附加資料探勘資訊的 R-tree 資料存取

陳靖國*
朝陽科技大學資訊管理系
jkchen@cyut.edu.tw

簡汶婷
朝陽科技大學資訊管理系
s9414633@cyut.edu.tw

摘要

資料探勘應用一直是熱門的研究主題，可在不同領域上擷取各種資訊，例如應用於空間資料索引。常用索引可分為點資料索引和空間資料索引，目前已有將資料探勘應用在點資料索引的研究，但少有應用在空間資料索引的研究。R-tree 是目前最受歡迎的空間資料索引，我們將資料探勘概念應用到 R-tree，使探勘項目具有空間物件特性，基本項目包括物件面積、平均面積、面積標準差、面積最小的物件、面積最大的物件，並依據空間物件具有重疊特性，另外提出三個特別項目，亦即涵蓋所有重疊區域的最小矩形、面積最大的重疊區域和同時被最多物件重疊的區域。預先計算葉節點的資料探勘項目，從 R-tree 底層逐層往上統計探勘項目，R-tree 每個節點皆包含彙總性資訊。最後以農藥污染區分佈為例，說明彙總性資訊提供的實際效益，亦即有些資訊在 R-tree 上層即可取得，省去搜尋和計算大部分下層資料的時間。

關鍵字：資料探勘，空間資料，R-tree，重疊

Abstract

Data mining application is a popular research topic. It can extract many kinds of information from different applications. For example, data mining can be applied to spatial data index. Common index can be classified to point data index and spatial data index. There are some researches of data mining applied to point data index at present, but rare are applied to spatial data index. R-tree is the most popular spatial data index at present. Therefore, we try to apply data mining to R-tree. To make mining items having spatial object characteristics. Basic mining items are object area, mean of object area, standard deviation of object area, the object with the smallest area, and the object with the largest area. According to the overlap character of spatial object, we add three special mining items, the smallest cover range of total objects Overlap Region, The Largest Area of Overlap Region and Overlap Region with the Most Overlap Objects. We calculate the values of the above mining items in the leaf nodes and repeat the action on the nodes level-by-level from bottom to the top of the tree. Finally, we use an example of pesticide residue contamination districts to explain the real effect of the statistical information in our research. We can get the statistical information from a node at an upper level of R-tree. Thus, the search and computing time of the data at lower levels can be omitted.

Keyword: data mining, spatial data, R-tree, overlap

1.前言

資料探勘是指從大量資料中，擷取出有意義的潛在資料[8]。資料探勘近來的應用相當熱門，各個領域都可以加以應用，例如：電子商務、財務應用、企業顧客關係管理、購物交叉分析、顧客關係行銷、決策支援、網頁搜尋、

地理資訊系統、全球衛星定位導航、影像資料查詢等應用[2,9,10,16,17,18]。若將資料探勘應用於物件資料查詢上，則能擴大查詢結果的附加效益。目前較具有代表性的物件資料索引有Grid file [13]、K-D Tree [1]、K-D-B Tree [14]和 R-Tree [7]，可以分成兩種[5]。一種是針對點資料(Point Data)建立索引，例如 Grid file、K-D Tree、K-D-B Tree，這些方法是以一個點資料代表一個物件，利用分割(Partition)的方法，將不同點資料所隸屬的空間做區分來建立索引。另一種是針對空間資料(Spatial Data)建立索引，例如 R-tree，以一個涵蓋空間物件的最小矩形 (Minimum Bounding Rectangle, 簡稱 MBR)代表一個空間物件，MBR 之間可能互相重疊。

目前已有資料探勘應用於索引的方法被提出[3,4,6,13,15,19]，但只應用於點資料索引，如 Statistical Information Grid (STING) [15]，但少有應用於空間資料索引的研究。R-Tree 是目前最受歡迎的空間資料索引，因此我們試著將資料探勘概念應用在空間資料索引上，並且以農藥污染區分佈為例子，說明本研究依據空間位置給予對應的彙總性資訊和空間資訊，例如特定縣市農藥污染區的總面積、平均面積、最大的污染區、最小的污染區、涵蓋所有污染區的最小範圍、污染面積最大的區域和污染最嚴重的區域，以輔助說明空間資料索引加上資料探勘能帶來的便利性和實用性。

