基于数据规范化的 co-location 模式挖掘算法

曾 新 李晓伟 杨 健
(大理大学数学与计算机学院 云南 大理 671003)

摘 要 在实际应用中,空间特征不仅包含空间信息,其特征实例还伴随着属性信息,这些属性信息对知识发现和科学决策具有重大作用。在现有的 co-location 模式挖掘算法中,计算两个不同特征实例的邻近距离时并未考虑实例不同属性的取值在邻近距离中所占的权重,导致部分属性权重过大,从而影响 co-location 模式挖掘的结果。对属性取值进行规范化,赋予所有属性相等的权重,并提出基于 join-based 的数据规范化算法 DNRA;同时,对距离阈值范围难以确定的问题进行了深入研究,推导出 DNRA 算法中距离阈值的取值范围,为用户选择适当的距离阈值提供帮助。最后,通过大量实验对 DNRA 算法的性能进行了分析比较。

关键词 co-location 模式,数据规范化,距离阈值,属性权重
中图法分类号 TP311.13 **文献标识码** A

Co-location Pattern Mining Algorithm Based on Data Normalization

ZENG Xin LI Xiao-wei YANG Jian
(College of Mathematics and Computer, Dali University, Dali, Yunnan 671003, China)

Abstract In the practical application, the spatial features not only contain the spatial information, but also the attribute information, which is important for the knowledge discovery and scientific decision. Existing co-location pattern mining algorithms do not consider the weight of instances of different attributes in the adjacent distance when calculating the adjacent distance of two different feature instances. It results in that the weight of partial attribute is too large and also affects the result of the co-location pattern mining. Standardizing the attribute values and giving an equal weight to all attributes, a data standardization algorithm DNRA based on join-based was put forward. Meanwhile, a deep research was given on the problem that the distance threshold was difficult to determine. The range of the distance threshold was derived in DNRA algorithm, helping the users to select the appropriate distance threshold. Finally, the performance of the DNRA algorithm was analyzed and compared by a large number of experiments.

Keywords Co-location pattern, Data standardization, Distance threshold, Attribute weight

1 引言

空间数据挖掘是指从空间数据库中获取潜在的、人们感兴趣的知识以及空间关系或者其他没有显示在存储空间数据库中的模式,用于理解空间数据的自身特性,并发现空间数据之间存在的隐含关系。

空间 co-location 模式是空间属性的一个子集,它们的实例在空间中频繁关联。例如:西尼罗河病毒通常发生在蚊子泛滥、饲养家禽的区域;植物学家们发现"半湿润常绿阔叶林"生长的地方 80%的概率会有"兰类"植物生长[1]。

欧几里得距离(俗称"乌鸦飞行"距离)是最流行的对象的相异性距离度量方法[2],在 co-location 模式挖掘中的应用较广泛。然而,大多数研究直接根据实例的属性初始值来计算不同实例间的欧几里得距离,并未考虑实例的不同属性的初始值对欧几里得距离的影响和距离阈值难以设置的问题。

例 1 $A.1$ 和 $B.1$ 是两个被数值属性 *sale* 和 *price* 描述的对象实例,$A.1$ 的属性值为(1000,24),$B.1$ 的属性值为(8000,60),欧几里得距离的计算式如下。

假设对象 $i=(x_{i1},x_{i2},\cdots,x_{ip})$ 和对象 $j=(x_{j1},x_{j2},\cdots,x_{jp})$ 被 p 个数值属性描述。对象 i 和 j 之间的欧几里得距离定义为:

