

杰卡德相似系数在推荐系统中的应用

张晓琳,付英姿,褚培肖

(昆明理工大学 理学院,云南 昆明 650500)

摘要:项目相似性度量是协同过滤系统的核心。相关研究中,基于物品协同过滤系统的相似性度量方法普遍使用余弦相似性。然而,在许多实际应用中,评价数据稀疏度过高,物品之间通过余弦相似度计算会产生误导性结果。文中将杰卡德相似性度量应用到基于物品的协同过滤系统中,并建立起相应的评价分析方法。与传统相似性度量方法相比,杰卡德方法完善了余弦相似性只考虑用户评分而忽略了其他信息量的弊端,特别适合于文中所应用的稀疏度过高的数据。最后通过实例说明上述方法的有效性。

关键词:推荐系统;协同过滤;杰卡德相似系数;余弦相似性

中图分类号:TP302.1

文献标识码:A

文章编号:1673-629X(2015)04-0158-04

doi:10.3969/j.issn.1673-629X.2015.04.036

Application of Jaccard Similarity Coefficient in Recommender System

ZHANG Xiao-lin, FU Ying-zi, CHU Pei-xiao

(Department of Science, Kunming University of Science and Technology,
Kunming 650500, China)

Abstract: Item similarity metric is the core of collaborative filtering system. In related research, cosine similarity has been regularly used in item-based collaborative filtering system. However, in many practical cases, cosine similarity leads to misleading results due to high sparsity. Jaccard similarity is used in item-based collaborative filtering in this article to build the corresponding evaluation and analysis methods. Compared with traditional similarity metric, Jaccard method improves the drawbacks of cosine similarity, which only takes user rating in consideration and ignores other information, and is suitable for considering the high sparse data. In the end, an example is used to prove the effectiveness of Jaccard similarity.

Key words: recommender systems; collaborative filtering; Jaccard similarity coefficient; cosine similarity

1 概述

推荐系统从 20 世纪 90 年代发展到今天已成为重要研究内容,得到越来越多的关注^[1]。目前主要的方法有:协同过滤推荐、基于内容的推荐及混合推荐^[2-3]。协同过滤是推荐系统中应用最多的技术之一。基于用户的协同过滤算法(UBCF)于 1992 年被提出并应用于邮件过滤系统,此后直到 2000 年都是该领域最流行的算法。UBCF 基本思想为:找到和目标用户兴趣相似的用户集合。并在这个集合中将喜欢但没有听说过的项目推荐给目标用户^[4]。但随着网站的用户数目越来越大,导致时间复杂度与空间复杂度的平方增长,且不能合理解释其推荐结果^[5]。因此,电子

商务公司亚马逊提出了另一种算法—基于物品的协同过滤算法(BCF)^[6]。但不足的是,BCF 中传统的相似性度量法只关注于用户评分,而忽略了其他信息量,例如评价商品的覆盖范围以及数目等^[7-8]。

Jaccard^[9]于在 1912 年提出杰卡德相似性度量方法并用于分析高山区的区系分布。Tan 等^[10]将此方法引入协同过滤系统并引入神经网络对其进行修正。近年来,新的研究思路及方法层出不穷,Deepa Anand 等^[11]提出利用稀疏数据来建立用户的局部和全局相似度。与传统相似性度量方法相比,杰卡德方法完善了余弦相似性只考虑用户评分而忽略了其他信息量的弊端,特别适合于文中所应用的稀疏度过高的数据^[12]。

收稿日期:2014-05-19

修回日期:2014-08-25

网络出版时间:2015-02-23

基金项目:国家自然科学基金资助项目(11201200)

作者简介:张晓琳(1989-),女,硕士研究生,CCF 会员,研究方向为多元统计分析;付英姿,副教授,研究生导师,研究方向为应用统计。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20150223.1233.011.html>

2 基于物品的协同过滤算法

基于物品的协同过滤算法基本思想可以理解为:通过分析用户的行为计算物品的相似程度,给用户推荐与他们之前喜欢的物品相似的物品^[13]。此算法主要步骤为:首先,考察喜欢项目 $A_i (i = 1, 2, \dots, n)$ 的目标用户,计算与第 i 个目标项目相似度 $\{s_{i1}, s_{i2}, \dots, s_{ik}\}$ 最高的 k 个项目 $\{i_1, i_2, \dots, i_k\}$; 然后,通过对这些类似项目计算加权平均生成推荐列表。下面分别详细介绍相似度计算和生成预测。

