模糊对象的空间 Co-location 模式挖掘研究

欧阳志平 王丽珍 陈红梅
(云南大学信息学院计算机科学与工程系 昆明 650091)

摘 要 空间 co-location 模式表示的是空间对象的实例在一个相同的区域内频繁地空间并置. 过去人们已经对确定及不确定数据的 co-location 模式挖掘问题进行了一些研究,但是针对模糊对象上进行的研究还没有. 模糊对象在许多领域里都有着非常重要的应用,比如生物医学图像数据库和 GIS. 该文研究模糊对象的空间 co-location 模式挖掘问题. 首先,定义模糊对象上空间 co-location 模式挖掘的相关概念,包括模糊参与率、模糊参与度等. 其次,提出 FB 算法挖掘模糊对象的 co-location 模式. 接着,提出了 3 种改进算法,包括剪枝对象、减少实例间连接、改进剪枝步,以提高挖掘性能、加快 co-location 规则的产生. 最后通过大量的实验说明 FB 算法及其改进算法的效果和效率.

关键词 模糊对象; co-location 模式; 空间数据挖掘; 模糊参与率; 减少连接
中图法分类号 TP311 **DOI 号**: 10.3724/SP.J.1016.2011.01947

Mining Spatial Co-Location Patterns for Fuzzy Objects

OUYANG Zhi-Ping WANG Li-Zhen CHEN Hong-Mei
(*Department of Computer Science and Engineering, School of Information Science and Engineering, Yunnan University, Kunming* 650091)

Abstract A spatial co-location pattern is a group of spatial objects whose instances are frequently located in the same region. The mining co-location pattern problem for certain and uncertain data had been investigated in the past, but not for fuzzy objects. Fuzzy objects could be applied to many areas such as biomedical image databases, GIS and more. This paper investigates the spatial co-location pattern mining problem for fuzzy objects. Firstly, it defines the related concepts of spatial co-location patterns mining on fuzzy objects, including fuzzy participation ratio, fuzzy participation index, etc. Secondly, this paper proposes an FB algorithm to mine co-location patterns from fuzzy objects. Then, three kinds of the improved algorithms, the pruning objects, reducing of the operation joining between spatial instances and optimizing the pruning steps, are put forward so as to improve the mining performance and accelerate the co-location rule generation. Finally, by extensive experiments, the efficiency and effectiveness of the algorithms are verified.

Keywords fuzzy objects; co-location patterns; spatial data mining; fuzzy participation index; reducing joining

1 引 言

空间 co-location 模式代表了一组空间对象的子集,它们的实例在空间中频繁地关联. 挖掘空间 co-location 模式就是在空间数据库中发现和挖掘空间对象之间的关联关系. 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

收稿日期:2011-06-25;最终修改稿收到日期:2011-08-08. 本课题得到国家自然科学基金项目(61063008)、云南省应用基础研究基金项目(2010CD025)、云南大学科学研究基金项目(2009F29Q)资助.**欧阳志平**,男,1985 年生,硕士研究生,主要研究方向为数据挖掘. E-mail: chxoyzp@126.com. **王丽珍**(通信作者),女,1962 年生,博士,教授,主要研究领域为数据库、数据挖掘和计算机算法. E-mail: lizhwang2005@126.com. **陈红梅**,女,1976 年生,博士研究生,讲师,主要研究方向为数据挖掘.

“半湿润常绿阔叶林”生长的地方 80%会有“兰类”植物生长[1].

在现实世界中,模糊对象无处不在,比如“漂亮的女人”、“高大的树”等,同时模糊对象在许多应用中也起着十分重要的作用,目前对模糊对象的研究范围十分广泛,但对 co-location 模式的研究还没有,针对这一情况,本文研究模糊对象上的 co-location 模式挖掘问题.

本文第 2 节为相关工作;第 3 节是相关定义及性质;第 4、5 节为算法和实验;第 6 节为结论.

2 相关工作

空间 co-location 模式挖掘问题是空间数据挖掘领域的一个重要研究方向,人们对确定数据上的 co-location 模式挖掘问题进行了深入的研究,并提出了很多算法,比如 join_based 算法[2]、partial-join 算法[3]、join-less 算法[4]、CPI-tree 算法[5]、order-clique-based 算法[6] 等. 在文献[2]中,给出了 co-location模式挖掘相关的一些定义,包括邻近关系、空间 co-location 模式、行实例、表实例、参与率、参与度以及 co-location 规则和条件概率等. 近年来,对不确定数据上的 co-location 模式挖掘的研究也越来越多,文献[7]提出了不确定集上的 UJoin_based 算法. 文献[8]研究了从区间数据表示的不确定对象中挖掘 co-location 模式. 虽然目前空间 co-location 模式挖掘算法很多,但对模糊对象的 co-location 挖掘算法还未见报道. 模糊对象[9]的研究目前主要集中在模糊对象的建模上,比如基本的类型和操作模型等. 文献[10-13]在 GIS 中对模糊对象做了大量的研究. 文献[14]研究模糊对象的 KNN 查找问题,提出了 AKNN 和 RKNN 算法.

3 相关定义及性质

本部分首先对模糊对象、模糊概率阈值、空间距离、模糊参与率以及模糊参与度进行定义,其次给出本文定义的模糊参与率及参与度所满足的一个性质.