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{ip}-x_{jp})^2}$$

根据欧几里得距离计算式计算对象实例 $A.1$ 和 $B.1$ 的欧几里得距离,如式(1)所示:

$$d(A.1,B.1)=\sqrt{(1000-8000)^2+(24-60)^2}=7000.093 \quad (1)$$

其中,*price* 属性值在 $d(A.1,B.1)$ 中所占的权重如下:

$$wr(price)=\frac{(24-60)^2}{(1000-8000)^2+(24-60)^2}=0.000026$$

如果直接忽略实例 $A.1$ 和 $B.1$ 的 *price* 属性,则计算结果如下:

$$d(A.1,B.1)=\sqrt{(1000-8000)^2}=7000 \quad (2)$$

由式(1)、式(2)和 $wr(price)$ 的计算结果可知:*sale* 属性在邻近度距离的计算中对结果的影响远大于 *price* 属性,即

本文受国家自然科学基金项目(71462001),云南省科技厅应用基础青年项目(2016FD071),云南省教育厅项目(2016ZZX192)资助。
曾 新(1986-),男,硕士,讲师,主要研究方向为空间数据挖掘,E-mail:hbzengxin@163.com;李晓伟(1985-),男,博士,讲师,主要研究方向为信息安全、计算机网络;杨 健(1976-),男,博士,副教授,CCF 会员,主要研究方向为云计算、数据安全与隐私保护。

sale 属性所占的权重要高于 *price* 属性所占的权重。这导致对象实例的某一重要属性(例如 *price* 属性)对实例间的邻近度计算毫无影响，并未实现属性权重一致的目标；同时，由于不同属性的取值范围相差较大，邻近距离阈值难以选择。因此，本文采用基于数据规范化的方法，研究属性权重一致和距离阈值范围可确定的空间 co-location 模式挖掘问题具有一定的应用价值。

本文第 2 节介绍 co-location 模式挖掘研究的相关工作；第 3 节介绍 co-location 模式的相关概念及性质；第 4 节给出基于数据规范化的 co-location 模式挖掘算法；第 5 节进行实验分析；最后总结全文。

2　co-location 模式挖掘研究的相关工作

自 Agrawal 等设计出 Apriori 算法以来，文献[3]提出了基于全连接的方式产生候选模式的 join-based 算法，即 k 阶模式产生 $k+1$ 阶候选模式，基于 k 阶的表实例连接产生 $k+1$ 阶表实例，该方法在数据集较大时，连接操作会产生巨大的开销。文献[4]提出 partial-join 算法，对实例进行分块处理，对块内、块间实例进行连接，从而减少了连接中的计算量。文献[5]提出一种基于星型邻近扩展的 join-less 算法以解决候选模式生成中的连接开销问题。文献[6]提出基于前缀树的 CPI-tree 算法，以树型结构表示空间对象实例间的邻近关系，co-location 表实例通过 CPI-tree 快速生成，算法的性能高于 join-less 算法。在 CPI-tree 算法的基础上，文献[7]和文献[8]分别提出了 iCPI-tree 和 Order-Clique-Based 算法，对前缀树的结构进行了进一步优化，取得了较高的效率。

近年来，空间数据呈爆炸式增长，从空间数据集中挖掘有价值的信息成为研究热点，且取得了不少的研究成果。文献[9]提出模糊对象的空间 co-location 模式挖掘基本算法(FB 算法)，并在 FB 算法的基础上扩展出了 3 种优化剪枝算法。文献[10]深入分析传统挖掘方式过度消耗时间和空间资源的根本原因，针对海量数据的挖掘问题，提出网格微分挖掘 co-location 模式的算法。空间特征不仅包含空间信息，还伴随着属性信息，属性信息对决策和知识发现具有重大意义，文献[11]针对带有模糊属性的空间数据集进行 co-location 模式挖掘的研究，并提出基于桶的候选模式产生算法。文献[12]针对空间 co-location 模式挖掘都只关注某一个时刻的空间 co-location 模式，而实际上数据库中的数据是随着时间改变的情况，提出 co-location 模式增量挖掘算法。空间数据挖掘的相关研究大多数是基于理想化数据和实例平等思想，而忽略了实际场景中存在的时间约束，文献[13]提出带时间约束的 co-location 模式挖掘方法，挖掘出更具实际意义的并置关系。除此之外，针对 co-location 模式的高效用问题，文献[14]给出特征效用，并提出挖掘高效用 co-location 模式的基础算法和剪枝算法。文献[15]给出特征实例效用，并利用领域驱动方法，提出挖掘高效用 co-location 模式的算法及剪枝算法。