2.1 项目相似性度量

传统 IBCF 中一个步骤是计算项目之间的相似度,如果不同用户对不同的项目欣赏程度一致,则项目的评分大致相同,那么认为这些项目相似程度高。相似性度量方法有很多,普遍使用的是余弦相似性。定义为:把用户评分看做 n 维向量,相似度为计算向量间的夹角余弦。项目 i 与 j 的相似度记为:

$$\text{sim}(i, j) = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{\|\vec{i}\| * \|\vec{j}\|} \quad (1)$$

其中, $\vec{i} \cdot \vec{j}$ 为 \vec{i} 与 \vec{j} 的内积; $\|\vec{i}\| * \|\vec{j}\|$ 为 \vec{i} 与 \vec{j} 的乘积。

除了余弦相似性外还有相关相似性:这种方法通过计算项目 i 与 j 的 Pearson 相关系数 $\text{corr}_{i,j}$ 衡量相似度。为了使计算精确必须先确定共同给 i 与 j 评分的用户集合 U , 相关相似性定义为:

$$\text{sim}(i, j) = \frac{\sum_{u \in U} (R_{ui} - \bar{R}_i)(R_{uj} - \bar{R}_j)}{\sqrt{\sum_{u \in U} (R_{ui} - \bar{R}_i)^2} \sqrt{\sum_{u \in U} (R_{uj} - \bar{R}_j)^2}} \quad (2)$$

其中, R_{ui} 为第 u 个用户的第 i 个项目评分; \bar{R}_i 为 u 个用户给第 i 个项目的平均评分。

基于物品协同过滤算法中的余弦相似性有一大缺点,没有考虑不同用户的不同评分尺度,修正的余弦相似性改善了上述问题,减去了用户对项目的平均评分。相似性记为:

$$\text{sim}(i, j) = \frac{\sum_{u \in U} (R_{ui} - \bar{R}_u)(R_{uj} - \bar{R}_u)}{\sqrt{\sum_{u \in U} (R_{ui} - \bar{R}_u)^2} \sqrt{\sum_{u \in U} (R_{uj} - \bar{R}_u)^2}} \quad (3)$$

其中, \bar{R}_u 为第 u 个用户评分的均值。

2.2 预测计算

在协同过滤系统中,生成并输出预测结果是重要的一步^[14]。只要通过相似度计算得到相似项目集,下一步就是确定目标用户评分并可以使用以下两种方法获得预测。

2.2.1 权重法

权重法为通过计算用户对与第 i 个项目相似项目的评分的权重和,得到给用户 u 关于项目 i 的预测。并且每一个评分被赋予项目 i 与 j 的相似度 s_{ij} 作为权重。如果与目标项目相似的项目数为 $k = 1, 2, \dots, N$ 个,则预测结果表示如下:

$$P_{ui} = \frac{\sum_{k=1}^N s_{ik} * R_{uk}}{\sum_{k=1}^N |s_{ik}|} \quad (4)$$

基本上,此种方法基于用户对相似项目的打分为。权重法基本指标为相似项目权重和,这样便确保了预测结果在预先确定范围内。

2.2.2 回归法

回归法与权重法类似,但是取代直接使用相似项目的评分,而改用基于回归模型得到的近似评分值。在实际操作中,使用余弦或者相关相似性得到的相似度,可能由于两个评分向量在欧氏空间中距离很远而得到相似度很高的误导性结果。回归法的基本思想和计算方法与权重法类似,但是取代第 u 个用户的第 k 个项目评分 R_{uk} , 使用线性回归方程得到的近似值 \bar{R}_{uk} 。将目标项目 i 和相似项目 N 相应的向量记为 R_i 和 R_N , 则线性回归方程表示如下:

$$\bar{R}_N = \alpha \bar{R}_i + \beta + \varepsilon \quad (5)$$

其中,参数 α 和 β 由两个评分向量决定; ε 为回归方程误差项。

与 UBCF 相比,IBCF 不会因为用户数目的增长,而导致时间复杂度与空间复杂度的平方增长,并合理解释了推荐结果。但不足的是,IBCF 中传统的相似性度量法只关注于用户评分,而忽略了其他信息量,例如评价商品的覆盖范围以及数目等。文中在 IBCF 基础上考虑杰卡德相似性作为相似性度量方法,并期待得到更优结果。

