u,v})}\sum_{i\in A_{u,v}}|r_{u,i}-r_{v,i}| \tag{3}$$

$$MSD=1-\frac{1}{num(A_{u,v})}\sum_{i\in A_{u,v}}(r_{u,i}-r_{v,i})^2 \tag{4}$$

考虑到 $(r_{u,i}r_{v,i}s_p^i s_p^i)/(d(v)\times d(i))$ 的值在[0,1]之间且偏大较好,而 MAD,MSD 与之组合均存在不足,例如 MAD 的值并非偏大较好,所以文中采用改进的 MAD(adjusted MAD,简称 AMAD)来处理评分 $r_{u,i}$ 和 $r_{v,i}$.其中,分母加 1 是为了防止评分相等的情况.表达式如下:

$$AMAD=\frac{1}{num(A_{u,v})}\sum_{i\in A_{u,v}}\frac{1}{|r_{u,i}-r_{v,i}|+1},\ AMAD\in[0,1] \tag{5}$$

综合表 3 和公式(5),给出本文 $sim(u,v)$的基本模型如下:

$$sim(u,v)=\frac{1}{4}\left(\begin{array}{l}\dfrac{1}{num(A_{u,v}^{++})}\sum_{i\in A_{u,v}^{++}}\dfrac{r_{u,i}r_{v,i}s_p^i s_p^i}{(|r_{u,i}-r_{v,i}|+1)d(v)d(i)}+\dfrac{1}{num(A_{u,v}^{+-})}\sum_{i\in A_{u,v}^{+-}}\dfrac{r_{u,i}r_{v,i}s_p^i s_n^i}{(|r_{u,i}-r_{v,i}|+1)d(v)d(i)}\\ +\dfrac{1}{num(A_{u,v}^{-+})}\sum_{i\in A_{u,v}^{-+}}\dfrac{r_{u,i}r_{v,i}s_n^i s_p^i}{(|r_{u,i}-r_{v,i}|+1)d(v)d(i)}+\dfrac{1}{num(A_{u,v}^{--})}\sum_{i\in A_{u,v}^{--}}\dfrac{r_{u,i}r_{v,i}s_n^i s_n^i}{(|r_{u,i}-r_{v,i}|+1)d(v)d(i)}\end{array}\right) \tag{6}$$

可以看出,上述 $sim(u,v)\in[0,1]$同样不具有对称性,即 $sim(u,v)\neq sim(v,u)$,这与扩散过程模型的不对称性有关.它体现的相似性来源于 4 个部分,其中的项目分别包含于 4 个集合中: $A_{u,v}^{++}, A_{u,v}^{+-}, A_{u,v}^{-+}$ 和 $A_{u,v}^{--}$.当且仅当条件 $A_{u,v}^{++}\neq\varnothing\vee A_{u,v}^{+-}\neq\varnothing\vee A_{u,v}^{-+}\neq\varnothing\vee A_{u,v}^{--}\neq\varnothing$ 成立时,$sim(u,v)$的值存在.其中,当 4 个集合中有∅存在时,比如 $A_{u,v}^{++}=\varnothing$,则令相应的 $\frac{1}{num(A_{u,v}^{++})}\sum_{i\in A_{u,v}^{++}}\frac{r_{u,i}r_{v,i}s_p^i s_p^i}{(|r_{u,i}-r_{v,i}|+1)d(v)d(i)}=0$.

例如,计算 u_1 与 u_8 之间的相似性:

$$A_{1,8}^{++}=\{2\},\ A_{1,8}^{+-}=\varnothing,\ A_{1,8}^{-+}=\{3\},\ A_{1,8}^{--}=\{4,5\},\ num(A_{1,8}^{++})=1,\ num(A_{1,8}^{-+})=1,\ num(A_{1,8}^{--})=2,$$

则 $sim(1,8)=(5\times5\times0)/(1\times44\times31)+(2\times4\times0.247)/(2\times44\times25)+(1\times1\times0)/(1\times88\times11)+(1\times1\times0.29)/(1\times88\times13)=0.0012$.

2.4 扩展模型

当推荐系统评分级别范围[m,M]较大时,说明此时评分级别比较多,用户对项目的评价也更加丰富.这时,如果再像上面一样将评分范围简单地划分为积极和非积极两个区间,显然过于笼统,不能准确地体现用户评分之间的奇异性,并且有的评分并不好判定是积极还是非积极的,可能是中性的,而且中性评分在数量上可能会更多一些.这样,为了准确地反映用户对项目的真实评价,提高推荐系统的推荐准确性,需要将[m,M]根据具体情况多分几个小区间,设总区间数目为 n,则划分后的评分范围[m,M]表示为

$$[m,l_1],(l_1,l_2],\ldots,(l_{n-2},l_{n-1}],(l_{n-1},M],$$

其中,$m<l_1<l_2<\ldots<l_{n-1}<M$,$n$ 为自然数.

可以看出,上述例子中的 R^+,R^-仅为此划分的一个特例,其中,$R^-=[m,3]$,$R^+=(3,M)$,$n=2$,$l_1=3$.

为方便起见,不妨令 $R^1=[m,l_1]$,$R^n=(l_{n-1},M]$,$R^j=(l_{j-1},l_j]$,$j\in\{2,3,\ldots,n-1\}$.

同时,定义 Q_j^i 为关于项目 i 评分 $r_{u,i}\in R^j$ 的用户集合,$Q_j^i=\{u\in U\mid r_{u,i}\in R^j\}$,$s_j^i$ 为对应 Q_j^i 关于项目 i 的评分奇异值,$s_j^i=1-num(Q_j^i)\Big/\sum_{j=1}^n num(Q_j^i)$,$A_{u,v}^{j,q}$ 为 u,v 公共评分项目集,$A_{u,v}^{j,q}=\{i\in I\mid r_{u,i}\in R^j\wedge r_{v,i}\in R^q\}$,其中 $j,q\in\{1,2,\ldots,n\}$.