3　co-location 模式的相关概念及性质

3.1　相关概念

定义 1(空间对象与空间对象的实例)　空间对象是指空间不同类别的事物，空间对象集 $F=\{f_1,f_2,\cdots,f_n\}$，其中 f_i 表示空间中的某个对象。在空间某个确定位置上的对象称为空间对象实例。

对象 A,B,C 分别有 2,3 和 2 个实例，它们的分布位置如图 1 所示，其中 $A.2$ 表示对象 A 的第 2 个实例。

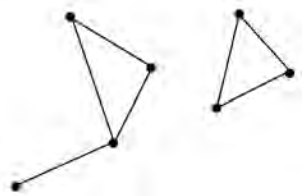

图 1　空间对象及其实例

定义 2(空间邻近关系 R)　R 用于表示空间对象实例之间的空间关系，包括距离关系、混合关系、空间拓扑关系等。利用欧几里得距离来表示两实例间的邻近关系 R，计算式如下：

$$R(A.1,B.2)\Leftrightarrow distance(A.1,B.2)\leqslant d$$

其中，d 为预设的欧几里得距离阈值，图 1 中 $A.1$ 和 $B.2$ 是 R 邻近的，且用实线连接。

定义 3(团、co-location 模式、行实例、表实例)　假设实例集 $I=\{i_1,i_2,\cdots,i_n\}$，如果 I 中的任何两个实例间都满足：

$$\{R(i_x,i_y)\mid 1\leqslant x\leqslant n,1\leqslant y\leqslant n\}$$

则称 I 为团，例如图 1 中$\{A.2,B.3,C.2\}$就是一个团。

空间 co-location 模式表示一组空间对象的集合，用 c 表示，其中 $c\subseteq F$，而 c 中对象的个数称为模式 c 的阶。例如：$c=\{A,B,C\}$是一个空间 co-location 模式，模式 c 的阶为 3。

如果存在一个团包含模式 c 的所有对象，并且此团的任何子集都不包含模式 c 的全部对象，则称此团为模式 c 的一个行实例。例如图 1 中$\{A.2,B.3,C.2\}$就是模式 c 的一个行实例。

模式 c 的表实例为其所有行实例的集合，记为 *table-instance*(c)。例如，图 1 中模式 c 的表实例为$\{\{A.1,B.2,C.1\},\{A.2,B.3,C.2\}\}$。

定义 4(参与度与参与率)　在空间 co-location 模式挖掘中，参与度和参与率是衡量 co-location 模式频繁程度的重要标准。

假设 f_i 为空间中的某个对象，f_i 在模式 $c=\{f_1,f_2,\cdots,f_k\}$中的参与率为：

$$PR(c,f_i)=\frac{\pi f_i(table-instance(c))}{|table-instance(f_i)|}$$

其中，π 是关系的投影操作，即 f_i 在模式 c 中的参与率就是 f_i 的实例在模式 c 的所有实例中不重复出现的个数与 f_i 总实例的个数的比值。

例 2　图 1 中模式 c 的表实例为$\{\{A.1,B.2,C.1\},\{A.2,B.3,C.2\}\}$，而对象 A,B,C 的实例个数分别为 2,3,2，根据参与率计算式：$PR(c,A)=\frac{2}{2}=1$，$PR(c,B)=\frac{2}{3}=0.67$，$PR(c,C)=\frac{2}{2}=1$。

模式 c 的参与度是其所有对象参与率的最小值，记为：$PI(c)=\min\limits_{1\leqslant i\leqslant k}\{PR(c,f_i)\}$，则例 2 中模式 c 的参与度：$PI(c)=\min(1,0.67,1)=0.67$。