2.相關工作

2.1 R-tree

R-tree是由Guttman [7]所提出的階層式架構的多維度索引技術，是一種類似B-tree的高度平衡樹(height-balanced tree)，可用應用於多維度空間物件，是一種動態索引資料結構。R-tree 由葉節點(leaf node)和中間節點(intermedia node)所組成，每個節點皆包含若干個entries，一個entry代表某個物件或某個子節點相關資訊，每一個entry包含涵蓋物件的最小矩形(Minimum Bounding Rectangle, 簡稱MBR)與指標(pointer)。葉節點的entry格式為*(I, tuple-identifier)*，*I*表示涵蓋空間物件的MBR，而*tuple-identifier*表示一個指向空間物件所在位址的指標；中間節點的entry格式為*(I, child-pointer)*，*I*表示涵蓋子節點的MBR，而*child-pointer*表示一個指向子節點所在位址的指標。在R-tree的動態索引資料結構中，根節點會涵蓋所有子節點所涵蓋的範圍，而葉節點的entries會儲存每個實際物件的資訊。假設*M*為每個節點所能擁有的最大entry數量(最大分支度)，而*m*為每個節點至少擁有的最小entry數量(最小分支度)，則R-tree必須滿足以下特性：(1)根節點至少有二個子節點，除非根節點同時也是葉節點。(2)每個非根節點的entries數量必須介於m和M之間。(3)所有葉節點都出現在同一層。

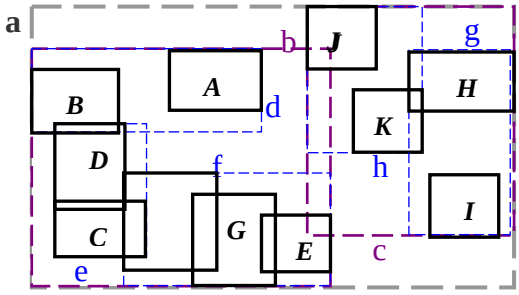

圖 1(a) 、散佈在資料空間的物件

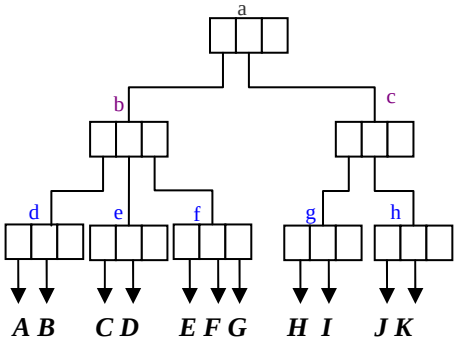

圖1(b)、對應的R-tree結構

圖1為*m*=2，*M*=3之R-tree範例，圖1(a)表示物件在資料空間散佈的情形，而圖1(b)表示對應的R-tree結構。其中*d*、*e*、*f*、*g*和*h*代表葉

節點，透過葉節點entry的*tuple-identifier*以存取*A*、*B*、…、*K*物件本身。*a*、*b*、*c*代表中間節點，其中根節點*a*包含*b*和*c*兩個節點，*b*包含*d*、*e*和*f*三個葉節點，*c*包含*g*和*h*兩個葉節點。從圖**1**可得知，根節點完全覆蓋子節點的MBR，並且允許MBR之間可有重疊情況發生。

2.2 STING

Statistical Information Grid (STING)是一種以網格(grid cell)為基礎的階層式統計資訊方法[15]，並且應用於點資料上。STING 將資料空間分割成許多小格子，我們稱之為網格，最低層網格會預先計算並儲存相關的統計資訊，由於 SITNG 具有階層式架構，因此高階層網格會彙總低階層網格的相關統計資訊。STING 階層式架構必須包含以下兩點：(1)根節點位於第一層，表示所有的資料空間；中間節點則位於第二層以後，依此類推。(2)除了葉節點，第 *i* 階層節點會具備第 *i+1* 階層節點的彙總性統計資訊。圖 2 表示 STING 於二維空間的階層式架構[15]。