3.1 相关定义

定义 1(模糊对象). 本文中的模糊对象,表示为一个空间中离散点的集合,定义如下:$A=\{\langle a,\mu(a)\rangle|\mu(a)>0\}$,其中 A 表示模糊对象,a 表示实例,$\mu(a)$表示实例 a 属于模糊对象 A 的隶属度.

图 1 所示为一个模糊对象 A,实例 $a1$、$a2$ 属于 A 的隶属度分别为 0.6,0.01.

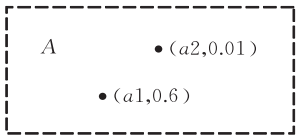


图 1 模糊对象

从图 1 中可以看出,实例 $a2$ 属于模糊对象 A 的隶属度非常低,对于这样的实例,我们希望计算 co-location 模式时把它排除,因为在现实生活中,我们认为隶属度非常低的实例对于模糊对象来说没有多大意义. 例如老年人,假设 80 岁隶属度为 0.8,50 岁为 0.02,对于 50 岁,我们认为它对于老年人这个对象贡献相当少,所以在有些应用中可以把它排除,而只关注那些符合实际应用的期望隶属度值条件的实例,于是我们有以下定义.

定义 2. 给定一个用户自定义的概率阈值 $f\_threshold$,称为模糊度阈值,集合 $A_{f\_threshold}=\{a|\mu(a)\geq f\_threshold\}$,表示满足用户自定义模糊度阈值的实例集.

如图 1 中,$A_{0.2}=\{a1\}$.

定义 3(空间邻近关系 R). 设 a,b 分别为模糊对象 A,B 的实例,它们之间的距离用欧几里德距离计算,表示为 $d(a,b)=\|a-b\|$. 由此可以定义空间邻近关系 R:若两个实例之间的欧几里德距离小于等于阈值 $dis\_threshold$,即 $d(a,b)\leq dis\_threshold$,则表示它们邻近.

当两个空间实例 a 和 b 之间满足空间邻近关系 R 时,称这两个空间实例 R 邻近,记为 $R(a,b)$,并在图中用线段连接它们(如图 2 所示).

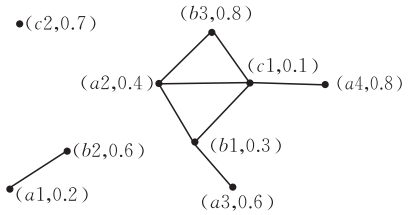


图 2 模糊对象参与度

一个空间 co-location 模式表示的是一组空间对象的集合. co-location 模式的长度称为此 co-location 模式的阶,即 co-location 模式里空间对象的个数. 例如 co-location 模式 $c=\{A,B,C\}$,则称模式 c 是 3 阶 co-location 模式.

设有空间实例集 $I=\{i_1,i_2,\cdots,i_l\}$,如果有 $\{R(i_j,i_k)|1\leq j\leq l,\ 1\leq k\leq l\}$,则称 I 是一个团

(clique). 如果团 I 包含了 co-location 模式 c 中的所有对象，并且 I 没有任何一个子集可以包含 c 中的所有对象，那么 I 是 co-location 模式 c 的一个行实例(称为 co-location 实例). co-location 模式 c 的所有行实例的集合称为表实例，记为 $table\_instance(c)$. 例如图 2 中，co-location $\{A,B,C\}$ 的表实例为 $\{\{a2,b1,c1\},\{a2,b3,c1\}\}$.

定义 4(模糊参与率). 设 f_i 为某个空间模糊对象，f_i 在 k 阶的 co_location 模式 c 中的模糊参与率表示为 $FPR(c,f_i)$，它是 f_i 的实例在空间 co-location 模式 c 的表实例中不重复出现的实例的隶属度之和与 f_i 中总实例个数的比率，公式如下：

$$FPR(c,f_i)=\frac{\sum 1\times\mu(a)}{|table\_instance(\{f_i\})|},$$

其中 $a\in\prod_{f_i}(table\_instance(c))$，而 $\prod$ 是关系投影操作.

定义 5(模糊参与度). 模糊对象的空间 co-location 模式 $c=\{f_1\cdots f_k\}$ 的模糊参与度表示为 $FPI(c)$，是模式中所有空间对象 FPR 值中的最小值，公式表示如下：

$$FPI(c)=\min_{i=1}^{k}\{FPR(c,f_i)\}.$$

设 min_prev 是用户给定的最小参与度阈值，当 $FPI(c)\geq min\_prev$ 时，称模糊对象的 co-location 模式 c 是频繁的.

例 1. 图 2 所示为 3 个空间模糊对象 $A=\{(a1,0.2),(a2,0.4),(a3,0.6),(a4,0.8)\}$，$B=\{(b1,0.3),(b2,0.6),(b3,0.8)\}$，$C=\{(c1,0.1),(c2,0.7)\}$，$R$ 关系用连线表示，假设 $min\_prev=0.2$. 则在 co-location 模式 $c=\{A,C\}$ 中，$FPR(c,A)=(1\times0.4+1\times0.8)/4=0.3$，而 $FPR(c,C)=(1\times0.1)/2=0.05$，则 $FPI(c)=\min\{0.3,0.05\}=0.05$，co-location 模式 c 为非频繁模式.