定义 5(频繁模式)　如果模式 c 的参与度 $PI(c)$大于或等于用户预先设定的最小参与度阈值 *min_prev*，则 c 是频繁模

式，否则 c 是非频繁模式。假设用户预先设定的 $min\_prev=0.5$，在例 2 中 $PI(c)=0.67$，由于 $PI(c)\geqslant min\_prev$，因此模式 $c=\{A,B,C\}$ 为频繁模式。

3.2 相关性质

定理 1 参与度和参与率随着 co-location 模式 c 的阶的增大而单调递减。

证明：假设模式 c 的行实例中包含某一空间对象 f_i 的实例，如果模式 c' 是 c 的子集，那么 f_i 的实例也一定被包含在 c' 的行实例中，反之则不然。因此空间对象的参与率随着模式阶的增长而递减。

假设 $c=\{f_1,f_2,\cdots,f_k\}$，$PI(c\cup f_{i+1})=\min\limits_{i=1}^{k+1}\{PR(c\cup f_{k+1},f_i)\}\leqslant\min\limits_{i=1}^{k}\{PR(c\cup f_{k+1},f_i)\}\leqslant\min\limits_{i=1}^{k}\{PR(c,f_i)\}=PI(c)$，因此模式 c 的参与度也是单调递减的。

4 基于数据规范化的 co-location 模式挖掘算法

4.1 相关定义及距离阈值范围

引入数据规范化，赋予对象实例的所有属性相等的权重，防止具有较大初始值域的属性与具有较小初始值域的属性相比权重过大，造成某一属性在邻近距离计算中可以忽略不计，这样会导致部分属性的信息丢失，影响人们对感兴趣的 co-location 模式的挖掘结果；同时，对于邻近距离阈值难以选择的问题，也进行了一定的研究。

定义 6[16]（最小-最大规范化） 假设 $\min_p$ 和 $\max_p$ 分别为属性 p 的最小值和最大值。最小-最大规范化通过公式：$v_p'=\frac{v_p-\min_p}{\max_p-\min_p}(new\_\max_p-new\_\min_p)+new\_\min_p$ 将属性 p 的值 v_p 映射到一个给定的新区间 $\lfloor new\_\min_p,new\_\max_p\rfloor$ 中的值 v_p'。

例 3 将例 1 中的实例 $A.1$ 和 $B.1$ 的 *sale* 和 *price* 属性值映射到区间 $[0,1]$ 中，假设 *sale* 属性值的初始取值范围为 $[1000,10000]$，而 *price* 属性值的初始取值范围为 $[10,100]$，则 $A.1$ 的数据规范化过程如下：

$$v'_{sale}=\frac{1000-1000}{10000-1000}(1-0)+0=0$$

$$v'_{price}=\frac{24-10}{100-10}(1-0)+0=0.16$$

$A.1$ 的属性值被规范化为 $(0,0.16)$。而 $B.1$ 的数据规范化过程如下：

$$v'_{sale}=\frac{8000-1000}{10000-1000}(1-0)+0=0.78$$

$$v'_{price}=\frac{60-10}{100-10}(1-0)+0=0.56$$

$B.1$ 的属性值被规范化为 $(0.78,0.56)$。此时，根据欧几里得距离的计算式重新计算 $A.1$ 和 $B.1$ 之间的距离：

$$d(A.1,B.1)=\sqrt{(0.78-0)^2+(0.56-0.16)^2}=0.88 \tag{3}$$

在数据规范化之后，*price* 属性值在 $d(A.1,B.1)$ 中所占的权重为：

$$wr(price)=\frac{(0.56-0.16)^2}{(0.78-0)^2+(0.56-0.16)^2}=0.21$$

如果忽略 *price* 属性，$A.1$ 和 $B.1$ 之间的距离为：

$$d(A.1,B.1)=\sqrt{(0.78-0)^2}=0.78 \tag{4}$$

由式(3)、式(4)和 $wr(price)$ 的计算结果分析可知：*price* 属性值对 $A.1$ 和 $B.1$ 的距离影响较大，不可忽略。

在不具备足够先验知识的情况下，用户面对众多属性的不同取值范围，很难给出恰当的距离阈值 d，可能会丢失一些有意义的邻近关系，并导致错失一些感兴趣的模式，因此确定距离阈值的范围显得尤为重要。