另外需要考虑扩散过程,由于改进后的扩散过程模型具有较好的可扩展性,这里不需要再作修改,可直接与上述奇异性相结合得出本文 $sim(u,v)$的扩展模型,表示为

$$sim(u,v)=1/n^2\times\sum_{j=1}^n\sum_{q=1}^n\frac{1}{num(A_{u,v}^{j,q})}\sum_{i\in A_{u,v}^{j,q}}\frac{r_{u,i}r_{v,i}s_j^i s_q^i}{(|r_{u,i}-r_{v,i}|+1)d(v)d(i)} \tag{7}$$

同样地,$sim(u,v)\in[0,1]$,其存在的条件是:对$\forall j,q\in\{1,2,\ldots,n\}$,$\exists\ A_{u,v}^{j,q}\neq\varnothing$, 当有∅存在时,例如 $A_{u,v}^{1,2}=\varnothing$, 则令

$$\frac{1}{num(A_{u,v}^{1,2})}\sum_{i\in A_{u,v}^{1,2}}\frac{r_{u,i}r_{v,i}s_1^i s_2^i}{(|r_{u,i}-r_{v,i}|+1)d(v)d(i)}=0.$$

为了进一步展示 $sim(u,v)$的扩展模型,下面以 Jester 数据集为例,其用户评分为[−10,10]内的连续数.为了与后文的实验保持一致,令 $n=3$,$R^1=[-10,-5]$,$R^2=(-5,5]$,$R^3=(5,10]$,则

$$sim(u,v)=1/3^2\times\sum_{j=1}^{3}\sum_{q=1}^{3}\frac{1}{num(A_{u,v}^{j,q})}\sum_{i\in A_{u,v}^{j,q}}\frac{r_{u,i}r_{v,i}s_j^i s_q^i}{(|r_{u,i}-r_{v,i}|+1)d(v)d(i)}.$$

2.5 预测与推荐过程

为目标用户 u 产生 N 个项目推荐,首先需要利用相似性度量为 u 寻找一定数量的用户近邻构成推荐近邻集合 K_u,K_u 中的用户相比其他用户与 u 具有更大的相似性值,则有下面的命题成立:

$$\forall x\in K_u,\ \forall y\in U-K_u\wedge y\neq u:\ sim(u,y)\leqslant sim(u,x).$$

然后,通过 K_u 中用户对项目 i 的评分,利用预测函数计算获得用户 u 对项目 i 的预测评分 $p_{u,i}$.通常的预测函数有:公式(8)为均值函数、公式(9)为权重函数、公式(10)为本文采用的改进型的权重函数,如下:

$$p_{u,i}=\frac{1}{num(G_{u,i})}\sum_{x\in G_{u,i}}r_{x,i} \tag{8}$$

$$p_{u,i}=\frac{1}{\sum_{x\in G_{u,i}}sim(u,x)}\sum_{x\in G_{u,i}}sim(u,x)r_{x,i} \tag{9}$$

$$p_{u,i}=\overline{r}_u+\frac{1}{\sum_{x\in G_{u,i}}sim(u,x)}\sum_{x\in G_{u,i}}sim(u,x)(r_{x,i}-\overline{r}_x) \tag{10}$$

其中,$\overline{r}_u$ 和 $\overline{r}_x$ 分别为评分均值,$G_{u,i}=\{x\in K_u|r_{x,i}\neq null\}$ 为 u 的推荐近邻集合中对项目 i 评分过的用户集合.可以看出,$p_{u,i}$ 存在的条件是 $G_{u,i}\neq\varnothing$,否则 $p_{u,i}=null$.因为 $G_{u,i}=\varnothing$ 时 K_u 中用户均未对 i 进行过评价,所以无法为 i 进行预测.

最后,为了获得关于用户 u 的 N 个推荐项目集合 N_u,定义 $E_u=\{i\in I|r_{u,i}=null\wedge p_{u,i}\neq null\}$ 为用户 u 未评分但可以获得评分预测的项目集合,并从中选取 N 个预测评分值最大的项目构成集合 N_u.特殊情况下,当 E_u 中项目的个数不足 N 时全部选取,此时有下面的命题成立:

$$N_u\subseteqq E_u,\ \forall x\in N_u,\ \forall y\in E_u-N_u:\ p_{u,y}\leqslant p_{u,x}.$$

3 质量评估标准

3.1 平均绝对误差(mean absolute error,简称MAE)

在推荐系统中,为了衡量相似性度量模型在项目预测方面的有效性,需要统计项目评分预测的偏差.常用的评价标准平均绝对误差 MAE 不仅易于理解,而且可以直观地对预测质量进行评价,所以在文中作为评价标准之一.

MAE 通过统计项目的预测评分与实际评分之间的差值来衡量预测的准确性,MAE 越小,准确性越高.这里,针对用户 u,定义 MAE_u 为项目评分预测的用户偏差,MAE 为系统偏差,其表达式如下:

令 H_u 为用户 u 评分且可以获得评分预测的项目集合,$H_u=\{i\in I|r_{u,i}\neq null\wedge p_{u,i}\neq null\}$,则

$$MAE_u=\frac{1}{num(H_u)}\sum_{i\in H_u}|r_{u,i}-p_{u,i}| \tag{11}$$

$$MAE=\frac{1}{num(U)}\sum_{i\in U}MAE_u \tag{12}$$

3.2 覆盖率(coverage)

除了 MAE 以外,推荐系统中关于预测质量还有另一个评价标准,即覆盖率,在文中作为评价标准用来衡量项目预测的全面性.它反映了未评分项目中能够获得评分预测的项目比例,其值越大,预测越全面.这里,针对用户 u,定义 $Coverage_u$ 为项目评分预测的用户覆盖率,$Coverage$ 为系统覆盖率,其表达式如下:

令 D_u 为用户 u 未评分的项目集合,$D_u=\{i\in I|r_{u,i}=null\}$,则

$$Coverage_u=\frac{num(E_u)}{num(D_u)},\text{当且仅当 } D_u\neq\varnothing \tag{13}$$

$$Coverage=\frac{1}{num(U)}\sum_{i\in U}Coverage_u \tag{14}$$

3.3 推荐精度(precision)与召回率(recall)

推荐系统根据产生的项目评分预测向用户进行推荐,为了衡量项目推荐的准确率和全面率,需要另外定义推荐精度 *Precision* 和召回率 *Recall*.第 2.5 节中已经定义了项目集合 E_u,这里将重新定义 $E_u=\{i\in I|r_{u,i}\neq null\wedge p_{u,i}\neq null\}$,并从中选取 N 个预测值最大的项目构成推荐项目集合 N_u;同时,在此基础上引入推荐阈值 ε,它是一个常数,用来判定 N_u 中的项目是否值得向用户推荐.若项目评分真实值大于等于 ε,则值得向用户 u 推荐该项目,由此类项目构成的集合定义为推荐成功的项目集合 V_u,$V_u=\{i\in N_u|r_{u,i}\geqslant\varepsilon\}$,余下的项目构成的集合定义为推荐失败的项目集合 F_u,$F_u=\{i\in N_u|r_{u,i}<\varepsilon\}$,以及 N_u 以外、E_u 中值得推荐却由于预测的原因没有获得推荐的项目,由此类项目构成的集合定义为遗憾项目集合 W_u,即 $W_u=\{i\in E_u-N_u|r_{u,i}\geqslant\varepsilon\}$.