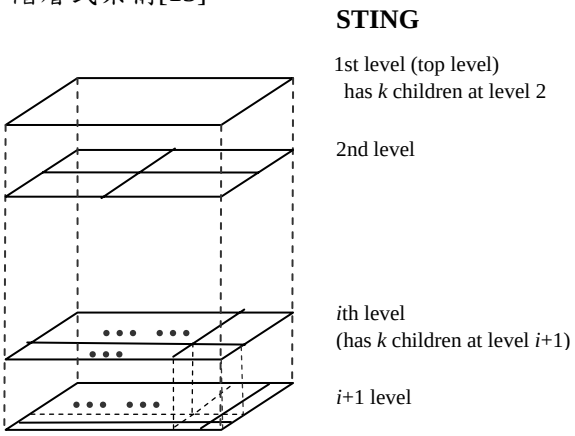

圖 2、STING 於二維空間的階層式架構

STING 提出五個主要資料探勘項目：(1) *n*：空間物件的個數。(2) *m*：空間物件的平均值。(3) *s*：空間物件的標準差。(4) *min*：空間物件裡最小的值。(5) *max*：空間物件裡最大的值。

3.資料探勘應用於空間資料

3.1 資料探勘項目

空間資料和點資料最明顯的差別為空間資料有面積資料的存在，而點資料沒有，所以面積相關的探勘是重要項目之一，因此我們以STING資料探勘項目為概念，將其應用於空間物件索引上，針對物件的面積提出下列資料探勘項目，分別為面積值(*area*,簡稱*a*)、平均面積值(*mean of area*,簡稱m_a)、面積值的標準差(*standard deviation of area*,簡稱s_a)、面積最小的物件名稱(*minimum area*,簡稱min_a)、面積最大的物件名稱(*maximum area*,簡稱max_a)。同時依據空間物件可能重疊的特性，再提出三個資料探勘項目，分別為涵蓋所有重疊區域的最小矩形(*Minimum Rectangle of Cover Total Overlap Region*,簡稱*MRCTOR*)、面積最大的重疊區域(*The Largest Area of Overlap Region*,簡稱*LAOR*)和同時被最多物件重疊的區域(*Overlap Region with the Most Overlap Objects*,簡稱*ORMOO*)，使資料探勘應用更加廣泛和實用。下述將依序說明這些資料探勘項目的定義與應用性。

(1) *a*：空間物件的總面積值

R-tree 以 MBR 記錄涵蓋物件的最小矩形，因此我們以 *a* 記錄空間物件的佔用面積，表示空間物件的總面積值。*a* 數據的計算公式為 $a=\sum_i a_i$，其中 a_i 為 *a* 的下一階層某個子節點所包含的面積值。亦即，*a* 是由下一階層所有的 a_i 所包含的面積值相加所得，而每一個 a_i 又是由下一階層節點 *i* 所有的子節點所包含的面積值相加所得，依此遞迴而得到總面積值。

(2) m_a：空間物件的平均面積值

我們以m_a代表空間物件的平均面積值。m_a數據的計算公式為 $m_a=\frac{\sum_i m_{ai}n_i}{n}$，其中 m_{ai}為含m_a的節點*p*下一階層某個子節點*i*所包含的平均面積值、n_i為子節點*i*所包含的個數、*n*為節點*p*的總個數。亦即，m_a是由下一階層所有的

n_i所包含的個數與m_{ai}所包含的平均面積值相乘的總和，再除以該階層n所包含的總個數所得。而每一個m_{ai}又是由下一階層節點i所有的子節點所包含的個數與平均面積值相乘的總和，再除以該階層所包含的個數所得，依此遞迴而得到平均面積值。

(3) s_a：空間物件面積值的標準差

我們以s_a代表空間物件的標準差，以該值判斷各個空間物件面積值的分散程度。s_a數據的計算公式為 $s_a = \sqrt{\frac{\sum_i (s_{ai}^2 + m_{ai}^2) n_i}{n} - m_a^{\,2}}$，其中 s_{ai} 為含s_a的節點p下一階層某個子節點i所包含面積值的標準差、m_{a_i}為子節點i所包含的平均面積值、n_i為子節點i所包含的個數、n為節點p的總個數。亦即，s_a是由下一階層所有的s_{ai}所包含的面積值的標準差平方與m_{ai}所包含的平均面積值平方相加，再乘n所包含的個數的總和，並除以該階層n所包含的總個數，減掉該階層m_a所包含的平均面積值平方後，再開根號所得。而每一個s_{ai}又是由下一階層節點i所有的子節點所包含的面積值的標準差平方與平均面積值平方相加，再與個數相乘，並除以該階層所包含的總個數，再與該階層所包含的平均面積值平方相減後開根號所得，依此遞迴而得到面積值的標準差。