定义 4 说明：假设模糊对象 A 有 4 个实例，图 3 表示的是 A 的一个实例 $a1$ 的模糊参与率与其隶属度的关系. 横坐标表示 $a1$ 属于对象 A 的隶属度值，纵坐标表示模糊参与率值，从图中可以看出，当实例 $a1$ 的隶属度增加时，它的模糊参与率也会相应增加，因为 $a1$ 的模糊参与率值为 $a1$ 的隶属度与对象 A 的总实例个数的比率. 当隶属度为 0 时表示 $a1$ 不属于对象 A，这时模糊参与率自然为 0，隶属度为 1 时表示 $a1$ 完全属于对象 A，这时它的模糊参与率与传统定义的值一样，为 0.25.

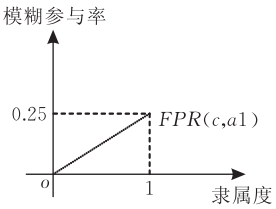


图 3　定义 4 说明

3.2　性　质

模糊参与率与参与度满足如下的一个引理和定理.

引理 1. 模糊参与率(FPR)和模糊参与度(FPI)随着 co-location 模式阶的增大单调递减.

证明. 假设某个模糊空间对象的实例包含在 co-location 模式 c 的实例中，那么当有 co-location 模式 $c1\subseteq c$，这个模糊空间对象的实例也一定包含在模式 $c1$ 的实例中，反之则不然，所以模糊参与率是单调递减的. 由于模式的模糊参与度取它包含的对象中最小参与率值，当模式的阶增大时，由于模糊参与率是递减的，所以 co-location 模式的模糊参与度也是单调递减的.

定理 1. 如果 k 阶 co-location 模式 c 是频繁的，那么它的任意 $k-1$ 阶的子 co-location 模式也是频繁的.

证明. 根据引理 1，k 阶 co-location 模式 c 的模糊参与度，要小于其 $k-1$ 阶子 co-location 模式，所以若 k 阶 co-location 模式 c 是频繁的，则 $k-1$ 阶的子 co-location 模式肯定也是频繁的.

利用定理 1 可以对候选模式进行剪枝处理，具体的剪枝过程见 4.4 节.

4　模糊对象 co-location 模式挖掘算法

首先给出一个模糊对象的 co-location 挖掘的基本算法——FB 算法，接着在其基础上提出 3 种改进算法，包括剪枝模糊对象、减少实例间的连接、优化剪枝步. 算法中模糊对象的实例按照实例的模糊度非递增排序.

4.1　FB 算法

FB 算法采用的是经典 join_based 算法的思想，循环执行以下 4 个步骤：(1) 产生候选 co-location 模式(包括连接步和剪枝步)；(2) 产生候选 co-location 模式的表实例；(3) 剪枝(利用用户自定义的参与度阈值 min_prev 进行剪枝)；(4) 产生 co-location 规则. 具体过程如算法 1 所示.

算法 1. FB 算法.

输入：空间模糊对象集 SF，空间实例集 SFI，参与度阈值 min_prev，条件概率阈值 min_conf，距离阈值 $dis\_threshold$，模糊度阈值 $f\_threshold$，满足模糊度阈值条件的模糊对象与实例的集合 F，FI

输出：co-location 规则集 FP

变量：k：co-location 模式的阶，C_k：k 阶 co-location 候选模式集，T_k：候选模式 C_k 的表实例集，P_k：k 阶 co-location 频繁集，R_k：k 阶 co-location 规则集

步骤：

1. F，FI =gen_fdata(SF，SFI，$f\_threshold$)；
2. $P_1=F$，$FP=\varnothing$；
3. for($k=2$；$P_{k-1}\neq\varnothing$；$k++$) do
   3.1. C_k=gen_candidate_co-location(k，P_{k-1})；
   3.2. T_k=gen_table_instance(C_k，T_{k-1})；
   3.3. P_k=sel_prev_co-location(C_k，T_k，min_prev)；
   3.4. R_k=sel_co-location_rule(P_k，T_k，min_conf)；
   3.5. $FP\leftarrow FP\cup R_k$；
4. return FP.

步 1 是根据文中定义的模糊度阈值得到满足条件的模糊对象和实例集；步 2 为初始化；步 3 迭代地生成频繁 co-location 模式集和规则集，其中，步 3.1 为生成 k 阶 co-location 候选模式集，步 3.2 为生成 k 阶 co-location 候选模式的表实例集，步 3.3 生成 k 阶频繁 co-location 模式集，步 3.4 生成 k 阶 co-location 规则集；步 4 返回结果.

4.2 剪枝模糊对象

由于 co-location 挖掘算法需要对模糊对象的大量实例之间进行距离计算以及连接操作，因此我们应当尽可能剪掉那些不可能存在于 co-location 模式中的模糊对象. 基于以上思考，论文提出一种有效的剪枝规则.

定理 2. 对于一个模糊对象 A，若它最大模糊参与率值小于给定的最小参与度阈值，则对象 A 不可能存在于任意的频繁 co-location 模式中.