引理 1 空间对象的所有实例的属性值经过数据规范化映射到区间 $[m,n]$（$0\leqslant m\leqslant n$），那么实例间的距离阈值范围为 $[0,\sqrt{2}(n-m)]$。

证明：假设实例 $A.1$ 和 $B.1$ 的属性值分别为 (x_1,y_1) 和 (x_2,y_2)，经过数据规范化之后，属性值分别为 (x_1',y_1') 和 (x_2',y_2')，属性值的取值范围为 $m\leqslant x_1',x_2',y_1',y_2'\leqslant n$。

$$0\leqslant|x_2'-x_1'|\leqslant|n-m|\Rightarrow 0\leqslant(x_2'-x_1')^2\leqslant(n-m)^2$$

$$0\leqslant|y_2'-y_1'|\leqslant|n-m|\Rightarrow 0\leqslant(y_2'-y_1')^2\leqslant(n-m)^2$$

则可以推导出欧几里得的取值范围为：

$$0\leqslant(x_2'-x_1')^2+(y_2'-y_1')^2\leqslant 2(n-m)^2\Rightarrow$$

$$0\leqslant d(A.1,B.1)=\sqrt{(x_2'-x_1')^2+(y_2'-y_1')^2}\leqslant\sqrt{2}(n-m)$$

4.2 算法

数据规范化算法（DNRA）以 join-based 算法为基础，通过前期对实例的属性值进行数据规范化，以使实例属性权重一致以及确定邻近距离阈值的取值范围，如算法 1 所示。

算法 1 数据规范化算法（DNRA）

输入：

对象集 $F=\{f_1,f_2,\cdots,f_n\}$，共有 n 个对象。实例集 $I=\{i_{11},i_{12},\cdots,i_{1m},\cdots,i_{nm}\}$，表示有 n 个对象，每个对象至多有 m 个实例；预先设定的邻近距离阈值 d；预先设定的频繁模式阈值 min_prev；数据规范化后的属性值区间[min，max]；中间变量；模式的阶 k，k 阶 co-location模式候选集 C_k，C_k 中 co-location 模式表实例的集合 Tab_k，k 阶频繁 co-location 模式集 P_k，邻近关系集 neiR

输出：

频繁 co-location 模式集合 P

算法过程：

1. 生成频繁 1 项集 P_1；
2. 初始化频繁 co-location 模式集 P 为空；
3. 利用〈最小-最大规范化〉方式将实例集 I 中实例的所有属性值映射到区间[min，max]；
4. 根据给定的邻近距离阈值 d 和欧几里得距离公式产生实例间的邻近关系集 neiR；
5. for(k=2；$P_{k-1}\neq\emptyset$；k++)
   - 5.1 将频繁(k−1)阶模式连接产生 k 阶候选模式集 C_k；
   - 5.2 将频繁模式的反单调性作为剪枝策略，对候选模式集 C_k 进行剪枝；
   - 5.3 通过(C_k，Tab_{k-1}，neiR)产生 Tab_k；
   - 5.4 根据候选模式集 C_k 和所有候选模式的表实例集 Tab_k，计算每个 k 阶候选模式的参与度，并将参与度大于或等于 min_prev 的模式放入 P_k 中；
   - 5.5 将 P_k 合并到 P 中；
6. 返回最终的频繁 co-location 模式集 P。