(4) min_a：空間物件中面積最小的物件名稱

我們以min_a代表所有的空間物件裡面積最小的物件名稱，將來執行查詢動作時，則能快速得知面積最小的空間物件是哪一個。其判斷的順序為最優先選擇面積最小的空間物件。若面積最小的空間物件有兩個以上，則從中選擇本身與其它空間物件重疊數量最多的該空間物件。由於空間物件之間的重疊是相當重要的一個特性，因此我們以重疊數量最多的空間物件做為判斷的準則之一。因此min_a必須包含總共三個探勘項目(O_{min}, a_{Omin}, N_{Omin})，其中O_{min}為最小物件名稱，a_{Omin}為O_{min}的面積值，N_{Omin}為與O_{min}重疊物件個數，後二項是為輔助判斷找出最小物件的數據，min_a的演算法如下。

Algorithm $min_a(p)$
Input: R-tree node pointer p;
Output: structure min_a;
Begin
  for each entry $entry_i$ of p, do
    find out the $entry_i$ with the minimum a_i;
      if $entry_i$ has more than one, then
        find out the $entry_i$ with the maximum number of overlap objects;
      end if;
  end for;
  return $entry_i$'s name, area and the number of overlap objects;
End min_a.

(5) max_a：空間物件中面積最大的物件名稱

我們以max_a代表所有的空間物件裡面積最大的物件名稱，將來執行查詢動作時，則能快速得知面積最大的空間物件是哪一個。同理，其判斷順序為最優先選擇面積最大的空間物件。若面積最大的空間物件有兩個以上，則從中選擇本身與其它空間物件重疊數量最多的該空間物件。因此max_a必須包含總共三個探勘項目(O_{max}, a_{Omax}, N_{Omax})，其中O_{max}為最大物件名稱，a_{Omax}為O_{max}的面積值，N_{Omax}為與O_{max}重疊物件個數，後二項是為輔助判斷找出最大物件的數據。max_a的演算法如下。

Algorithm $max_a(p)$
Input: R-tree node pointer p;
Output: structure max_a
Begin
  for each entry $entry_i$ of p, do
    find out the $entry_i$ with the maximum a_i;
      if $entry_i$ has more than one, then

find out $entry_i$ with the maximum numbers of overlap objects;
end if;
end for;
return $entry_i$'s name, area and the number of overlap objects;
End max_a.

(6) *MRCTOR*：涵蓋特定範圍內所有重疊區域的最小矩形

我們以*MRCTOR*代表涵蓋所有重疊區域的最小矩形(*Minimum Rectangle of Cover Total Overlap Region, MRCTOR*)，透過該項目能立即得知包含特定範圍裡全部的重疊區域的最小矩形範圍。其判斷的方法為找出該節點內空間物件MBR所有重疊的區域，再以一個最小矩形涵蓋這些區域，並且在該父節點上記錄該矩形範圍，因此*MRCTOR*為 $((x_1,y_1),(x_2,y_2))$，(x_1,y_1)為*MRCTOR*之左下角座標值，(x_2,y_2)為*MRCTOR*之右上角座標值，*MRCTOR*的演算法如下。

Algorithm *MRCTOR*(*p*)
Input: R-tree node pointer *p*;
Output: structure *MRCTOR*;
Begin
for each entry $entry_i$ of *p*, do
find out all overlap regions;
end for;
use a minimum rectangle *M* to completely cover all overlap regions;
assume *M*'s left lower coordinate is (x_1, y_1) and *M*'s right upper coordinate is (x_2, y_2) return $(M.x_1, M.y_1, M.x_2, M.y_2)$;
End *MRCTOR*.

