一种基于用户情境聚类的个性化推荐算法

吴 楠,秦 锋,姜太平
(安徽工业大学 计算机科学与技术学院,安徽 马鞍山 243032)

摘 要:面向基于情境感知的推荐问题,提出一种基于用户情境聚类的个性化推荐算法。该算法利用情境预过滤的思想,首先运用模糊聚类的方法对历史数据集中用户的情境进行聚类,构造与当前用户情境相似度较高的用户集合,再与传统的基于用户的协同过滤算法相结合进行个性化推荐。实验采用公开数据集,结果表明该算法在多维情境信息条件下可用,并且推荐准确度要高于传统协同过滤算法,在聚类粒度不同的情况下对推荐结果也会产生不同的影响。

关键词:情境感知;聚类;个性化推荐
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2014)10-0106-04
doi:10.3969/j.issn.1673-629X.2014.10.025

A Personalized Recommendation Algorithm Based on User Context Clustering

WU Nan,QIN Feng,JIANG Tai-ping
(College of Computer Science and Technology,Anhui University of Technology,
Ma'anshan 243032,China)

Abstract:Towards the problem of recommendation based on context-aware,propose a personalized recommendation algorithm based on user context clustering. This algorithm uses the idea of contextual pre-filtering. A method of fuzzy clustering is used on the users' context in history data set first to construct the user set which is similar with current user context,and then combined with user-based collaborative filtering algorithm for personalized recommendation. The proposed methodology is tested using public datasets and the results show that it can be used in multidimensional context information. The recommendation precision is higher than traditional collaborative filtering algorithm. And find the change of the cluster size has an impact on the recommendation result.

Key words:context-aware;clustering;personalized recommendation

0 引 言

推荐系统的研究致力于缓解由于网络信息爆炸式增长带来的"信息过载"问题"。根据用户所处的情境和兴趣来推荐个性化的服务越来越被学术界和工业界所重视。情境(Context)是可以被用来表征一个实体状态的任何信息,实体可以是人,地点,或者被认为与用户和应用交互相关的物体,包括用户和应用自身。Schmidt 等人将情境分为与人有关的情境和与物理设备相关的情境,即内部情境与外部情境。随着无线通信技术和移动终端技术的不断发展,很大程度上促进了情境感知计算在各行各业中的实际应用。智能手机、掌上电脑等移动终端可以提供丰富的外部情境信息,如时间、地理位置、温度、天气,网速、当前设备性能参数等。用户自身提供的诸如年龄、性别、职业、学历等组成了内部情境信息。从情境信息中挖掘出有价值的信息,为人机交互智能化提供支撑[1],是情境感知计算的研究热点[2]。传统推荐系统建立在 U(用户)×I(项目)→R(评分)基础上。面向不同领域的项目推荐时,依靠传统的推荐算法往往不能产生较好的推荐效果。学者 Admavicius 和 Tuzhilin 将情境信息引入到传统的推荐系统中[3],将 $U\times I$ 扩展为 $U\times C\times I$ 模型,提出了情境感知推荐系统(CARS)这一概念。国内外的研究学者对如何合理利用情境信息以提高推荐系统的精确度做了大量的工作。文献[4]采用求余弦相似度的

收稿日期:2013-11-19 **修回日期**:2014-02-25 **网络出版时间**:2014-07-28
基金项目:安徽省教育自然科学重点项目(KJ2011A039)
作者简介:吴 楠(1989-),男,硕士研究生,研究方向为移动学习、数据挖掘;秦 锋,教授,研究方向为人工智能、机器学习、数据挖掘;姜太平,副教授,CCF 高级会员,研究方向为模式识别与图像处理。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140728.1221.004.html

方法,将当前情境与历史情境进行比较,并利用历史用户的偏好记录来对当前用户的喜好(评分)进行预测。文献[5]提出了一种情境信息聚类与基于项目的推荐算法结合的方法。文献[6]将项目按照情境进行划分,结合协同过滤算法进行推荐。文献[7-8]分别利用 SVM 和粗糙集的方法对情境建模并与协同过滤方法结合进行推荐。这些研究在某些特定的领域起到了比较好的效果,但是需要将用户与大量历史数据集记录进行比较,严重影响了算法的时效性。