由此定义 $Precision_u$ 为项目推荐的用户精度,*Precision* 为系统精度,其值越大,推荐准确率越高;$Recall_u$ 为用户召回率,*Recall* 为系统召回率,其值越大,推荐全面率越高,其表达式分别如下:

$$Precision_u=\frac{num(V_u)}{N} \tag{15}$$

$$Precision=\frac{1}{num(U)}\sum_{u\in U}Precision_u \tag{16}$$

$$Recall_u=\frac{num(V_u)}{num(V_u)+num(W_u)} \tag{17}$$

$$Recall=\frac{1}{num(U)}\sum_{u\in U}Recall_u \tag{18}$$

4 实验与结果评估

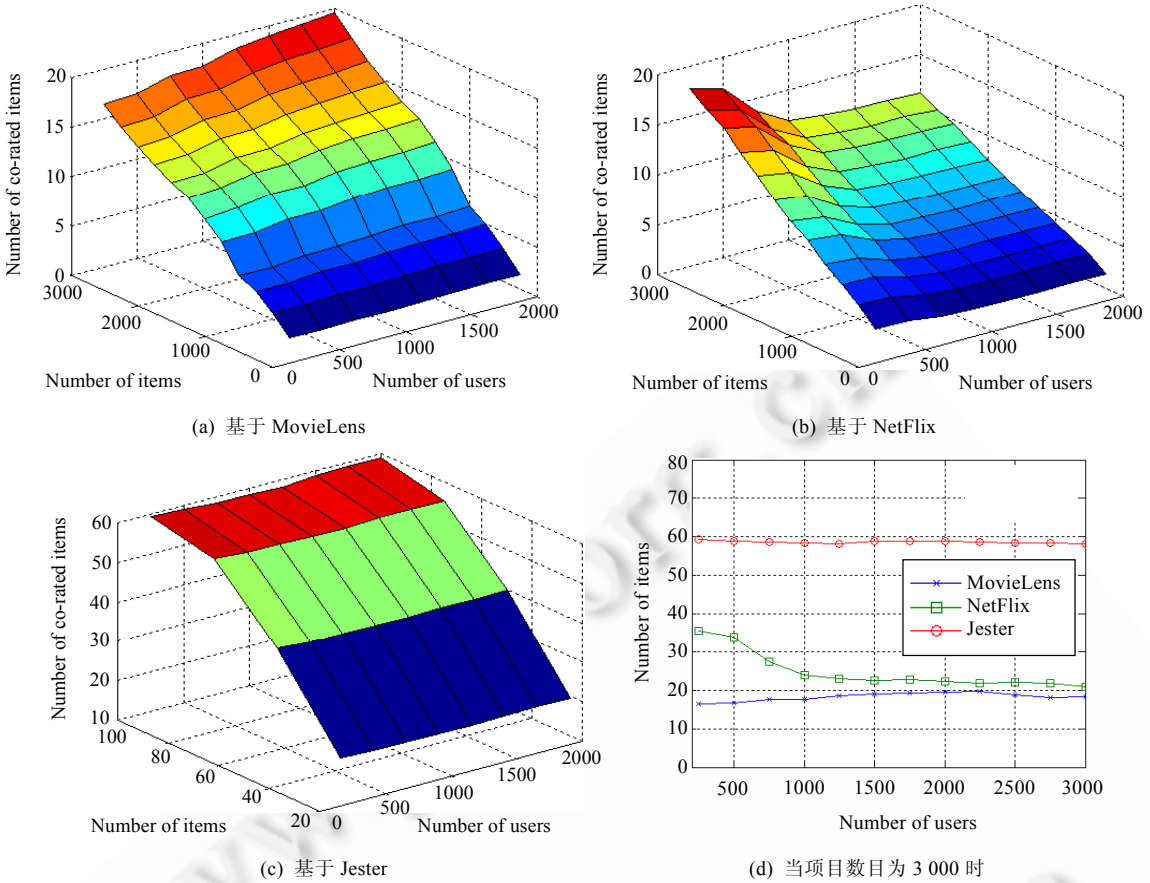
4.1 相似性度量的改进

PC 及其衍生模型在衡量用户相似性方面比其他模型具有更加优越的性能,许多新提出的相似性度量模型将其作为比较对象,用来证明所提出的模型具有更好的性能表现.例如,文献[4]对目前常用的相似性度量(PC,COS,CPC,SPR)进行对比验证,PC 脱颖而出,并将提出的模型 SM 与 PC 在 MAE,Coverage,Precision,Recall 这 4 个方面基于不同的数据集进行性能比较,实验结果表明前者表现更优;文献[5]借鉴二分网络思想,提出了基于多渠道扩散的 Diffusion 模型,比 PC 在 MAE,RMSE(root mean square error)方面和复杂度方面具有优越性.本文在上述两种算法的基础上进行融合,提出了 CFSDP 模型,并将其与 SM,Diffusion 模型进行比较.

值得注意的是,在 CFSDP 模型计算相似性的过程中,相似性值与用户间公共评价项目数目并没有必然的联系,相似性值大的,其公共评价项目数目并不一定大.而根据日常经验,用户间的相似度并不能通过一两次的项目评分来衡量,因为即使两次的评分完全相同,此时相似性值非常大,也有可能出于偶然.鉴于此,有必要利用公共评价项目数目对 CFSDP 模型进行改进,通常的处理方法是,取公共评价项目数目阈值 μ:当公共评价项目数目超过 μ 时,相似性值保持不变;但当小于 μ 时,处理方法有两种:(1) 相似性值随公共评价项目数目减小而减小[30],文中采用 $\frac{num(A_{u,v})}{\mu}\times sim(u,v)$;(2) 相似性值直接为 0.