圖 3 為涵蓋所有重疊區域的最小矩形之範例，其中節點A涵蓋物件名稱分別為 1、2、3、4 和 5 這五個空間物件，圖 3(a)表示所有重疊的區域，分別為(物件 3 與物件 4)和(物件 3 與物件 5)的MBR彼此有重疊，我們用斜線加以標示。圖 3(b)表示涵蓋所有重疊區域的最小矩形即為涵蓋所有重疊區域的最小矩形，如斜線所標示，並且在節點A上記錄該矩形範圍，即 $MRCTOR = ((x_1,y_1),(x_2,y_2))$。

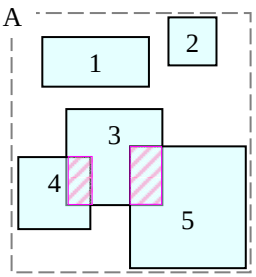

圖 3(a)、所有重疊的區域

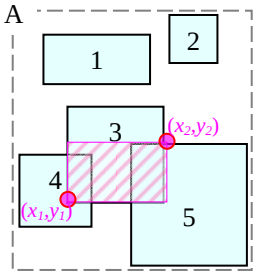

圖3(b)、涵蓋所有重疊區域的最小矩形

(7) *LAOR*：特定範圍裡面積最大的重疊區域

我們以*LAOR*代表面積最大的重疊區域(*The Largest Area of Overlap Region, LAOR*)，透過該項目能立即得知特定範圍內重疊面積最大的區域。其判斷的方法為找出該節點內空間物件MBR所有重疊的區域，從中選擇重疊面積最大的區域，並且在該父節點上記錄該矩形範圍。若重疊面積最大的區域有兩個以上，則從中選擇同時被最多物件覆蓋的重疊區域，並且在該父節點上記錄該矩形範圍。因此*LAOR*必須包含總共二個探勘項目$(Rect_{LAOR}, N_{RectLAOR})$，其中$Rect_{LAOR} = ((x_1,y_1),(x_2,y_2))$，$(x_1,y_1)$為*LAOR*之左下角座標值，$(x_2,y_2)$為*LAOR*之右上角座標值，$N_{RectLAOR}$為與$Rect_{LAOR}$重疊物件個數，且為輔助判斷找出面積最大的重疊區域的數據，*LAOR*的演算法如下。

Algorithm *LAOR*(*p*)
Input: R-tree node pointer *p*;

Output: structure *LAOR*;
Begin
structure *block*{(x_1,y_1),(x_2,y_2)}; //用以記錄重疊區域的範圍//
for each entry $entry_i$ of *p*, do
find out the *block* with the largest overlap region;
if *block* has more than one, then
find out *block* with the maximum number of overlap objects;
end if;
end for;
use a minimum rectangle *L* to cover *block*'s overlap regions;
assume *L*'s left lower coordinate is (x_1, y_1) and *L*'s right upper coordinate is (x_2, y_2) return (*L*.x_1, *L*.y_1, *L*.x_2, *L*.y_2) and the number of overlap objects;
End *LAOR*.

圖 4 為面積最大的重疊區域之範例，其中節點A涵蓋物件名稱分別為 1、2、3、4 和 5 這五個空間物件，圖 4(a)表示所有重疊的區域，分別為(物件 1 與物件 3) 、(物件 1 與物件 4)、(物件 3 與物件 4)、(物件 1、物件 3 與物件 4)和(物件 2 與物件 5)的MBR彼此有重疊，如斜線所標示。圖 4(b)表示面積最大的重疊區域，如斜線所標示，由於圖 4(a)五組的重疊區域湊巧皆是重疊面積最大的區域，所以進一步從中選擇重疊物件個數最多的區域，因此面積最大的重疊區域為(物件 1、物件 3 與物件 4)共同重疊的區域，並且在節點A上記錄該矩形範圍和重疊物件個數，即*LAOR* = ((x_1,y_1),(x_2,y_2)), 3)。