3 杰卡德相似性

考虑以下情况:如果图书 R_1 被 15 个用户评价,图书 R_2 被 32 个用户评价,图书 R_3 被 100 个用户评价, R_1 和 R_2 被其中的 10 个用户给出了同样的评价, R_1 和 R_3 被其中的 12 个用户给出了同样的评价。那么,虽然 R_1 和 R_3 同样的评价数多于 R_1 和 R_2 ,但是 R_2 和 R_1 具有相似评价的数目占总数的比例远大于 R_3 ,于是认为 R_1 和 R_2 的相似度其实要高于 R_1 和 R_3 。在计算余弦相似性时,忽略了上述情况的讨论。为了完善此类问题,文中将基于杰卡德指数修正 IBCF 算法。

杰卡德指数,又称杰卡德相似性系数,是用于比较有限样本集的相似性与差异性的统计量。存在两个集

合 A, B , 杰卡德指数记为 $J(A, B)$, 定义如下:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (6)$$

如果集合 A, B 均为空集, 定义 $J(A, B) = 1$ 。显然 $0 \leq J(A, B) \leq 1$ 。J值越大, 两样本相似性越大。

杰卡德距离, 用于测量样本集之间的差异性, 与杰卡德指数互补, 记为 $d_J(A, B)$, 定义如下:

$$d_J(A, B) = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \quad (7)$$

杰卡德距离的另一种解释为: 对称差占两个集合并集的比例。

给定两个集合 A, B , 均含有 n 个共同的属性, 每个属性取值为0或1。杰卡德系数是测评此类数据的有效工具。文中定义: M_{11} 为集合 A 与 B 的属性值都是1的个数; M_{01} 为样本 A 的属性值是1, 且样本 B 的属性值是0的个数; M_{10} 为样本 A 的属性值是0, 且样本 B 的属性值是1的个数; M_{00} 为样本 A 与 B 的属性值都是0的个数。则有如下结论:

$$\begin{aligned} M_{11} + M_{01} + M_{10} + M_{00} &= n \\ J &= \frac{M_{11}}{M_{11} + M_{01} + M_{10}} \\ d_J &= \frac{M_{01} + M_{10}}{M_{11} + M_{01} + M_{10}} \end{aligned} \quad (8)$$

文中将基于杰卡德相似系数的推荐算法记为LBCF。

4 评价指标

推荐系统评价指标有很多种, 涵盖了评价推荐系统的各方面性能。指标中的一些用于定性描述, 一些用于定量计算, 还有用于离线实验计算, 有些只能在线测评。

预测准确度量是重要的离线测评指标。文中使用的是预测准确度量中的评分预测和TopN推荐。

4.1 评分预测

在推荐系统的研究中, 评价系统质量使用的标准通常可以分为两类: 统计精度度量和决策支持精度度量。统计精度度量通过比较推荐评分和实际用户评分评估一个系统的精确度。预测的与实际的用户评分的平均绝对误差 (Mean Absolute Error, MAE) 是普遍使用的指标。令用户 u 对项目 i 的实际评分值记为 p_i , 预测评分值记为 q_i , 则MAE的定义为:

$$MAE = \frac{\sum_{i=1}^N |p_i - q_i|}{N} \quad (9)$$

均方根误差 (RMSE) 也经常用于统计精度度量。定义为:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (p_i - q_i)^2}{N}} \quad (10)$$

MAE与RMSE的值越小, 准确性越好, 推荐质量越高。

4.2 TopN推荐

系统对用户推荐时会提供一个因人而异的推荐列表, 被称为TopN推荐。此推荐的预测准确率指标为召回率 (Recall) 和准确率 (Precision)。如果对用户 u 推荐 N 个物品 $R(u)$, 预测用户喜欢的物品集合为 $T(u)$, 召回率和准确率分别定义为:

$$Recall = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |T(u)|} \quad (11)$$

$$Precision = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |R(u)|} \quad (12)$$