接下来,本文将验证两种处理方法的优劣,但是需要首先确定阈值 μ,为此需要统计 3 个数据集 MovieLens,NetFlix,Jester 中的用户公共评价项目平均数目,如图 1 所示.

由图 1(a)~图 1(c)可以看出,用户公共评价项目平均数目随项目数目的增加而逐渐增多,而与用户数量的变化关系不大.图 1(d)不仅进一步说明了这一点,而且显示了当项目数目为 3 000 时,MovieLens,NetFlix 和 Jester 的平均公共评价项目平均数目分别约为 17,22,60.根据这些数据,μ 取值不应太大.接下来,本文将对[3,10]间的整数 μ 值进行 MAE 验证,如图 2 所示.



(a) 基于 MovieLens (b) 基于 NetFlix

(c) 基于 Jester (d) 当项目数目为 3 000 时

Fig.1 Averages of $num(A_{u,v})$s

图 1 公共项目平均数 $num(A_{u,v})$s

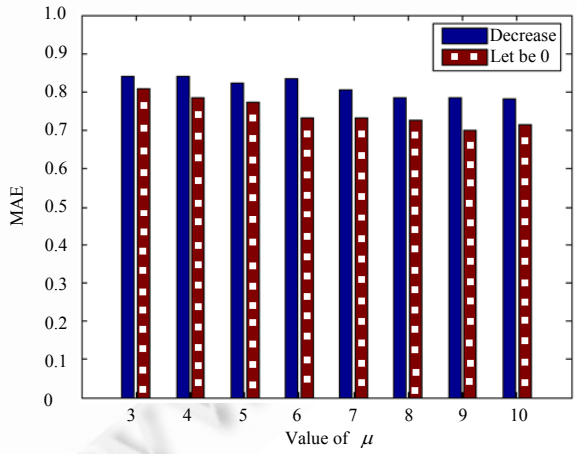


Fig.2 Comparative MAE results based on MovieLens

图 2 基于 MovieLens 的 MAE 比较结果

由图 2 可以看出,方法(2)在不同μ值时,与方法(1)相比均表现更优,因此,本文在实验中会采用方法(1)将用户间公共评价项目数目小于μ的相似性值均置为 0.但方法(1)中,不同μ值的性能差异并不大,由于μ为 9 时表现较

好,同时为了简单,在实验中μ均取 9.

4.2 实验面临的挑战及对策

实验中采用的评分数据集分别为 Movielens,NetFlix 和 Jester,其基本参数见表 4.

Table 4 Basic parameters of data sets

表 4 评分数据集的基本参数

	用户数	项目数	评分数	[m,M]	稀疏性(%)
MovieLens	6 040	3 706	1 000 209	1~5	95.53
NetFlix	480 189	17 770	100 480 507	1~5	98.82
Jester	19 986	100	1 810 455	[−10,10]	9.41

实验过程中,数据集通常分为训练集和测试集两部分,前者用于模型的构建,比重应该大一些,这样获得的模型更加稳定;后者用于模型的测试,检验模型的性能.3 个数据集 MovieLens,NetFlix 和 Jester,本文对于用户,从用户数据集中分别随机选取 60%的用户作为训练集,余下的作为测试集使用.对于项目,由于 MovieLens,Jester 项目数目比较少,分别为 3 706,100,从项目数据集中选择全部项目;而 NetFlix 项目数目较大,随机选择 60%的项目作为训练集使用,同时也作为测试集使用.

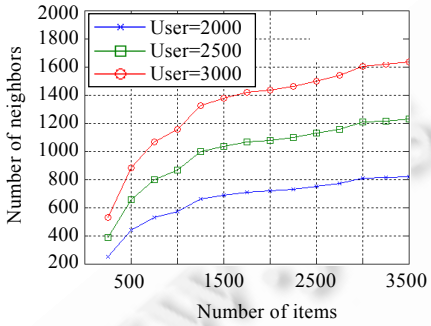


Fig.3 Averages of users' neighbors based on Diffusion

图 3 基于 Diffusion 模型的用户平均近邻

实验通常选取目标用户的近邻个数作为不同模型的比较条件,比如近邻个数为 1 500,2 000,步长 50.但这种做法在本文中显得不太合适,因为对于 Diffusion 模型来说,由于本身对数据稀疏性的要求比较苛刻,尤其是在 MovieLens 中,如图 3 所示,用户平均近邻数目要达到 1 500,2 000 是很困难的.

由图 3 可以看出,Diffusion 模型的平均近邻数目随用户项目数目的增长而呈现准对数增长的趋势,最终,平均近邻数目达到 1 500 是在项目数目为 3 500、用户数目为 3 000 的时候.这说明通常的模型比较条件在用作此时模型之间的比较时面临着挑战.

为此,本文采用用户近邻数目百分数 p 作为项目预测的比较标准,因为其不仅可以满足用户平均近邻数目比较小时的要求,而且同样适用于用户近邻数目参差不齐的情况.在实验中,p 的取值范围为 0.1~0.8,步长为 0.05.例如,当 p=0.12 时,按照相似性值大小,取每个用户近邻总量的前 12%作为参与推荐的推荐近邻集合 K_u.这样做还有另外一个好处:不再因为数据集的变化而需要设置新的参数.同时,为了考察不同的项目推荐数目 N 对推荐结果的影响,本文采用 N 作为项目推荐的比较标准,参照文献[4,25,28]中关于项目推荐参数ε和 N 的取值.这样,实验及实验中需要用到的主要参数见表 5.

Table 5 Main parameters used in experiments

表 5 实验中用到的主要参数

	MAE, Coverage			μ	Precision, Recall			训练用户(%)	训练项目(%)
	p	N	步长		p	N	ε		
MovieLens	[0.1,0.8]	10	0.05	9	0.5	6~20	5	60	100
NetFlix	[0.1,0.8]	10	0.05	9	0.5	6~20	5	60	60
Jester	[0.1,0.8]	10	0.05	9	0.5	6~20	5	60	100

4.3 实验对比与结果分析

实验中,利用 MAE,Coverage,Precision,Recall 这 4 项评价指标对 3 个模型进行比较,分为两部分:

(1) 基本模型比较:基于数据集 MovieLens,NetFlix 和 Jester,如图 4~图 6 所示;

(2) 扩展模型比较:检验所提模型的适应性,基于数据集 Jester,如图 7 所示.