文中在前人工作的基础上提出一种基于用户情境聚类的个性化推荐算法,把大量数据的比较操作与协同过滤操作分离开来。在获得相似度较高的用户集合的同时,减少了用于协同过滤的用户数目,即缩短了算法的执行时间。该算法利用模糊聚类的方法首先获得相似度较高的用户集合,并与传统基于用户的协同过滤算法相结合进行推荐。实验使用公开数据集,结果表明该算法可用,并且相对于传统协同过滤算法能有效提高推荐的质量。

1　基于用户的协同过滤算法

目前推荐系统常用到的推荐算法有三种[9]:基于用户的协同过滤算法、基于项目的协同过滤算法以及混合推荐算法。基于用户的协同过滤算法主要包括两个步骤:

(1)找到和当前用户兴趣度相似的用户集合;

(2)找到这个集合中的用户感兴趣的,且当前用户没有使用过的项目推荐给当前用户。

用户兴趣相似度计算主要包括三种:余弦相似度、关联度(Pearson 相关系数)和余弦相似度修正方法,这三种启发式方法的思想是通过考察与用户 u 相似的用户 v 对项目 i 的评价来预测 u 对项目 i 的感兴趣程度。文献[10]提出的公式是一种修正的余弦相似度计算方法:

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in N(u)\cap N(v)}\frac{1}{\log(1+|N(i)|)}}{\sqrt{|N(u)||N(v)|}} \tag{1}$$

公式(1)通过分子削弱了热门项目对相似度的影响,使得用户相似度更接近准确。公式(2)利用求 Pearson 相关系数的方法:

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in N(u)\cap N(v)}(r_{(u,i)}-\bar{r}_u)(r_{(v,i)}-\bar{r}_v)}{\sqrt{\sum_{i\in N(u)\cap N(v)}(r_{(u,i)}-\bar{r})^2\times\sum_{i\in N(u)\cap N(v)}(r_{(v,i)}-\bar{r})^2}} \tag{2}$$

其中,$r_{(u,i)}$ 与 $r_{(v,i)}$ 代表用户 u、v 对共同选择的项目 i 的评分;$\bar{r}_u$,$\bar{r}_v$ 代表用户 u 和用户 v 对其自身选择项目的平均评分。

由公式(2)不难看出,用户相似度计算是时间复杂度比较大的操作,尤其在用户数较多的情况下。因此利用用户情境信息来过滤掉不相关的偏好信息(即用户操作的历史记录),对于缩短算法的处理时间,提高推荐准确度是非常有必要的。Adomavicius 等人提出了情境感知推荐系统的三种范式:情境预过滤、情境后过滤与情境建模。其中情境预过滤是指在生成推荐结果之前,首先利用情境信息过滤掉不相关的用户数据,构建和当前情境相似的数据集合。然后利用传统推荐技术处理筛选后的数据集合,生成推荐。该思想已被实践证明在特定领域能起到比较好的效果。

2　用户情境信息的聚类算法

一个由移动终端采集的情境信息通常由多个情境变量组成(如位置、天气、时间、温度、设备信息等)。因此一条情境信息包含的数据类型有多种,如区间型(如温度)、二元型(如性别)、比例型、标称型(如职业)、序列型、混合型等。所以对用户情境信息的聚类可以看作是一种对混合型变量组合的相似度的计算。由于每个情境变量的量纲和数量级均不同,因此必须要先进行数据规格化[11],使得每一变量在统一的数值范围之内,如[0,1]。可以运用统计学中提供的标准差规格、极大值规格和平均绝对偏差规格等方法进行上述处理。对于处理后的数据采用基于距离的相似度计算公式来计算两个情境信息之间的距离。

$$d(i,j)=\frac{\sum_{f=1}^{p}\delta_{ij}^{(f)}d_{ij}^{(f)}}{\sum_{f=1}^{p}\delta_{ij}^{(f)}} \tag{3}$$

其中,$\delta_{ij}^{(f)}$ 是变量 f 的权重;$d_{ij}^{(f)}$ 是情境 i 和 j 在变量 f 上的距离。

对于不同数据类型变量之间的距离计算,借鉴文献[12]中的方法,这里就不再赘述了。在多属性情境信息条件下,由于事先并不能确定各个属性是否全部有用,各个属性所占权重如何,即区别不同情境信息之间的界限是模糊的,因而可以采取模糊聚类的方法,即利用规格化后的情境信息,计算出两两情境信息之间的距离,在此基础上建立模糊相似矩阵 $\boldsymbol{R}$,利用传递闭包的方法构造模糊等价矩阵 $\boldsymbol{t}(\boldsymbol{R})=\boldsymbol{R}^{2^k}(k\geqslant 1)$。最后再取不同的相似度 $\lambda\in[0,1]$ 值来获得动态聚类的结果。