证明. 反证法. 假设模糊对象 A 存在于某个频繁的 co-location 模式 c 中，则我们可以得到 $FPR(c,A)\geq min\_prev$. 模糊对象 A 在 c 中的最大参与率为它的所有实例均在 co-location 模式 c 的表实例中，根据模糊参与率的定义，最大模糊参与率值等于模糊对象 A 的所有实例的隶属度之和与对象 A 的实例数目的比率，由定理条件可知，它小于给定的最小支持度阈值，这时可以得到 $FPR(c,A)<min\_prev$，与假设矛盾，所以对象 A 不可能存在于任意的 co-location 模式中.

例 2. 图 2 中，设 $min\_prev=0.6$. 假设模糊对象 C 的所有实例均在 co-location 模式的行实例中，C 的最大参与率值等于 $0.1/2+0.7/2=0.4<0.6$，由定理 2 可知，C 不可能存在于任意的 co-location 模式中，这时把对象 C 剪枝掉.

利用定理 2，可以在模式挖掘前对对象进行初步剪枝，降低算法的时间复杂性. 具体见算法 2，其中步骤 1 为计算每一个模糊对象的最大参与率.

算法 2. 剪枝模糊对象算法.

输入：模糊对象集 F，实例集 FI，参与度阈值 min_prev，对象的最大参与率 T_PR

输出：剪枝模糊对象后的 F，FI

步骤：

1. for all fuzzy object $f\in F$ do
   1.1. 计算 f 的最大参与率 T_PR
   1.2. if $T\_PR<min\_prev$ then
   $F=F-\{f\}$；$FI=FI-\{a\}$(其中 $a\in f$)；
2. return F，FI.

4.3 减少实例间连接

尽管通过 3.2 节中的改进算法可以减少实例之间的计算量，但 co-location 模式挖掘过程中仍然有大量实例之间的连接操作，相当耗时. 在 FB 算法基于参与度的剪枝过程中，首先要生成候选模式的表实例，再基于参与度阈值来对候选模式进行剪枝. 我们在研究中发现，可以在表实例生成最开始阶段就对某些不可能满足参与度阈值的候选模式进行剪枝，避免表实例之间大量不必要的连接操作，大大提高算法的效率. 下面给出定理 3，它是减少实例间连接算法的依据.

定理 3. 在 co-location 模式 c 中，假设模糊对象 $A\in c$，如果 A 在 c 的表实例中的实例满足 $\max\{\mu(a)\}<min\_prev$，其中 a 是对象 A 的实例，则 co-location 模式 c 可以被剪枝掉.

证明. 因为

$$FPR(c,A)\leq\frac{\sum_{i=1}^{n}\mu(a_i)}{|table\_instance(A)|}\leq\frac{n\times\max\{\mu(a_i)\}}{|table\_instance(A)|}=\frac{n\times\max\{\mu(a_i)\}}{n}=\max\{\mu(a_i)\},$$

所以当 $\max\{\mu(a)\}<min\_prev$ 时 co-location 模式可以被剪枝(假设模糊对象 A 的实例数是 n).

例 3. 图 2 中，考虑 2 阶 co-location 模式 $c=\{A,B\}$，每个对象的实例序按隶属度非递增进行排序，假设 $min\_prev=0.7$. 对象 A 与对象 B 实例的 R

关系中，$a4$ 与 B 的实例不存在 R 关系，与 B 存在 R 关系的最大隶属度实例为 $a3$，而 $a3$ 的隶属度为 $0.6<min\_prev$，所以模式 $c=\{A,B\}$ 可以被提前剪枝，这样就不用计算实例 $a2$，$a1$ 与对象 B 的实例之间的关系.

从定理 3 可知，只要模式中任何一个对象，它在模式表实例中的实例的最大模糊度小于参与度阈值，则可以对该模式进行剪枝. 于是可以得到减少实例间连接的改进算法，具体见算法 3，其中变量 max_flag 为判断是否最大模糊度实例条件，因为对象的实例顺序已经按照实例模糊度非递增排序，所以在实例连接过程中，第一次实例间的连接即为模式每个对象具有最大模糊度的实例间的连接.

算法 3.　减少实例间的连接算法.

输入：频繁 k 阶模式 $p1_k$，$p2_k$，模式 $p1_k$，$p2_k$ 的表实例 $t1_k$，$t2_k$，$p1_k$，$p2_k$ 连接生成的 $k+1$ 阶候选模式 $c3_{k+1}$

输出：候选模式 $c3_{k+1}$ 的表实例集 $t3_{k+1}$

步骤：

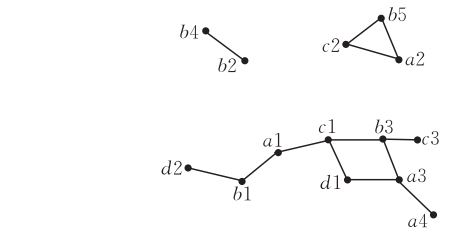
1. $max_flag=$true;
2. for($i=1$; $i\leq t1_k.count$; $i++$)
   2.1. for($j=1$; $j\leq t2_k.count$; $j++$)
      2.1.1. if(($t1_k[i].instance_1=t2_k[j].instance_1\cdots t1_k[i].instance_{k-1}=t2_k[j].instance_{k-1}$) and ($t1_k[i].instance_k$, $t2_k[j].instance_k$) $\in R$) then
         2.1.1.1. if ($max_flag=$true and ($t1_k[i].instance<min\_prev$ or $t2_k[j].instance<min\_prev$)) then goto loop1;
         2.1.1.2. else $t3_{k+1}=t3_{k+1}\cup\{t1_k[i].instance$ join $t2_k[j].instance\}$;
         2.1.1.3. $max_falg=$false;
3. loop1：return $t3_{k+1}$.