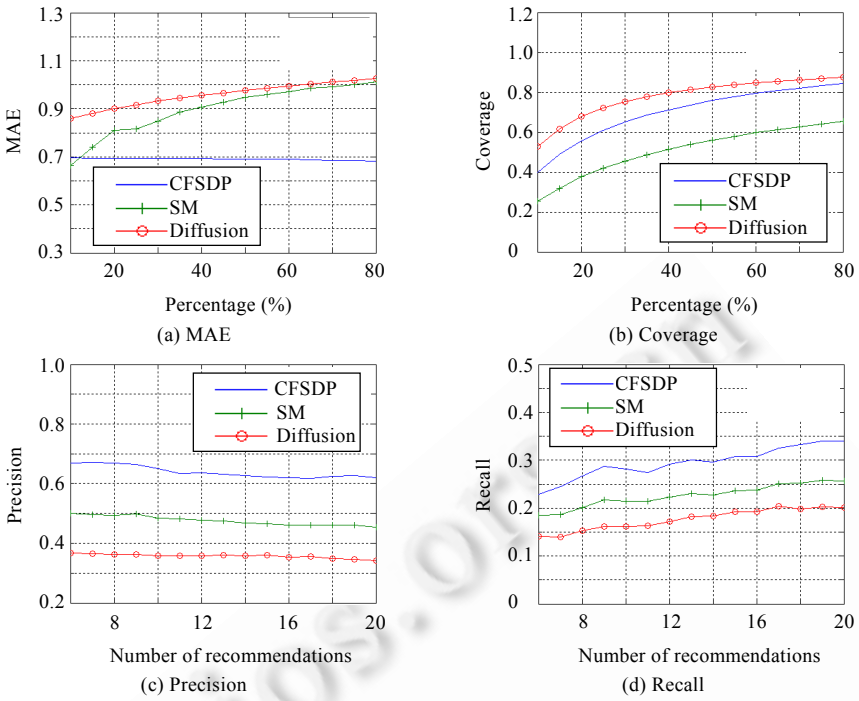


Fig.4 Comparisons based on MovieLens

图 4 基于 MovieLens 的比较

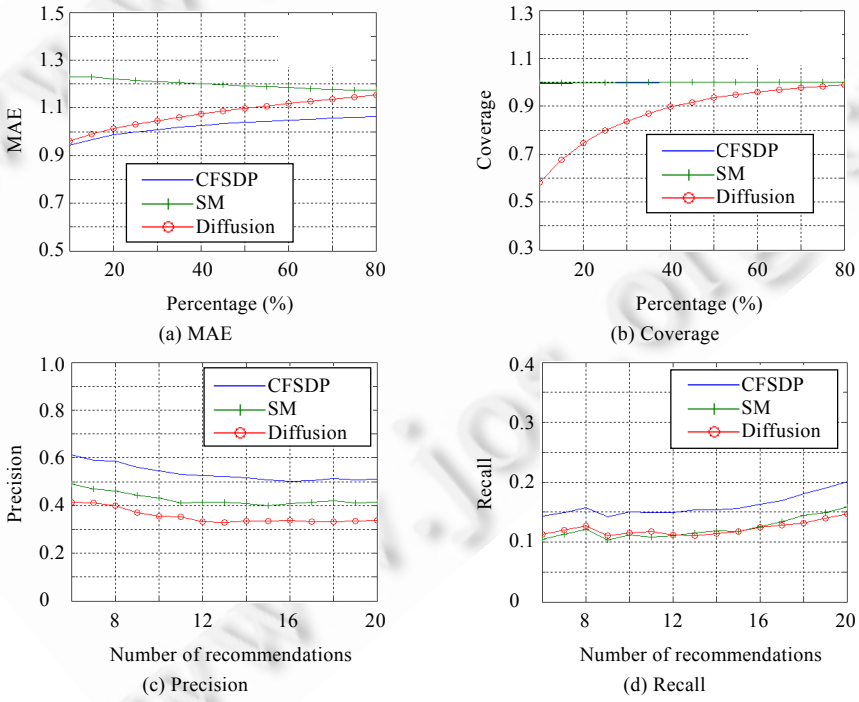
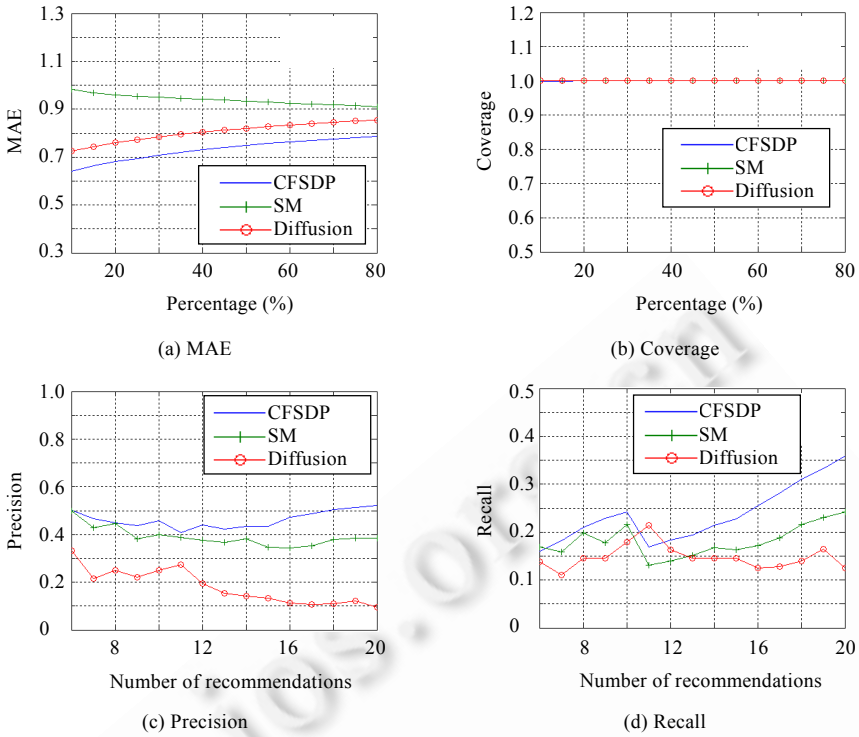