3　基于用户情境信息聚类的个性化推荐算法

Chen 在文献[13]中提出的“相似用户具有相似

的偏好”并不充分,还应当关注“其他用户在与活动用户当前情境相似的情境条件下对项目的偏好”。文献[14]提出一种基于用户情境相似度的协同过滤算法,该算法利用情境预过滤范式的思想,利用求相关系数的方法构造出相似情境集合,从而把多维推荐问题转化为二维推荐问题,进而使用基于用户的协同过滤算法进行推荐操作。该算法相对于传统的用户推荐算法保证了较好的推荐准确率。

但是算法若应用于多维情境信息的处理,时效性将更为低下。将该推荐算法应用到移动推荐系统之中将很难满足实时性的需求。文中针对该算法在情境预过滤步骤上需要与大量数据集进行比较的问题进行改进,提出一种基于用户情境聚类的个性化推荐算法。

3.1 算法设计思想

算法基于用户(User)-情境(Context)-项目(Item)三维模型(简写成 $U \times \boldsymbol{C} \times I$),其中 $\boldsymbol{C}$ 表征 n 维情境向量,$\boldsymbol{C} = \{c_1, \cdots, c_i, \cdots, c_n\}$,$c_i$ 代表情境信息的某一个属性分量(如时间、地理位置、天气、职业信息等)。抽取历史数据集中情境集 $\boldsymbol{C}$ 部分进行聚类分析,产生分类结果。之后将用户当前情境 c 根据历史情境的分类结果进行判别分析,得到分类结果 C_t,将多维的 c 用一维的 C_t 代替。抽取数据集中情境类别为 C_t 的 $U \times I \times R$ 数据,即进行情境预过滤,之后再通过基于用户的协同过滤算法以获得推荐项目集。

3.2 算法过程描述

基于上述设计思想,文中提出的基于用户情境聚类的协同过滤算法由两个过程组成。下面将算法的输入、输出与步骤分别进行阐述。

3.2.1 用户情境模糊聚类算法

算法输入:用户历史操作记录集 $H(U \times \boldsymbol{C} \times I$ 三维模型结构),其中包含 n 条记录,每条记录中 $\boldsymbol{C}$ 部分包含 m 个情境属性。

算法输出:聚类结果 $C_t = \{C_{t1}, \cdots, C_{ti}, \cdots, C_{tm}\}$。

算法步骤如下:

(1)抽取数据集 H 中的 $\boldsymbol{C}$ 部分,利用 n 个样本的 m 个情境属性建立 $n \times m$ 维矩阵;

(2)用标准差规格化的方法,即使用公式(4)将各个情境属性值规格化到[0,1]区间之内,更新矩阵,用 c'_{ij} 代替原矩阵中的元素 c_{ij};

$$c'_{ij} = \frac{c_{ij} - \overline{c_j}}{\sigma_j} \tag{4}$$

其中,$\overline{c_j} = \frac{1}{n}\sum_{i=1}^{n} c_{ij}$;$\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(c_{ij} - \overline{c_j})^2}$。

(3)计算两两情境信息之间每个属性变量的欧几里得距离,利用公式(3),合并生成一个距离值,由于暂时没有很好的方法确定每种情境属性所占有的权重,因此假定各个属性所占权重相等,即公式(3)中权重因子 δ 统一使用值 1;

(4)由情境变量之间的距离值建立 $n \times n$ 维矩阵。该矩阵具有对称性,为便于处理,将其简化成上(下)三角矩阵 $\boldsymbol{R}$。用传递闭包的方法,即 $t(\boldsymbol{R}) = \boldsymbol{R}^{2^k}(k \geqslant 1)$ 构造模糊等价矩阵;

(5)确定相似度 λ 的值,将相似度大于等于 λ 的情境信息归为一类,生成聚类结果,并为每一个情境信息添加聚类结果 C_t。

3.2.2 基于用户情境聚类的协同过滤算法

算法输入:用户历史操作记录集 $H(U \times \boldsymbol{C} \times I$ 三维模型结构),用户 u 当前的情境 c 以及项目资源集合。