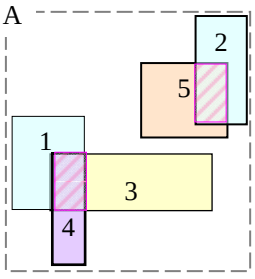

圖4(a)、所有重疊的區域

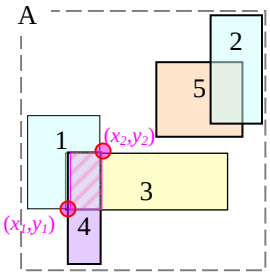

圖4(b)、面積最大的重疊區域

(8) *ORMOO*：特定範圍裡同時被最多物件重疊的區域

我們以*ORMOO*代表同時被最多物件重疊的區域(*Overlap Region with the Most Overlap Objects, ORMOO*)，透過該項目能立即得知特定範圍內被最多物件同時重疊的區域。其判斷的方法為找出該節點內空間物件MBR所有重疊的區域，從中選擇被最多物件同時覆蓋的重疊區域，並在其父節點上記錄該矩形範圍。若有兩個以上同時被最多物件覆蓋的重疊區域，則從中選擇最大的重疊區域，並在其父節點上記錄該矩形範圍。因此*ORMOO*必須包含總共二個探勘項目($Rect_{ORMOO}$, $N_{RectORMOO}$)，其中$Rect_{ORMOO}$ = ((x_1,y_1),(x_2,y_2))，(x_1,y_1)為*ORMOO*之左下角座標值，(x_2,y_2)為*ORMOO*之右上角座標值，並且為輔助判斷找出面積最大的重疊區域的數據，$N_{RectORMOO}$為與$Rect_{ORMOO}$重疊物件個數，*ORMOO*的演算法如下。

Algorithm *ORMOO*(*p*)
Input: R-tree node pointer *p*;
Output: structure *ORMOO*;
Begin
structure *block*{(x_1,y_1),(x_2,y_2)}; //用以記錄重疊區域的範圍//
for each $entry_i$ of *p*, do
find out the *block* with the maximum number of overlap objects;
if *block* has more than one, then

find out the *block* with the largest overlap region;

end if;

end for;

use a minimum rectangle *O* to cover *block*'s overlap regions;

assume *O*'s left lower coordinate is (x_1, y_1) and *O*'s right upper coordinate is (x_2, y_2) return ($O.x_1$, $O.y_1$, $O.x_2$, $O.y_2$) and the number of overlap objects;

End *ORMOO*.

圖 5 為同時被最多物件重疊的區域之範例，其中節點A涵蓋物件名稱分別為 1、2、3、4 和 5 這五個空間物件，圖 5(a)表示所有重疊的區域，分別為(物件 1 與物件 4)和(物件 2 與物件 5)的MBR彼此有重疊，如斜線所標示。圖 5(b)表示同時被最多物件重疊的區域，如斜線所標示，由於圖 5(a)二組的重疊區域湊巧皆是重疊物件個數最多的區域，所以進一步從中選擇重疊面積最大的區域，因此面積最大的重疊區域為(物件 2 與物件 5)的區域，並且在節點A上記錄該矩形範圍和重疊物件個數，即 *ORMOO* = ((x_1,y_1),(x_2,y_2)), 2)。

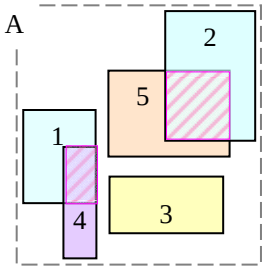

圖5(a)、所有重疊的區域

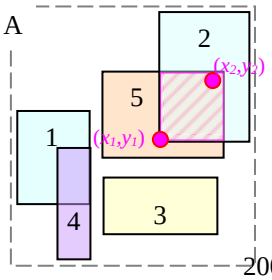

圖5(b)、同時被最多物件重疊的區域

3.2 實例介紹

我們變更R-tree節點的資料結構，以包含上述的資料探勘項目，變更後R-tree節點的資料結構為(*E[1,2,⋯,M],a,* m_a , s_a , min_a , max_a , *MRCTOR, LAOR, ORMOO*)，其中*E[i]*代表子節點的entry，當子節點為非葉節點時，則每個*E[i]*包含一對的(I_i, *child-pointer*$_i$)，當子節點為葉節點時，則每個*E[i]*包含一對的(I_i, *tuple-identifier*$_i$)，*M*值是每個節點所包含的 entries最大數量。從葉節點取得物件相關基本資料後，再經過簡單運算以得到彙總性資訊，並且在父節點記錄上述資訊，如此反覆處理，從底層開始逐漸往上加總和記錄彙總性資訊，持續至最上層根節點為止。