Fig.5 Comparisons based on NetFlix

图 5 基于 NetFlix 的比较



(a) MAE (b) Coverage

(c) Precision (d) Recall

Fig.6 Comparisons based on Jester

图 6 基于 Jester 的比较

4.3.1 基本模型比较

图 4、图 5 分别展示了基于 MovieLens,NetFlix 数据集的 3 种模型比较的实验结果.可以看出,文中提出的 CFSDP 模型与 SM 和 Diffusion 相比,在 4 个指标上,除了 Coverage 在 MovieLens 数据集上比 Diffusion 略逊色以外,均表现出良好的性能,并且随着 p 和 N 值的增大,一直保持着优势.在总体性能上,文中模型占优、Diffusion 较差,这主要是由于本文模型借鉴了 SM 和 Diffusion 的优点,在避免恶化数据稀疏性问题的同时对相似性值做了一定程度的改进.反观 Diffusion,其仅仅考虑了用户之间的相同评分,而对其他评分则置之不理,这不仅严重影响了近邻的选择,而且更加暴露了数据本身的稀疏性问题,从而使得各方面性能下降,在与其他模型的比较中处于劣势.

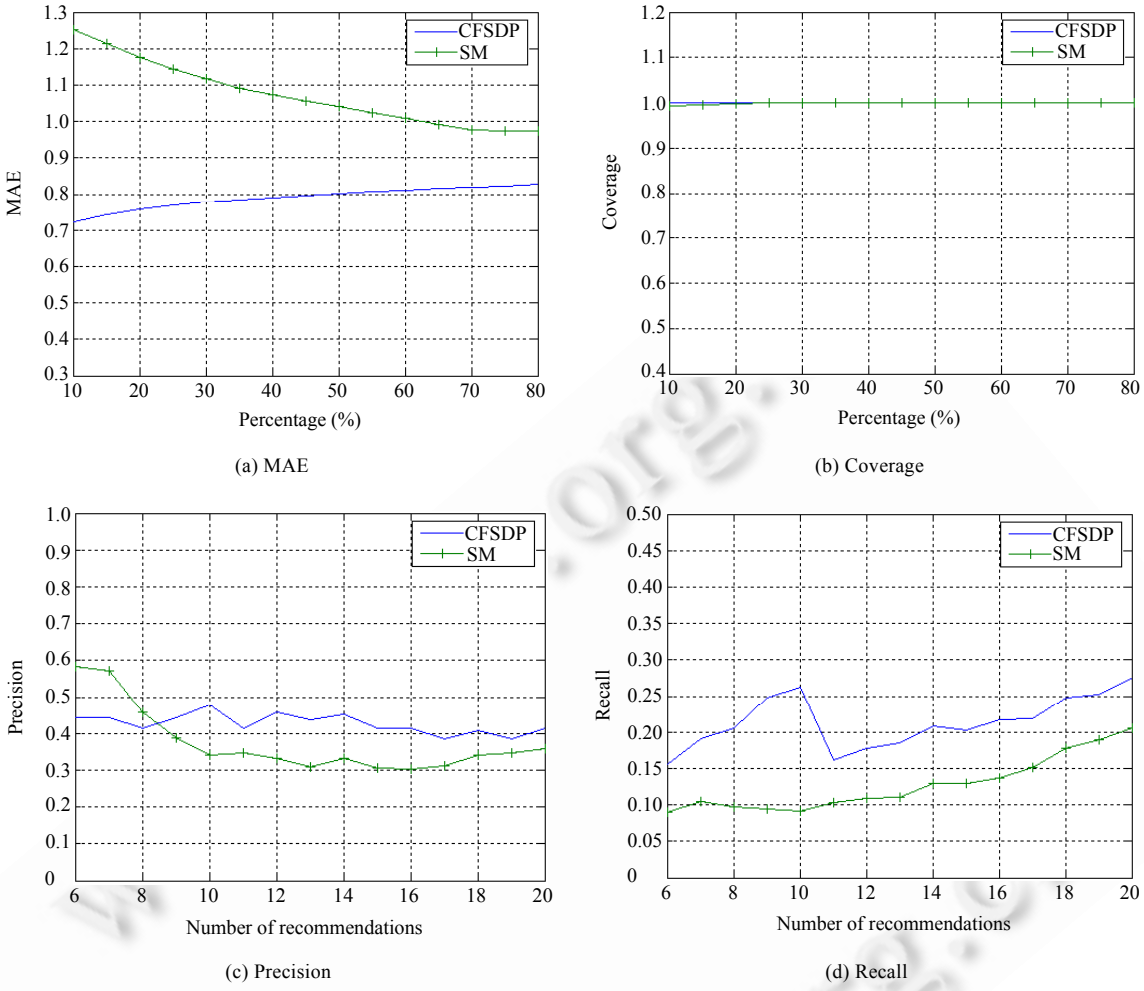
由于 Jester 数据集与其他数据集的不同,原因在于其评分级别范围[−10,10]是连续的,这就要求模型具有一定的适应性.为了做对比实验,实验中需要对 Diffusion 进行一定的改进:并非仅仅考虑相同的项目评分,而是将相同评分的概念扩展为相同范围内的评分,文中对评分进行向大取整,如 9.2 和 9.5 取整为 10,被视为“相同评分”参与相似度计算.图 6 展示了基于 Jester 数据集的 3 种模型比较的实验结果,结果显示:在同等条件下,文中提出的 CFSDP 模型与其他模型相比在 4 个指标上一样表现较好,并且一直保持着优势,某些情况下优势还比较明显,例如,在 MAE 上,相对 SM 和 Diffusion,CFSDP 分别改进了约 10%,23%.

4.3.2 扩展模型比较

由于 CFSDP,SM 两种模型均具有可扩展的能力,以及 Jester 数据集本身的独特性,因此基于 Jester 数据集来检验两种模型扩展后的性能,如图 7 所示.

图 7 展示了两种模型比较的实验结果,可以看出,CFSDP,SM 扩展模型除了在 Coverage 方面相同以外,在其他 3 个方面 MAE,Precision,Recall,前者均比后者表现得稍好一些,这主要是由于前者不仅继承了后者的优点,而且利用公共评分项目进一步优化了相似性值,从而使得模型的性能进一步的提高.这也证明了 CFSDP 模型具有

良好的扩展性.