算法输出:用户 u 在当前情境 c 下的项目资源 Top-N 推荐列表 i。

算法步骤如下:

(1)用 3.2.1 节介绍的方法对用户情境进行模糊聚类操作获得聚类结果,若此步已执行则跳转至步骤(2);

(2)将当前情境 c 同样利用公式(4)进行标准规格化处理后,使用步骤(1)中已分好的类进行判别分析,通过计算 c 到各个聚类中心点的欧几里得距离,取距离值最小的,判断其所属情境类型 C_t;

(3)获得记录集 H 中 C_t 对应的 $U \times I$ 数据集,即将 $U \times \boldsymbol{C} \times I$ 三维模型降为 $U \times I$ 的二维模型;

(4)对第(3)步产生的 $U \times I$ 模型数据集,利用基于用户的协同过滤算法,使用公式(5)计算用户对项目资源的评分;

$$p_{u,i} = \overline{r_u} + \frac{\sum_{v \in U} \mathrm{sim}(u,v)(r_{u,i} - \overline{r_v})}{\sum_{v \in U} \mathrm{sim}(u,v)} \tag{5}$$

其中,$\mathrm{sim}(u,v)$ 的计算方法见公式(2)。

最后将预测评分较高的 Top-N 项目集合 i 推荐给用户 u。

4 实验设计与结果分析

4.1 数据集

传统推荐系统的数据集没有包含情境因素,并且由于情境因素的多样性以及获得情境数据涉及到法律、隐私权等可行性问题,目前情境感知推荐系统的研究一直缺乏公开、通用、有效的数据集。

文中采用 MovieLens-100K 数据集,该数据集来源于美国明尼苏达大学计算机科学与工程系的 GroupLens Research 项目组,包括了 1 000 个用户对 1 700 部电影的 100 000 条记录。虽然该数据集中缺乏外部情境信息,但是包含了用户性别、年龄、职业、邮编地址等内部情境信息,并且数据集已经被清理与量化,可以

被推荐算法直接使用。此外该数据集中每名用户至少提供了 20 次的评分记录,一定程度上缓解了数据稀疏性问题。

综上所述,该数据集可较为理想地作为本算法测试数据集。

4.2　评价标准

个性化推荐算法的评测标准有很多,文中依据实验设计与数据集的选择,选用统计学中的 MAE(Mean Absolute Error)指标来对推荐结果的质量进行评测。推荐算法对用户 u 推荐 N 个项目,即 Top-N 集合,记为 $N(u)$。则 MAE 的计算公式为:

$$\mathrm{MAE} = \left(\sum_{i=1}^{N} | p_i - q_i | \right) / N \qquad (6)$$

其中,p_i 和 q_i 分别代表对 $N(u)$ 每一项的实际评分和推荐算法的预计评分。

MAE 的值越小表示推荐的误差越小,准确度越高。文中实验选用 N 为 10 的项目推荐集合。

4.3　实验结果分析

实验用 Matlab 软件进行了模拟与测试,将文中提出的算法与传统基于用户的协同过滤算法在 MAE 指标上进行了比较,并取用户情境模糊聚类中不同的相似度 λ 值获得了 5 组数据。

结果如图 1 所示。

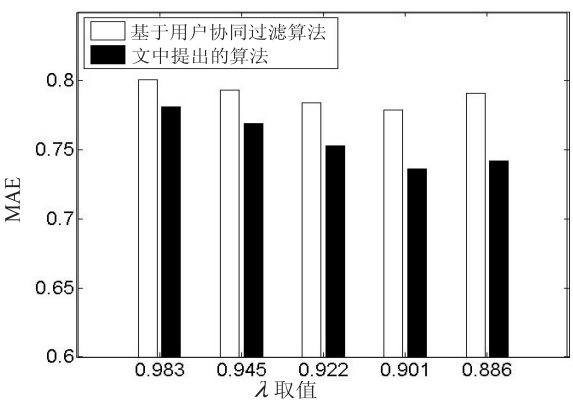

图 1　文中算法与协同过滤算法的 MAE 比较

图中横坐标是用户情景聚类时截取的 5 种不同相似度,纵坐标是在当前聚类状态下推荐的 Top-10 项目集中项目预测评分与真实评分的 MAE 值。

从图 1 中可以看出,文中提出的算法在 MAE 这项指标上要优于传统协同过滤算法。说明在相似情境条件下的用户兴趣爱好也较为相似,考虑情境因素不仅可以减少推荐算法的运算量,而且可以提高推荐的准确度,由图中可知,实验在相似度 λ 值为 0.901 的情况下取得了最好的推荐效果。同时还观察到这几组数据中相似度 λ 值取最大时,推荐的效果并不是最佳。