召回率描述的是有多少比例的用户—物品评分记录包含在最终的推荐列表中, 而准确率描述最终的推荐列表中有多少比例是发生过的用户—物品评分记录。

5 实例分析

5.1 数据集

文中采用的图书漂流 (BX) 数据包含278 858名带有基本信息的匿名用户对271 379册图书提供的1 149 780次 (显/隐性) 评分。BX数据量过大, 文中选取1 093名用户对1 144册图书的24 100条有效评分数据作为实验数据集。在实验中将1 093名用户的80%作为训练集, 20%作为测试集。

基于正态原则, 大部分用户对图书的评分数值介于5~8分, 如对某册图书特别的感兴趣, 用户会给9~10分, 特别不感兴趣会给1~4分。利用杰卡德相似性系数描述两用户的相似度, 首先需要对图书评分进行预处理。

分数在1~4分范围认为对该图书评价不高, 统一用0值表示;

分数在5~10分认为喜欢该图书, 用1值表示。

此时, 某用户对图书的评价可以用0和1的集合来表示^[15]。

5.2 结果分析

对图书数据分别使用UBCF、IBCF及LBCF三种模型进行评分预测, 得到MAE及RMSE值如表1所示, 统计精度直方图如图1所示。

表1 模型统计精度度量

	MAE	RMSE
LBCF	1.052 997	1.618 945
IBCF	1.101 242	1.733 533
UBCF	1.212 734	1.636 076

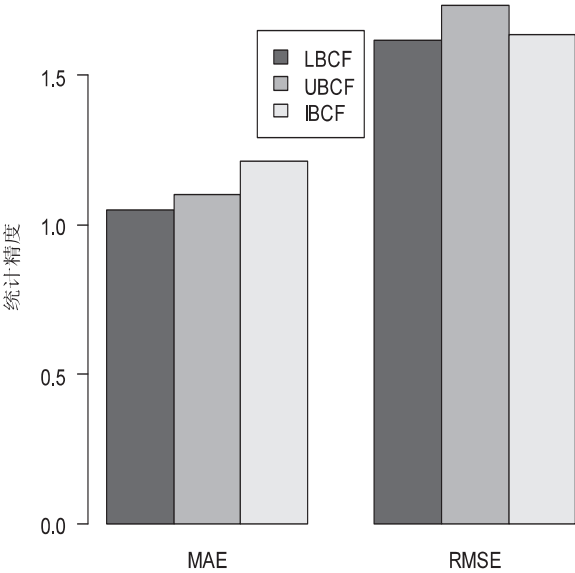


图1 模型统计精度直方图

如图1所示,UBCF、IBCF及LBCF算法的MAE值顺序降低,但差别不明显。其中,LBCF的MAE值分别比UBCF及IBCF低0.160和0.048。而这三种算法的RMSE值基本持平的情况下,LBCF的度量值略低于其他两种算法0.017及0.115。显然,LBCF预测精度最高。值得注意的是,由于数据的稀疏程度过高,通过余弦相似算法得出的物品相似度可能不准确,IBCF的均方根误差要高于UBCF。

在推荐的物品数量不同的情况下,LBCF、UBCF和IBCF的召回率和精确度如表2和表3所示。折线图如图2所示

表2 LBCF/IBCF/UBCF在不同N下的Recall值 %

N	LBCF	IBCF	UBCF
1	0.255	0.113	0.623
2	0.764	0.226	0.990
3	1.132	0.368	1.415
4	1.613	0.659	1.669
5	1.924	0.792	1.952
6	2.462	1.019	2.235
7	2.971	1.217	2.462
8	3.480	1.330	2.688
9	3.877	1.500	2.830
10	4.358	1.613	3.056

表3 LBCF/IBCF/UBCF在不同N下的Precision值 %

N	LBCF	IBCF	UBCF
1	4.265	1.896	10.092
2	6.398	1.896	8.028
3	6.319	2.054	7.645
4	6.754	2.725	6.766
5	6.445	2.655	6.330
6	6.872	2.844	6.040
7	7.114	2.911	5.701
8	7.300	2.784	5.447
9	7.233	2.791	5.097
10	7.323	2.701	4.954