(a) MAE (b) Coverage

(c) Precision (d) Recall

Fig.7 Comparisons of the extended models

图 7 扩展模型的比较

4.4 时间复杂度比较

上述 4 项指标在一定程度上可以理解为衡量推荐系统准确性和全面性的度量标准,但在实际应用中,还必须考虑另外一个重要因素,即模型的时间复杂度,具体从以下两个阶段进行分析:

(1) 离线阶段:对于 $m\times n$ 阶的二维用户项目矩阵(为方便说明,这里重新定义 m 为用户数目,n 为项目数目),在经过 $2n$ 次的除法操作获得项目奇异性值和用户项目连通度以后,本文提出的模型还需要 $5n$ 次乘法操作和 n 次除法操作,以获得两个用户间的相似度.由于需要计算系统中所有用户之间的相似度且该相似度不具有对称性,所以总共需要计算 $m(m-1)$次.因此,该阶段模型的计算复杂度为 $O(m^2n)$,这个过程需要花费很长的时间,不便于统计.

(2) 在线阶段:在获得用户间的相似度以后,本文模型需要选取百分数 p 的用户作为推荐近邻,然后通过常数次操作为目标用户推荐 N 个项目,因此,该阶段模型的计算复杂度为 $O(1)$.

对于不同的模型,在线阶段的时间花费实际上为用户获得推荐的等待时间,这个时间对于用户来说比较敏

感,对于系统也至关重要.为便于统计比较,本文接下来,将在系统测试集中随机选取不同的 100×1000 数据集,对这段时间进行 10 次记录取均值,以此作为不同模型的运行时间,整理结果见表 6,其中,“–”表示未作该实验.

Table 6 Comparison of running times (s)

表 6 运行时间比较 (秒)

	CFSDP	SM	Diffusion	扩展 CFSDP	扩展 SM
MovieLens	159.5	105.7	177.3	–	–
NetFlix	102.2	119.0	161.5	–	–
Jester	62.2	65.3	57.9	62.6	61.4

可以看出,在不同数据集的比较中,本文模型的运行时间与其他模型相比虽然不占有优势,但差距并不明显,甚至在 NetFlix 中还表现最优.因此,综合模型前面在预测和推荐方面取得的优秀性能,CFSDP 在运行时间开销上是合理的.

5 总结及未来工作

推荐系统中包含了大量的上下文信息,传统的相似性度量模型例如 PC,并没有对其充分发掘并加以利用,从而为模型的不断改进提供了空间,如当前的多渠道扩散过程模型和奇异性模型,就是充分利用评分内在信息的成功范例.本文以此作为借鉴,提出了融合奇异性和扩散过程的协同过滤模型 CFSDP,用来度量用户之间的相似性.并基于 MovieLens,NetFlix,Jester 这 3 个数据集,利用不同指标将所提出的模型与原来的模型进行对比.实验结果表明,本文模型在 MAE,Coverage,Precision,Recall 这 4 个方面获得了改进,并且具有良好的扩展性.值得一提的是,CFSDP 的时间花费并不高,在某些时候还是最优的.

需要说明的是,本文模型还有许多需要改进的地方:(1) 评分级别范围的划分问题.文中仅对两类具有代表性的评分级别范围进行了简单划分,实际应用中并非如此,因此,未来需要对评分级别范围划分问题进行专门的论证,以求获得更优的划分方案.(2) 项目评分的更新问题.文中在计算用户相似性值时,相当于仅仅考虑了 1 个周期内项目评分没有发生变化的情形,而没有考虑在不同周期间项目评分动态性的特点,这与实际情况是不是完全相符的.未来的工作将把这一因素纳入到模型中,从而为模型的实际应用起到促进作用.

References:

[1] Jiang F, Gao M. Collaborative filtering approach based on item and personalized contextual information. In: Yu F, ed. Proc. of the 2009 Int'l Symp. on Intelligent Information Systems and Applications. Academy Publisher, 2009. 63−66.

[2] Xiao RL, Hong FL, Xiong JB, Zheng XJ, Zhang ZQ. Syncretizing context information into the collaborative filtering recommendation. In: Chen JX, ed. Proc. of the 1st Int'l Workshop on Database Technology and Applications. IEEE Computer Society, 2009. 33−36. [doi: 10.1109/DBTA.2009.57]

[3] Yao Z, Wu Y, Chang N. Collaborative filtering recommender algorithm for integrating item category and contextual information. Computer Integrated Manufacturing Systems, 2008,14(7):1449−1456 (in Chinese with English abstract).

[4] Bobadilla J, Ortega F, Hernando A. A collaborative filtering similarity measure based on singularities. Information Processing and Management, 2011,48(2):204−217. [doi: 10.1016/j.ipm.2011.03.007]

[5] Shang MS, Jin CH, Zhou T, Zhang YC. Collaborative filtering based on multi-channel diffusion. Physica A—Statistical Mechanics and Its Applications, 2009,388(23):4867−4871. [doi: 10.1016/j.physa.2009.08.011]

[6] Xu HL, Wu X, Li XD, Yan BP. Comparison study of Internet recommendation system. Ruan Jian Xue Bao/Journal of Software, 2009,20(2):350−362 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3388.htm [doi: 10.3724/SP.J.1001.2009.03388]

[7] Wang HF, Wu CT. A strategy-oriented operation module for recommender systems in e-commerce. Computers & Operations Research, 2012,39(8):1837−1849. [doi: 10.1016/j.cor.2010.03.011]

[8] Cai XC, Bain M, Krzywicki A, Wobcke W, Kim YS, Compton P, Mahidadia A. Collaborative filtering for people to people recommendation in social networks. In: Li JY, ed. Advances in Artificial Intelligence (Ai 2010). Berlin: Springer-Verlag, 2010. 476−485. [doi: 10.1007/978-3-642-17432-2_48]

[9] Kwon HJ, Hong KS. Personalized smart TV program recommender based on collaborative filtering and a novel similarity method. IEEE Trans. on Consumer Electronics, 2011,57(3):1416−1423. [doi: 10.1109/TCE.2011.6018902]

[10] Ghauth KI, Abdullah NA. Learning materials recommendation using good learners' ratings and content-based filtering. Educational Technology Research and Development, 2010,58(6):711−727. [doi: 10.1007/s11423-010-9155-4]

[11] Uddin MN, Shrestha J, Jo GS. Enhanced content-based filtering using diverse collaborative prediction for movie recommendation. In: Nguyen NT, eds. Proc. of the 2009 1st Asian Conf. on Intelligent Information and Database Systems. IEEE, 2009. 132−137. [doi: 10.1109/ACIIDS.2009.77]