4.4　优化剪枝步

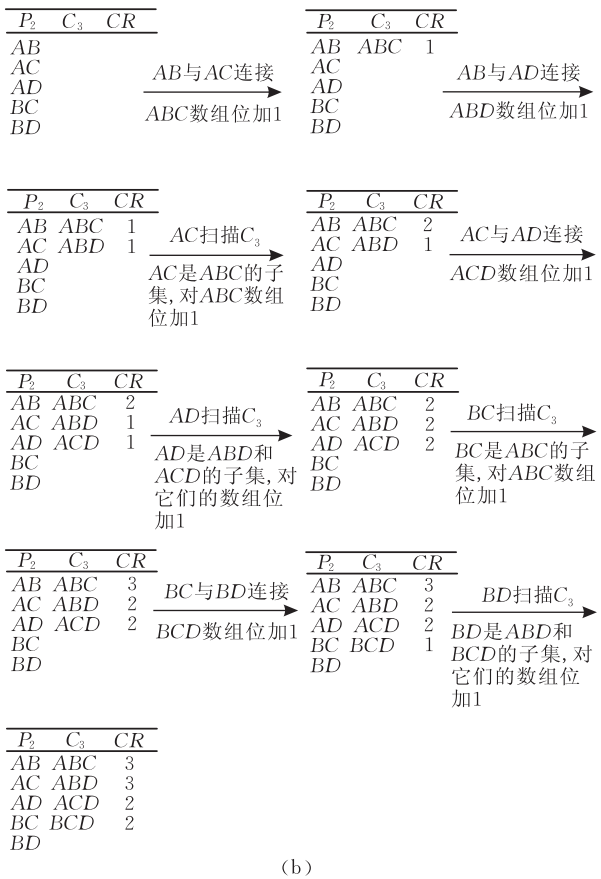
这一小节，本文将提出一种新的改进算法，对 FB 算法中生成候选模式过程中的剪枝步进行优化，在 FB 算法中剪枝步的做法如下：删除所有 C_k 中这样的候选模式 c，c 的 $k-1$ 阶子集不是频繁的. 本文根据第 2 部分中的定理 1 提出一种新的剪枝策略，它不用去比较 c 的每一个 $k-1$ 阶子集是否频繁，就可以对候选模式进行剪枝. 实验和理论分析表明，新的剪枝步策略要比传统的策略具有更好的时间复杂性.

对候选 k 阶模式 C_k 增加一个相应的计数数组. 该计数数组产生的过程和功能简单描述如下：当频繁模式进行连接产生候选 co-location 模式时，增加一个计数数组. 连接过程中每产生一个候选 co-location 模式，数组就相应地增加一位，用来存储该候选模式的频繁 $k-1$ 阶子模式的个数. 由定理 1 可知，每一个 k 阶候选模式，其 $(k-1)$ 阶子模式都是频繁的且为 k 个，所以最后可以根据数组相应位的值来判断对应的候选模式是否频繁. 以下是具体的处理方法和过程.

图 4(a) 中，假设已得到 2 阶频繁模式为 $P_2=\{AB,AC,AD,BC,BD\}$，这时基于 P_2 连接产生 3 阶候选模式过程为：首先增加一个计数数组 CR，对于 P_2 中的模式，计算 AB，由于 C_3 为空，所以对数组 CR 不做任何处理. 接着 AB 开始与 P_2 中其它频繁模式连接产生候选模式，AB 和 AC 连接产生候选



(a)



(b)

图 4　优化剪枝步算法

模式 ABC，这时相应地将数组 CR 增加 1 位，并将其值加 1. AB 与 AD 连接产生候选模式 ABD，数组 CR 增加 1 位，并将其值加 1. AB 连接完成，开始考虑 AC，先扫描 3 阶候选模式，由于 AC 是 ABC 的子集，所以 $CR[1]$加 1. 接着 AC 与 AD 进行连接产生候选模式 ACD，$CR[3]$加 1，依此循环，直到 P_2 中所有频繁模式连接完成，详细过程见图 4(b). 最后 $C_3=\{ABC,ABD,ACD,BCD\}$，数组 $CR=\{3,3,2,2\}$，由于 ACD 和 BCD 计数位不为 3，所以对其进行剪枝.

对于每一个 k 阶候选模式，若其频繁子模式个数不为 k，则对其进行剪枝. 通过证明，新的剪枝步策略比传统的具有更好的时间复杂性，以下给出证明.

证明. 假设 C_k 中有 m 个模式，P_{k-1} 中有 n 个频繁($k-1$)阶模式，在传统剪枝步中，每一个 k 阶模式有 k 个($k-1$)阶子模式需要与 P_{k-1} 进行比较，其时间复杂度为 $k\times m\times n$，而在新的剪枝步策略中，只需对 P_{k-1} 中每一个模式扫描一遍 C_k，其时间复杂度小于 $m\times n$，由此可知新策略的时间复杂度更优. 改进算法的伪代码见算法 4.