虽然聚类的结果从理论上来说更加精确,但是相似度很大时,聚类粒度变小,此时类中成员数量少,造成协同过滤推荐时的数据稀疏性问题,对推荐质量造成了负面影响。

5　结束语

情境感知推荐系统的研究是一个新兴领域,目前的研究与应用还很不成熟。文中在借鉴前人工作的基础上提出一种基于用户情境信息聚类的协同过滤推荐算法,该算法利用情境预过滤的思想,通过对用户情境聚类的方法将 $U\times \boldsymbol{C} \times I$ 模型降为 $U\times I$ 模型。

通过实验得出,相对传统的推荐算法,利用情境信息可以提高预测的精准度;在聚类相似度值选取不同的情况下对推荐的准确度有所影响。

因此在今后的工作中,如何确定聚类相似度的值是一个需要考虑的问题。此外情境感知推荐系统应用在不同领域时,各个情境参数所占的权重也有所不同。在计算聚类相似度的过程中为各情境参数分配相应的权重,才能获得对于推荐而言更好的聚类结果。

参考文献:

[1] 杨文漪,叶　丹,肖　波,等. 情景感知系统的数据挖掘模型研究[C]//Proceedings of the 31st Chinese control conference. Hefei:[s. n.],2012:3789-3793.

[2] 李　春,朱珍民,叶　剑,等. 个性化服务研究综述[J]. 计算机应用研究,2009,26(11):4001-4005.

[3] Adomavicius G,Tuzhilin A. Context-aware recommender systems[M]//Recommender systems handbook. [s. l.]:[s. n.],2011.

[4] Shin D,Lee J W,Yeon J,et al. Context-aware recommendation by aggregating user context[C]//Proc of the CEC 2009. Washington:IEEE Computer Society,2009:423-430.

[5] 周　涛,李　华. 基于用户情景的协同过滤推荐[J]. 计算机应用,2010,30(4):1076-1078.

[6] Baltrunas L,Ricci F. Context-based splitting of item ratings in collaborative filtering[C]//Proc of the RecSys 2009. New York:ACM Press,2009:245-248.

[7] Oku K, Nakajima S, Miyazaki J, et al. Context-aware SVM for context-dependent information recommendation [C]//Proc of the MDM. [s. l.]:[s. n.],2006:109-112.

[8] Huang Zhengxing,Lu Xudong,Duan Huilong. Context-aware recommendation using rough set model and collaborative filtering[J]. Artificial Intelligence Review,2011,35(1):85-99.

[9] 王立才,孟祥武,张玉洁. 上下文感知推荐系统[J]. 软件学报,2012,23(1):1-20.

[10] 项　亮. 推荐系统实践[M]. 北京:人民邮电出版社,

(下转第 114 页)