[12] Nguyen AT, Denos N, Berrut C. Improving new user recommendations with rule-based induction on cold user data. In: Proc. of the 2007 ACM Conf. on Recommender Systems (Recsys 2007). ACM Press, 2007. 121−128. [doi: 10.1145/1297231.1297251]

[13] Chun J, Oh JY, Kwon S, Kim D. Simulating the effectiveness of using association rules for recommendation systems. In: Baik DK, ed. Proc. of the Systems Modeling and Simulation: Theory and Applications. Berlin: Springer-Verlag, 2005. 306−314. [doi: 10.1007/978-3-540-30585-9_34]

[14] Qiu LY, Benbasat I. A study of demographic embodiments of product recommendation agents in electronic commerce. Int'l Journal of Human-Computer Studies, 2010,68(10):669−688. [doi: 10.1016/j.ijhcs.2010.05.005]

[15] Chen T, He L. Collaborative filtering based on demographic attribute vector. In: Luo Q, eds. Proc. of the 2009 ETP Int'l Conf. on Future Computer and Communication. IEEE Computer Society, 2009. 225−229. [doi: 10.1109/FCC.2009.68]

[16] Jia CX, Liu RR, Sun D, Wang BH. A new weighting method in network-based recommendation. Physica A-Statistical Mechanics and Its Applications, 2008,387(23):5887−5891. [doi: 10.1016/j.physa.2008.06.046]

[17] Zhou T, Ren J, Medo M, Zhang YC. Bipartite network projection and personal recommendation. Physical Review E, 2007,76(4): 1−7. [10.1103/PhysRevE.76.046115]

[18] Yamashita A, Kawamura H, Suzuki K. Adaptive fusion method for user-based and item-based collaborative filtering. Advances in Complex Systems, 2011,14(2):133−149. [doi: 10.1142/S0219525911003001]

[19] Liu ZB, Qu WY, Li HT, Xie CS. A hybrid collaborative filtering recommendation mechanism for P2P networks. Future Generation Computer Systems, 2010,26(8):1409−1417. [doi: 10.1016/j.future.2010.04.002]

[20] Barragans-Martinez AB, Costa-Montenegro E, Burguillo JC, Rey-Lopez M, Mikic-Fonte FA, Peleteiro A. A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition. Information Sciences, 2010,180(22):4290−4311. [doi: 10.1016/j.ins.2010.07.024]

[21] Pham MC, Cao YW, Klamma R, Jarke M. A clustering approach for collaborative filtering recommendation using social network analysis. Journal of Universal Computer Science, 2011,17(4):583−604. [doi: 10.3217/jucs-017-04-0583]

[22] Liu FK, Lee HJ. Use of social network information to enhance collaborative filtering performance. Expert Systems with Applications, 2010,37(7):4772−4778. [doi: 10.1016/j.eswa.2009.12.061]

[23] Kagie M, van der Loos M, van Wezel M. Including item characteristics in the probabilistic latent semantic analysis model for collaborative filtering. AI Communications, 2009,22(4):249−265. [doi: 10.3233/AIC-2009-0467]

[24] Bobadilla J, Hernando A, Ortega F, Bernal J. A framework for collaborative filtering recommender systems. Expert Systems with Applications, 2011,38(12):14609−14623. [doi: 10.1016/j.eswa.2011.05.021]

[25] Bobadilla J, Hernando A, Ortega F, Gutierrez A. Collaborative filtering based on significances. Information Sciences, 2012,185(1): 1−17. [doi: 10.1016/j.ins.2011.09.014]

[26] Zhang GW, Li DY, Li P, Kang JC, Chen GS. A collaborative filtering recommendation algorithm based on cloud model. Ruan Jian Xue Bao/Journal of Software, 2007,18(10):2403−2411 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/2403.htm [doi: 10.1360/jos182403]

[27] Ahn HJ. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. Information Sciences, 2008,178(1):37−51. [doi: 10.1016/j.ins.2007.07.024]

[28] Bobadilla J, Serradilla F, Bernal J. A new collaborative filtering metric that improves the behavior of recommender systems. Knowledge-Based Systems, 2010,23(6):520−528. [doi: 10.1016/j.knosys.2010.03.009]

[29] Hofmann T. Latent semantic models for collaborative filtering. ACM Trans. on Information Systems, 2004,22(1):89−115. [doi: 10.1145/963770.963774]

[30] Kwon K, Cho JY, Park Y. Multidimensional credibility model for neighbor selection in collaborative recommendation. Expert Systems with Applications, 2009,36(3):7114−7122. [doi: 10.1016/j.eswa.2008.08.071]

附中文参考文献:

[3] 姚忠,吴跃,常娜.集成项目类别与语境信息的协同过滤推荐算法.计算机集成制造系统,2008,14(7):1449−1456.

[6] 许海玲,吴潇,李晓东,阎保平.互联网推荐系统比较研究.软件学报,2009,20(2):350−362. http://www.jos.org.cn/1000-9825/3388.htm [doi: 10.3724/SP.J.1001.2009.03388]

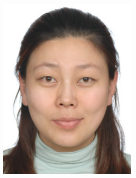
[26] 张光卫,李德毅,李鹏,康建初,陈桂生.基于云模型的协同过滤推荐算法.软件学报,2007,18(10):2403−2411. http://www.jos.org.cn/1000-9825/18/2403.htm [doi: 10.1360/jos182403]

杨兴耀(1984－),男,湖北襄阳人,博士生,CCF 学生会员,主要研究领域为推荐系统,网格计算与云计算,可信计算.
E-mail: yangxy@xju.edu.cn

于炯(1964－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为网络安全,网格与分布式计算.
E-mail: yujiong@xju.edu.cn

吐尔根·依布拉音(1958－),男,博士,教授,博士生导师,主要研究领域为计算机应用,自然语言处理,软件工程.
E-mail: turgun@xju.edu.cn

廖彬(1986－),男,博士生,CCF 学生会员,主要研究领域为数据库,网格与云计算.
E-mail: liaobin665@163.com



钱育蓉(1980－),女,博士,副教授,CCF 会员,主要研究领域为遥感图像处理,模式识别,数据挖掘.
E-mail: qyr@xju.edu.cn