算法 4. 优化剪枝步算法.

输入：k 阶频繁模式集 P_k，存储对应的候选模式的频繁子集数目 $CR[]$

输出：$k+1$ 阶候选模式集 C_{k+1}

步骤：

```
1. for(i=1;i≤P_k.count;i++)
 1.1. for all C_{k+1}[x]∈C_{k+1}
  /*判断频繁模式是否为 k+1 阶候选模式的子集*/
  if(P_k[i]中对象都在 C_{k+1}[x]中) then
   CR[x]=CR[x]+1;
 1.2. for(j=i+1; j≤C_k.count; j++)
                          /*k 阶频繁模式间连接*/
  if(P_k[i].object_1=
     P_k[j].object_1…P_k[i].object_{k-1}=
     P_k[j].object_{k-1}) then
    C_{k+1}=C_{k+1}∪{P_k[i] join P_k[j]};
2. for all C_{k+1}[y]∈C_{k+1}
                    /*判断 k+1 阶候选模式是否频繁*/
 if CR[y]≠k then
  remove C_{k+1}[y] from C_{k+1};
3. return C_{k+1}.
```

5 实验与分析

在本节中，本文做了大量实验来验证所提出的 FB 算法和改进算法的有效性，并用文中提出的算法与传统算法的挖掘结果进行了实验比较. 所有算法均采用 C# 编写，并在 AMD Athlon 1.8 GHz CPU 和 512 MB memory 的计算机上运行.

实验所采用的实例数据均是随机产生的并均匀分布在 100×100 空间里，模糊对象的数目为 10，模糊度的值从 0 到 1，也是随机产生的. 表 1 给出了实验的参数以及默认值.

表 1 实验数据的参数说明

参数	默认值
实例数目	500
参与度阈值	0.5
距离阈值	15
模糊度阈值	0

5.1 FB 算法与其改进算法的性能比较

在这一小节中，本文将对模糊对象 co-location 挖掘算法(FB)与其改进算法进行比较，改进算法包括剪枝对象算法(PO)、剪枝对象基础上的减少实例间连接的算法(PO_RI)、剪枝对象且减少实例间连接基础上的优化剪枝步算法(PO_RI_PC).

5.1.1 实例数目对算法的影响

首先考查实例数目对算法的影响，实例数目从 500 增加到 5000. 从图 5 可以看出，随着实例数目增加，所有算法的运行时间都增加，因为随着实例数目的增加，算法实例间距离计算、连接操作也会增加. PO_RI 和 PO_RI_PC 算法由于采用了减少实例间的连接改进算法，故随着实例数目的增加，它的处理时间上升的幅度比较小，这也说明了减少实例间连接改进算法的高效性.

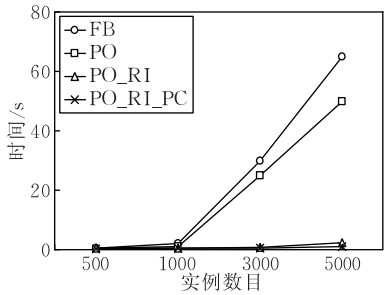


图 5 实例数目对算法的影响

5.1.2 参与度阈值对算法的影响

下面研究参与度阈值对算法的影响，参与度阈值的变化从 0.8 到 0.2. 从图 6 中可以看出，FB 算法和 PO 算法的运行时间随着参与度阈值的降低而急剧上升，而 PO_RI 算法和 PO_RI_PC 算法的运行时间一直保持平稳，这是因为随着参与度阈值的降低，更多的 co-location 模式满足参与度阈值条件，

使得 FB 和 PO 算法运行时间急剧上升，而 PO_RI 和 PO_RI_PC 算法由于采用了减少实例间的连接改进算法，所以运行时间一直较平稳. 从图中还可以看出当参与度阈值在 0.4 到 0.2 区间时，FB 和 PO 算法运行时间基本一样，这是因为 PO 算法采用的是剪枝模糊对象改进算法，它依赖于参与度阈值，当自定义的参与度阈值非常低时，算法很难剪枝掉很多的模糊对象，所以两个算法的运行时间相差不大.

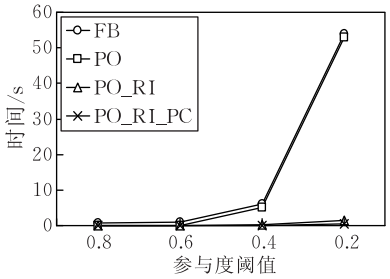


图 6　参与度阈值对算法的影响

5.1.3　距离阈值对算法的影响

接下来考虑距离阈值对算法运行时间的影响，距离阈值变化从 10～40. 从图 7 中可以看出，类似于实例数目对算法的影响，随着距离阈值增大，FB 和 PO 算法运行时间快速上升，而 PO_RI 和 PO_RI_PC 算法上升较平稳.