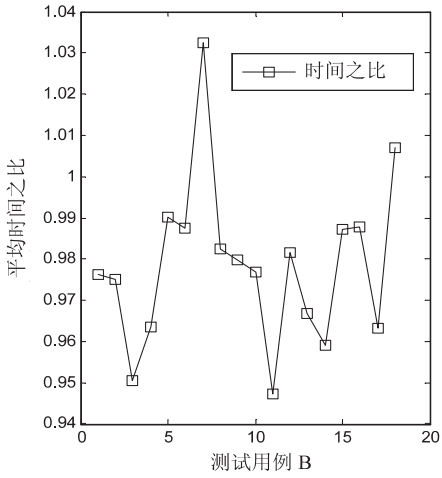

图 3 用例 B 时间比较

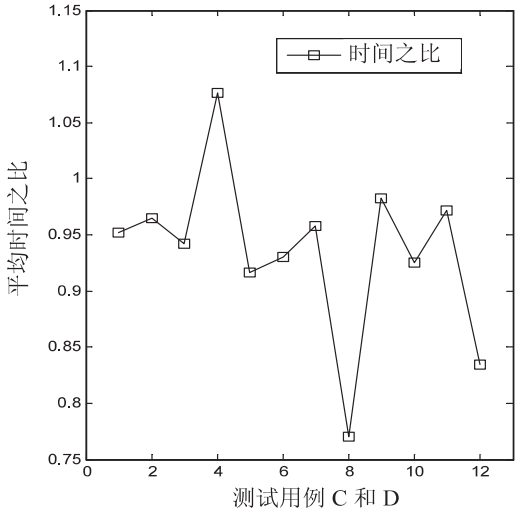

图 4 用例 C 和 D 时间比较

5 结束语

文中通过对基本遗传算法进行改进,在群体初始化时用相关的数学性质排除一些点,减少了候选点的规模,降低了染色体的长度。在个体进行交叉、变异时采用自适应方法,使进化过程可以得到有效的控制,对最优个体进行模拟退火操作,提出了求解 Steiner 最小树的混合遗传算法(GSA)。通过实验数据的对比表明,在大多数情况下 GSA 算法比 GA 算法在搜索速度上更快,得到的最优解更接近于精确解,并且解的稳定性更高。

参考文献:

[1] Hakimi S L. Steiner's problem in graphs and its implications [J]. Networks,1971,1(2):113-133.

[2] Karp R M. Reducibility among combinatorial problems, complexity of computer computations [M]. New York: Plenum Press,1972.

[3] Hwang F K, Richards D S. Steiner tree problems [J]. Networks,1992,22(1):55-89.

[4] Winter P. Steiner problem in networks: a survey [J]. Networks,1987,17(2):129-167.

[5] 梁 旭,黄 明. 现代智能优化混合算法及其应用[M]. 北京:电子工业出版社,2011.

[6] Esbensen H. Computing near-optimal solutions to the Steiner problem in a graph using a genetic algorithm [J]. Networks, 1995,26(4):173-185.

[7] 杨文国,郭田德. 求解最小 Steiner 树的蚁群优化算法及其收敛性[J]. 应用数学学报,2006,29(2):352-361.

[8] Ribeiro C C, Souza M C. Tabu search for the Steiner problem in graphs [J]. Networks,2000,36:138-146.

[9] 熊小华,刘艳芳,宁爱兵. 图的 Steiner 最小树的竞争决策算法[J]. 上海理工大学学报,2012,34(5):461-465.

[10] Lawler E L. Combinatorial optimization: networks and matroids [M]. New York: Holt, Rinehart and Winston,1976.

[11] 周 明,孙树栋. 遗传算法原理及应用[M]. 北京:国防工业出版社,1999.

[12] Kou L, Markowsky G, Berman L. A fast algorithm for Steiner trees [J]. Acta Informatica,1981,15(2):141-145.

[13] 栾庆林,卢辉斌. 自适应遗传算法优化神经网络的入侵检测研究[J]. 计算机工程与设计,2008,29(12):3022-3025.

[14] 雷英杰. MATLAB 遗传算法工具箱及应用[M]. 西安:西安电子科技大学出版社,2005.

[15] 王海英. 图论算法及其 MATLAB 实现[M]. 北京:北京航空航天大学出版社,2010.

[16] Beasley J E. OR-library: distributing test problems by electronic mail [EB/OL]. 1989. http://people.brunel.ac.uk/~mastjjb/jeb/orlib/steininfo.html.

(上接第 109 页)

2012:44-50.

[11] 李向阳,李玲娟,陈建新,等. 面向情境感知的不确定性数据融合策略[J]. 计算机技术与发展,2012,22(2):127-130.

[12] 汤效琴,戴汝源,徐 琪. 数据挖掘中变量聚类方法的应用研究[J]. 计算机工程与应用,2004,40(24):171-173.

[13] Chen A. Context-aware collaborative filtering system: predicting the user's preferences in ubiquitous computing environment [C]//Proc of the LoCA 2005. [s. l.]: [s. n.],2005: 244-253.

[14] 徐风苓,孟祥武,王立才. 基于移动用户上下文相似度的协同过滤推荐算法[J]. 电子与信息学报,2011,33(11):2785-2789.

一种基于用户情境聚类的个性化推荐算法

作者： 吴楠， 秦锋， 姜太平， WU Nan， QIN Feng， JIANG Tai-ping
作者单位： 安徽工业大学 计算机科学与技术学院,安徽 马鞍山,243032
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(10)