5 实验分析

为了评估 DNRA 算法的性能，本文采用随机生成的数据集进行了大量实验，并对实验结果进行了分析对比，最后得出结论。实验的硬件平台为 Intel core i3 处理器、4 GB 内存、64

位 Windows 7 操作系统；软件编程环境为 Python 2.7 版本。

实验中原始数据集的描述如表 1 所列。

表 1　原始数据集参数

参数	参数值
对象个数	10
属性个数	2
属性 1 范围	[100，1000]
属性 2 范围	[1，100]
距离阈值	25
参与度阈值	0.7
实例个数	100/200/300/400

原始数据集中的实例属性值经过数据规范化后的描述如表 2 所列。

表 2　规范化数据集参数

参数	参数值
对象个数	10
属性个数	2
属性范围	[0，1]
距离阈值	0.1
参与度阈值	0.7
实例个数	100/200/300/400

从表 1 和表 2 中可以看出，原始数据集和规范化数据集都有 10 个不同的对象，每个对象有 2 个属性，实验所用实例个数依次为 100，200，300 和 400。

5.1　属性权重对频繁模式数目的影响

采用默认的原始数据集参数和规范化数据集参数，比较 join-based 算法和数据规范化算法（DNRA）分别在原始数据集和规范化数据集上的频繁模式挖掘结果，如图 2 所示。

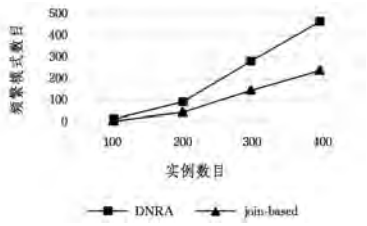

图 2　属性权重对频繁模式数目的影响

原始数据集通过数据规范化之后得到的规范化数据集在 DNRA 上挖掘的频繁模式数目多于原始数据集直接在 join-based 算法上挖掘的频繁模式数目，因此属性权重的巨大差异会导致挖掘的频繁模式数目减少。

5.2　DNRA 与 join-based 算法挖掘效率的比较

采用默认的原始数据集参数和规范化数据集参数，依次取实例数目为 100，200，300 和 400 来比较 DNRA 和 join-based 算法的挖掘效率，如图 3 所示。

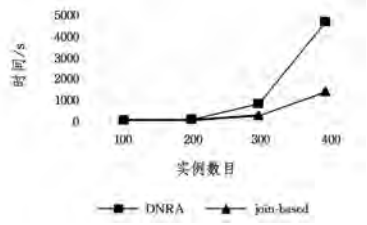

图 3　算法挖掘效率的比较

从图 3 中可以看出，随着实例数目的不断增加，DNRA 算法比 join-based 算法耗费更多的时间，这与图 2 中 DNRA 算法挖掘出更多的频繁模式一致。

5.3　距离阈值对 DNRA 算法的影响

根据引理 1 中距离阈值的取值范围和规范化数据集中属性的取值范围，计算得到实规范化数据集的距离阈值范围为 $[0,\sqrt{2}]$，我们取距离阈值为 0.05，0.1，0.15 和 0.2 分别进行实验，采用实例数目为 200 的规范化数据集，参与度阈值为 0.7，实验结果如图 4、图 5 所示：

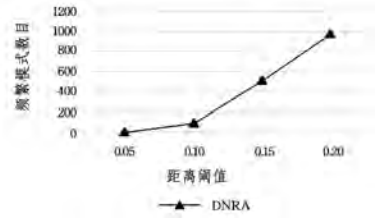

图 4　距离阈值对频繁模式数目的影响

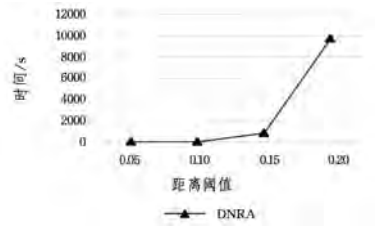

图 5　距离阈值对算法效率的影响

随着距离阈值的增大，不同对象实例间的邻近关系对增多，导致候选模式的行实例数目增多及参与度值增大，因此频繁数目也会随之增大，耗费的时间也随之增多。

5.4　参与度阈值对 DNRA 算法的影响

采用实例数目为 200 的规范化数据集，距离阈值为 0.1，我们取参与度阈值为 0.6，0.7，0.8 和 0.9 分别进行实验，实验结果如图 6 和图 7 所示。