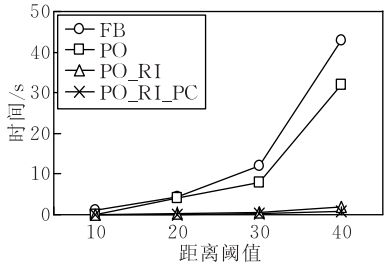


图 7　距离阈值对算法的影响

5.1.4　模糊度阈值对算法的影响

本组实验的最后考虑模糊度阈值对算法的影响，模糊度阈值从 0 到 0.5 变化. 从图 8 中可以看出，当模糊度从 0 到 0.2 时，4 个算法的运行时间均上升，而在 0.2 以后，4 个算法的运行时间又开始下降，这是因为把模糊对象的那些不满足模糊度阈值的，具有低模糊度的实例剪去后，剪枝掉实例的对象的参与率值会增大，这时满足参与度阈值条件的模式数目会增多，使得算法运行时间上升. 但是随着模糊度阈值的不断增大，要剪去对象的实例也越多，这就意味着一个对象具有的实例数目越来越少，就会减少模式产生的数目，使得算法的运行时间下降.

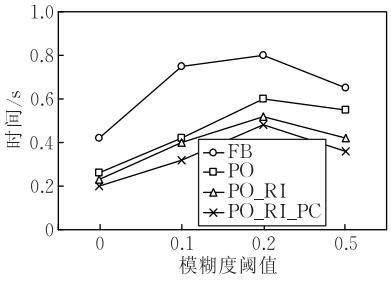


图 8　模糊度阈值对算法的影响

5.2　模糊挖掘算法与传统算法的比较

在这一小节，本文用模糊对象 co-location 挖掘算法，与传统挖掘算法的结果进行实验比较. 这里传统算法采用的是 co-location 挖掘算法中最为经典的 join_based 算法[2]. 考虑实例个数、参与度阈值、距离阈值、模糊度阈值对两种算法的影响，其中模糊对象 co-location 模式挖掘算法采用的是 PO_RI_PC 算法.

5.2.1　实例个数对算法的影响

首先，考虑实例个数对两种不同算法的影响，实例数目从 500 增加到 5000. 从图 9 中可以看出，两种算法的模式数目都随着实例数目的增加而增加. Join_based 算法产生的模式数目远大于 PO_RI_PC 算法，这是因为 PO_RI_PC 算法模式参与度的计算是根据实例的模糊度来计算的，而传统算法中没有区分实例之间模糊度的差别，这导致了后者的参与度要高于前者，在相同参与度阈值条件下，后者频繁模式数目要明显高于前者.

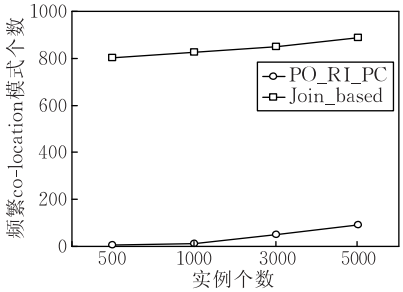


图 9　实例个数对算法的影响

5.2.2　参与度阈值对算法的影响

接下来考虑参与度阈值对算法的影响，阈值变化从 0.8 到 0.2. 从图 10 中可以看出两种算法的模

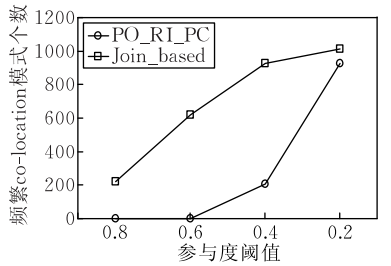


图 10　参与度阈值对算法的影响

式数目随着阈值降低都在增加，在低参与度阈值条件下，比如阈值为 0.2，两者的频繁模式数目比较接近.

5.2.3 距离阈值对算法的影响

接下来考虑距离阈值对算法的影响，距离阈值的变化从 10～40. 从图 11 中可以看出，两种算法的模式数目都上升，但 PO_RI_PC 算法上升较慢，因为限制频繁模式个数的因素还有参与度.

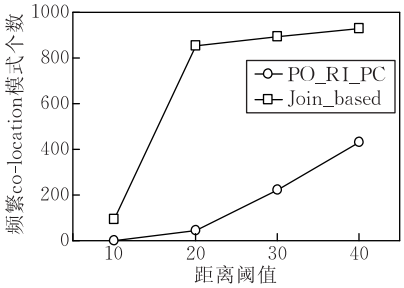


图 11 距离阈值对算法的影响

5.2.4 模糊度阈值对算法的影响

最后来看模糊度阈值对算法的影响，模糊度阈值的变化从 0～0.5. 从图 12 中可以看出，Join_based 算法的模式个数保持不变，因为模糊度没有参与模式参与度的计算中. 而 PO_RI_PC 算法生成的模式个数开始上升，后来出现下降，原因可以见实验 5.1.4 节中的分析.

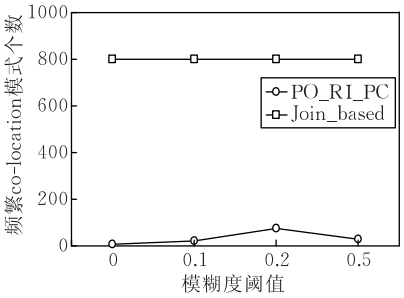


图 12 模糊度阈值对算法的影响

6 结 论

尽管空间 co-location 模式挖掘是一种非常有价值的空间挖掘，而且模糊对象也经常出现在许多重要的应用中，但是目前对于模糊对象的 co-location 模式挖掘的研究还未见报道. 本文针对模糊对象的空间 co-location 模式挖掘问题，提出了一种基本挖掘算法——FB 算法，为了提高算法的挖掘效率，文中提出了 3 种改进算法，包括剪枝对象、减少实例间连接、改进剪枝步. 通过大量的实验表明，本文提出的算法及改进算法是非常有效的. 下一步的工作将在此论文的基础上，考虑模糊度阈值在一个范围内变化时的空间 co-location 模式挖掘问题.