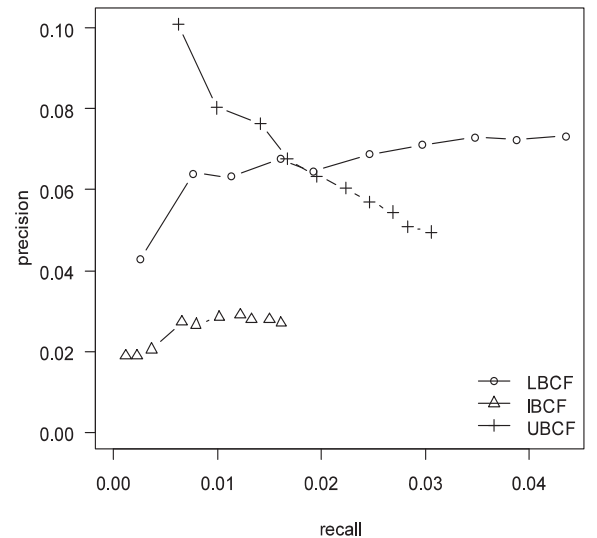


图2 三种推荐算法的准确率/召回率折线图

从图中可以看出,IBCF的推荐效果不如UBCF,这是因为此次实验的数据稀疏度过高,通过余弦相似度算法不能有效地计算出物品之间的相似程度。当推荐数量较少时,UBCF的准确率和召回率要优于LBCF,这是因为单一用户的评价数据过少,依靠相似用户的偏好可以得出精确的推荐,但是当推荐数量增加时,用户兴趣相似度就不能精确地表示出单一用户的偏好,所以基于物品的协同过滤算法要好于用户的协同过滤算法,LBCF明显要优于UBCF。从推荐的评价结果来看,在此次实验所采用数据集稀疏度较高的情况下,基于杰卡德相似系数的相似度度量方法要优于基于余弦相似性的度量方法。主要表现在评价数据稀疏度过高,依靠杰卡德相似算法进行推荐的效果更加准确,可满足推荐系统初期数据不足时的精准推荐。

6 结束语

文中一改传统的余弦相似算法,使用适合数据特
(下转第165页)

包、驱动程序以及系统自检软件。

固态存储模块卸载软件运行在卸载装置的工控机上,为用户提供固态存储模块的文件读写功能,将记录的数据下载到本地计算机或通过以太网 TCP/IP 协议转存到网络存储器上^[13]。

该配套软件的具体设计另文描述,不是文中重点,因此不再赘述。

4 结束语

鉴于对机载 FC 设备的采集记录测试等配套设备的迫切需求,文中提出了一种 FC 记录仪的设计方案,并详述了系统架构以及实现。该记录仪将 FC 协议处理、高速数据采集、海量数据记录等复杂功能高度融为一体,采用全硬件电路方式将多路高速数据记录和采集功能进行分立和隔离处理,最大程度地提高了系统带宽,增强了系统的灵活性,填补了国内空白。

参考文献:

[1] ANSI. Fiber Channel Framing and Signaling-2 (FC-FS-2) Rev0.01[S]. USA:ANSI,2003.

[2] ANSI. Fiber Channel Physical and Signaling Interface (FC-PH) X3[S]. USA:ANSI,1994.

[3] ANSI. Fiber Channel Avionics Environment-Anonymous Sub-

(上接第 161 页)

点的杰卡德相似算法作为推荐模型的物品相似度量算法。并且针对图书的相似评价数据,修正了杰卡德相似算法,以提高推荐的准确度。最后通过一实例说明了该方法的有效性。文中的后续研究工作包括预测物品评分,神经网络计算模型,信息熵在推荐系统中的应用等,希望在不久的将来可以进一步扩展以上结论。

参考文献:

[1] Ricci F,Rokach L,Shapira B,et al. Recommender systems handbook[M]. [s.l.]:Springer,2010.

[2] Hill W,Stead L,Rosenstein M,et al. Recommending and evaluating choices in a virtual community of use [C]//Proc of CHI. [s.l.]:[s.n.],1995:194-201.

[3] 崔春生,李光,吴祈宗. 基于 Vague 集的电子商务推荐系统研究[J]. 计算机工程与应用,2011,47(10):237-239.