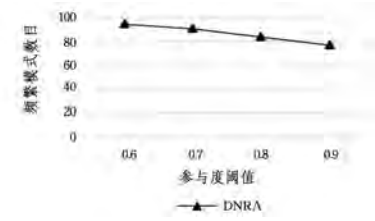

图 6　参与度阈值对频繁模式数目的影响

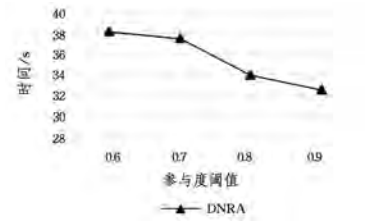

图 7　参与度阈值对算法效率的影响

在其他条件一定的情况下，随着参与度阈值的增加，频繁模式的数目递减，算法的执行时间也随之减小。

结束语　在计算实例对邻近距离的过程中，将实例的不同属性取值规范化到给定的范围内，使所有属性取值具有相等的权重，同时在规范化数据范围内计算邻近距离阈值的取值范围，为用户设置合适的距离阈值提供帮助，从而挖掘出用户感兴趣的频繁模式，为科学决策提供依据。现实生活中的数据大多含有某类约束条件，后期将考虑如何在带有条件的规范化数据集上挖掘用户感兴趣的 co-location 模式。

参 考 文 献

[1] 王丽珍,周丽华,陈红梅,等.数据仓库与数据挖掘原理及应用(第2版)[M].北京:科学出版社,2009:1-19.

[2] HAN J,KAMBER M,PEI J. Data mining:concept and techniques(Third Edition)[M]. Beijing:China Machine Press,2006:1-23.

[3] HUANG Y,SHEKHAR S,XIONG H. Discovering Co-location Patterns from Spatial Data Sets:A General Approach[C]// IEEE Transactions on knowledge and Data Enginnering (TKDE). 2004:1472-1485.

[4] YOO J S,SHEKHAR S. A partial Join Approach for Mining Co-location Patterns[C]// Proc. of the 12th Annual ACM Int. Workshop on Geographic Information Systems. Washington DC,USA,2004:241-249.

[5] YOO J S,SHEKHAR S,CELIK M. A join-less approach for co-location pattern mining:A summary of results[C]// Proc. of the 5th IEEE Int. Conf. on Data Mining. Washington:IEEE Computer Society,2005:813-816.

[6] WANG L Z,BAO Y Z,LU J,et al. A new join-less approach for co-location pattern mining[C]// IEEE International Conference on Computer and Information Technology (CIT2008). Washington,2008:197-202.

[7] WANG L Z,BAO Y Z,LU Z Y. Efficient discovery of spatial co-location patterns using the iCPI-tree[J]. The Open Information Systems Journal,2009,3(1):69-80.

[8] WANG L Z,ZHOU L H,LU J,et al. An order-clique-based approach for mining maximal co-locations[J]. Information Sciences,2009,179(19):3370-3382.

[9] 欧阳志平,王丽珍,陈红梅.模糊对象的空间 co-location 模式挖掘研究[J].计算机学报,2011,34(10):1947-1955.

[10] 姚华传,王丽珍,陈红梅,等.面向海量数据的空间 co-location 模式挖掘新算法[J].计算机科学与探索,2015,9(1):24-35.

[11] 吴萍萍,王丽珍,周永恒.带模糊属性的空间 co-location 模式挖掘研究[J].计算机科学与探索,2013,7(4):348-358.