参 考 文 献

[1] Wang Li-Zhen, Zhou Li-Hua, Chen Hong-Mei et al. The Principle and Applications of Data Warehouse and Data Mining. 2nd Edition. Beijing: Science Press, 2009 (in Chinese)
(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. 北京：科学出版社，2009)

[2] Huang Y, Shekhar S, Xiong H. Discovering colocation patterns from spatial data sets: A general approach. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(12): 1472-1485

[3] Yoo J S, Shekhar S. A partial join approach for mining co-location patterns//Proceedings of the ACM International Symposium on Advances in Geographic Information Systems (ACMGIS). Washington, USA, 2004: 241-249

[4] Yoo J S, Shekhar S, Celik M. A join-less approach for co-location pattern mining: A summary of results//Proceedings of the IEEE International Conference on Data Mining (ICDM). Houston, USA, 2005: 813-816

[5] Wang Li-Zhen, Bao Yu-Zhen, Lu J, Yip J. A new join-less approach for co-location pattern mining//Proceedings of the IEEE 8th International Conference on Computer and Information Technology (CIT 2008). Sydney, Australia, 2008: 197-202

[6] Wang Li-Zhen, Zhou Li-Hua, Lu J. Yip J. An order-clique-based approach for mining maximal co-locations. Information Sciences, 2009, 179(19): 3370-3382

[7] Lu Ye, Wang Li-Zhen, Zhang Xiao-Feng. Mining frequent Co-location patterns from uncertain data. Journal of Frontiers of Computer Science and Technology, 2009, 3(6): 656-664 (in Chinese)
(陆叶，王丽珍，张晓峰. 从不确定数据集中挖掘频繁 Co-location 模式. 计算机科学与探索，2009，3(6)：656-664)

[8] Wang Li-Zhen, Chen Hong-Mei, Zhao Li-Hong et al. Efficiently mining co-location rules on interval data//Proceedings of the 6th International Conference on Advanced Data Mining and Applications (ADMA 2010). Chongqing, China, 2010: 477-488

[9] Zadeh L. Fuzzy sets. Information and Control, 1965, 8(3): 338-353

[10] Altman D. Fuzzy set theoretic approaches for handling imprecision in spatial analysis. International Journal of Geographical Information Science, 1994, 8(3): 271-289

[11] Schneider M. Fuzzy topological predicates, their properties, and their integration into query languages//Proceedings of the ACM International Symposium on Advances in Geographic Information Systems (ACMGIS). New York, USA, 2001: 9-14

[12] Schneider M. Uncertainty management for spatial data in databases: Fuzzy spatial data types//Proceedings of the International Symposium on Advances in Spatial Databases. Berlin, Germany, 1999: 330-351

[13] Tang X, Kainz W. Analysis of topological relations between fuzzy regions in a general fuzzy topological space//Proceedings of the Symposium on Geospatial Theory, Processing and Applications. Ottawa, Canada, 2002: 114-123

[14] Zheng Kai, Fung Pui-Cheong, Zhou Xiao-Fang. *K*-nearest neighbor search for fuzzy objects//Proceedings of the Special Interest Group on Management of Data (SIGMOD'10). Indiana, USA, 2010: 699-710

OUYANG Zhi-Ping, born in 1985, M. S. candidate. His main research interests include spatial data warehouse and spatial data mining.

WANG Li-Zhen, born in 1962, professor, Ph. D. supervisor. Her main research interests include data warehouse, data mining and computer algorithms.

CHEN Hong-Mei, born in 1976, Ph. D. candidate, lecturer. Her main research interests include spatial data warehouse and spatial data mining.

Background

The mining co-location pattern is one of the most important areas in spatial databases, due to its broad range of applications including location-based services (LBS), public health, transportations and so on. In the past, a lot of researchers had studied this topic for certain and uncertain data. Many algorithms had been proposed, such as join-based, partial-join, join-less, CPI-tree, order-clique-base, etc., but there is not exploration for mining co-locations on fuzzy data. As is known to all, the fuzzy data can be seen everywhere in our lives. So in this paper, we study the problem of mining co-location pattern from fuzzy objects.

Firstly, we define some related concepts, such as fuzzy participation ratio, fuzzy participation index and so on. Based on these definitions, we put forward an FB algorithm. Then, we optimize FB algorithm in three different stages: (1) In order to reduce the number of fuzzy objects, an improving algorithm called pruning objects is proposed; (2) we put forward an effective pruning algorithm: reducing the operations of joining between instances, by which co-location patterns can be pruned in advance; (3) an algorithm of optimizing the pruning steps is raised to improve the efficiency of the traditional pruned steps.

This work is supported by the National Natural Science Foundation of China (61063008), the Application Basic Research Foundation of Yunnan Province (2010CD025), and the Scientific Research Foundation of Yunnan University (2009F29Q).