[4] 许海玲,吴潇,李晓东,等. 互联网推荐系统比较研究[J]. 软件学报,2009,20(2):350-362.

[5] Bobadilla J,Ortega F, Hernando A. A collaborative filtering similarity measure based on singularities[J]. Information Processing and Management,2012,48:204-217.

scriberMessaging (FC-AE-ASM), Rev1. 2[S]. USA:ANSI, 2006.

[4] 田泽,韩炜,蔡叶芳,等. 基于 FC 接口的 SoC 软硬件协同设计验证平台构建与实现[C]//第十三届计算机工程与工艺会议论文集. 西安:西北工业大学出版社,2009.

[5] 李攀,田泽,蔡叶芳,等. 基于 FPGA 的双通道 FC 数据采集卡设计[J]. 计算机技术与发展,2013,23(7):179-182.

[6] 黎小玉,田泽,王泉,等. 基于 SoC_FC 芯片的电源管理系统设计与实现[J]. 计算机技术与发展,2010,20(8):247-249.

[7] 王红春. 基于 FC 的航电数字视频传输技术研究[J]. 计算机技术与发展,2010,20(5):250-252.

[8] 刘鑫,陆文娟. 光纤通道在航空电子环境的应用及关键技术研究[J]. 光通信技术,2006,30(6):55-58.

[9] 廖寅龙,田泽. FC 网络通信中 PCIe 的接口的设计与实现[J]. 航空计算技术,2010,40(4):127-130.

[10] 张志,翟正军,李想. 航空电子光纤通道协议分析与接口卡设计[J]. 测控技术,2010,29(2):99-101.

[11] 卢光军. 一种高速采集记录设备的实现[J]. 计算机应用与软件,2009,26(9):175-176.

[12] 黄浩益,黄栋杉,徐晓飞. 光纤通道技术在航电系统中的应用[J]. 航空电子技术,2005,36(3):9-14.

[13] 曹阳,李文峰,陈震宇,等. 航空发动机试验数据采集分析系统设计与实现[J]. 航空发动机,2010,36(6):36-38.

[6] Greg L,Brent S,York J. Amazon. com recommendations:item-to-item collaborative filtering[J]. IEEE Internet Computing,2003,7(1):76-80.

[7] 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.

[8] 张光卫,李德毅,李鹏,等. 基于云模型的协同过滤推荐算法[J]. 软件学报,2007,18(10):2403-2411.

[9] Jaccard P. The distribution of the flora in the alpine zone[J]. New Phytologist,1912,11(2):37-50.

[10] Tan Pangning, Steinbach M, Kumar V. Introduction to data mining[M]. [s.l.]:Addison Wesley,2005.

[11] Anand D,Bharadwaj K K. Utilizing various sparsity measures for enhancing accuracy of collaborative recommender systems based on local and global similarities[J]. Expert Systems with Applications,2011,38(5):5101-5109.

[12] 李聪,梁昌勇,杨善林. 电子商务协同过滤稀疏性研究:一个分类视角[J]. 管理工程学报,2011,25(1):94-101.

[13] 项亮. 推荐系统实践[M]. 北京:人民邮电出版社,2012.

[14] 张春生,李艳,图雅. 基于属性拓展的数据挖掘预处理技术研究[J]. 计算机技术与发展,2014,24(3):79-81.

[15] 孙青云,王俊峰,赵宗渠,等. 一种基于模拟登录的微博数据采集方案[J]. 计算机技术与发展,2014,24(3):6-10.

杰卡德相似系数在推荐系统中的应用

作者：[张晓琳](#)，[付英姿](#)，[褚培肖](#)，[ZHANG Xiao-lin](#)，[FU Ying-zi](#)，[CHU Pei-xiao](#)

作者单位：[昆明理工大学 理学院, 云南 昆明, 650500](#)

刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2015(4)

引用本文格式：[张晓琳](#). [付英姿](#). [褚培肖](#). [ZHANG Xiao-lin](#). [FU Ying-zi](#). [CHU Pei-xiao](#) [杰卡德相似系数在推荐系统中的应用](#) [期刊论文] - [计算机技术与发展](#) 2015(4)