[12] 芦俊丽,王丽珍,肖清,等.空间 co-location 模式增量挖掘及演化分析[J].软件学报,2014,12(25):190-199.

[13] 曾新,杨健.带时间约束的 co-location 模式挖掘[J].计算机科学,2016,43(2):293-296.

[14] 杨世晟,王丽珍,芦俊丽,等.空间高效用 co-location 模式挖掘技术初探[J].小型微型计算机系统,2014,35(10):2302-2307.

[15] 江万国,王丽珍,方圆,等.领域驱动的高效用 co-location 模式挖掘方法[J].计算机应用,2017,37(2):322-328.

[16] HAN J W,KAMBER M,PEI J.数据挖掘概念与技术(第3版)[M].北京:机械工业出版社,2014:74-76.

(上接第 470 页)

参 考 文 献

[1] ADOMAVICIUS G,TUZHILIN A. Toward the next generation of recommender systems:A survey of the state-of-the-art and possible extensions[C]// Proceedings of the IEEE Transactions Knowledge and Data Engineering. 2005:734-749.

[2] LÜ L,MEDO M,YEUNG C H,et al. Recommender systems [J]. Physics Reports,2012,519(1):1-49.

[3] SU X,KHOSHGOFTAAR T M. A survey of collaborative filtering techniques [J]. Advances in Artificial Intelligence,2009,2009(12):4.

[4] WEI C,HSU W,LEE M L. A unified framework for recommendationsbased on quaternary semantic analysis[C]// Proceedings of the 34th International ACM SIGIR Conference on Research and Development InInformation Retrieval. Beijing,China,2011:1023-1032.

[5] WANG L C,MENG X W,ZHANG Y J. Context-Aware recommender systems:A survey of the state-of-the-art and possible extensions[J]. Journal of Software,2012,23(1):1-20.

[6] LIN J,SUGIYAMA K,KAN M Y,et al. Addressing cold-start in apprecommendation:latent user models constructed from twitterfollowers[C]// Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin,Ireland,2013:283-292.

[7] MISTRY O,SEN S. Tag recommendation for social book marking:Probabilistic approaches [J]. Multiagent and Grid Systems,2012,8(2):143-163.

[8] 于洪,李俊华.一种解决新项目冷启动问题的推荐算法[J].软件学报,2015,26(6):1395-1408.

[9] ZHANG Z K,ZHOU T,ZHANG Y C. Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs [J]. Physica A:Statistical Mechanics and its Applications,2010,389(1):179-186.

[10] ZHANG Z K,LIU C,ZHANG Y C,et al. Solving the cold-start problem in recommender systems with social tags [J]. EPL (Europhysics Letters),2010,92(2):28002.

[11] ZHANG Z K,ZHOU T,ZHANG Y C. Tag-Aware recommender systems:A state-of-the-art survey [J]. Journal of Computer Science and Technology,2011,26(5):767-777.

[12] JOMSRI P,SANGUANSINTUKUL S,CHOOCHAIWATTANA W. A framework for tag-based research paper recommender system:An IR approach[C]// Proceedings of the 2010 IEEE 24th Int'l Conf. on Advanced Information Networking and Applications Workshops. 2010:103-108.

[13] 蔡强,韩东梅,李海生,等.基于标签和协同过滤的个性化资源推荐[J].计算机科学,2014,41(1):69-71,110.

[14] 李慧,马小平,胡云,等.融合主题与语言模型的个性化标签推荐方法研究[J].计算机科学,2015,42(8):70-74.

[15] 叶剑虹,叶双.基于混合模式的流媒体缓存调度算法[J].计算机科学,2013,40(2):61-64.

[16] KIDEOK C,HAKYUNG J,et al. How can an ISP merge with a CDN? [J]. IEEE Communications,2011,49(10):156-162.

[17] 李瑞敏,林鸿飞,闫俊.基于用户-标签-项目语义挖掘的个性化音乐推荐[J].计算机研究与发展,2014(10):2270-2276.