我們舉例說明資料探勘應用於R-tree所帶來的附加效益，為了增加可讀性與容易理解，我們以簡單數據表達其便利性。假設針對台灣某二個縣市的農藥污染區分佈做資料探勘，圖 6(a)表示農藥污染區的分佈情形，圖 6(b)表示相對應的 3 層R-tree空間物件索引，每一個節點至少包含二個子節點或物件最多包含三個。圖 6 表示針對台灣某二個縣市農藥污染做探討，根節點(Root)包含***a***和***b***二個節點，分別代表***a***縣和***b***縣的農藥污染區，其中***a***縣的農藥污染區又涵蓋***c***、***d***和***e***三個節點，分別代表***c***鄉鎮、***d***鄉鎮和***e***鄉鎮的農藥污染區，***b***縣的農藥污染區又涵蓋***f***和***g***二個節點，分別代表***f***鄉鎮和***g***鄉鎮的農藥污染區。其中***c***鄉鎮的農藥污染區包含O_1、O_2和O_3三個節點，分別代表O_1、O_2和O_3農藥污染區，***d***鄉鎮的

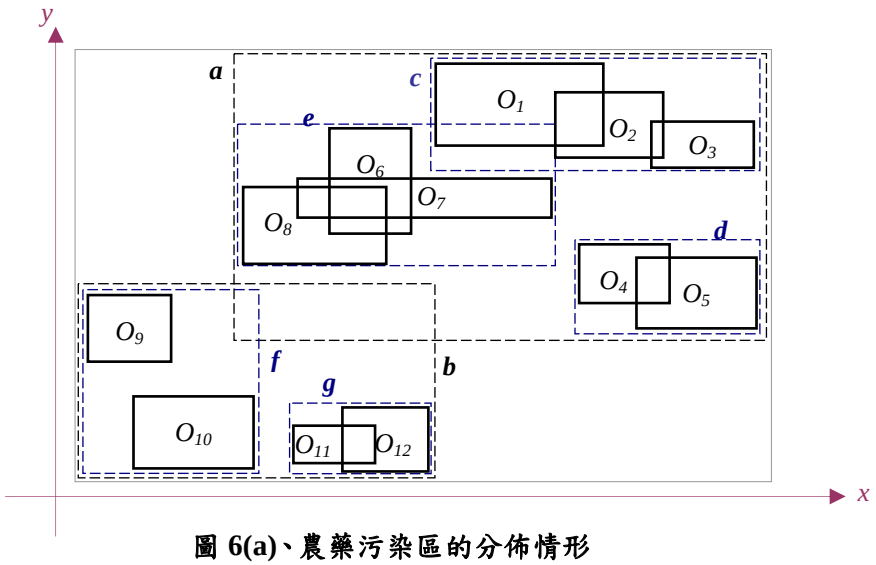

圖 6(a)、農藥污染區的分佈情形

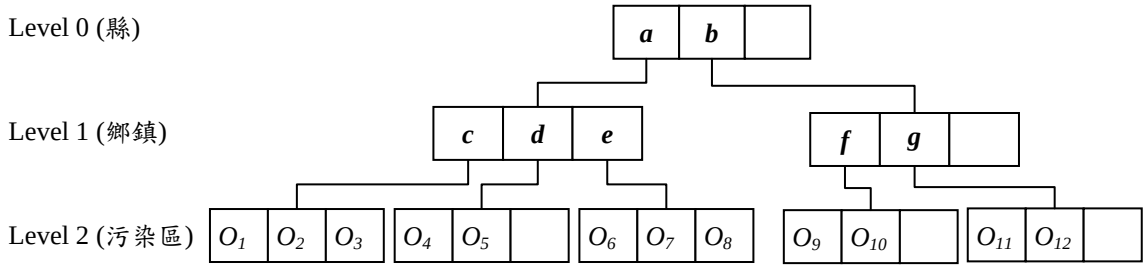

圖 6(b)、對應的 R-tree 空間物件索引結構

農藥污染區包含O_4和O_5二個節點，分別代表O_4和O_5農藥污染區，***e***鄉鎮的農藥污染區包含O_6、O_7和O_8三個節點，分別代表O_6、O_7和O_8農藥污染區，***f***鄉鎮的農藥污染區包含O_9和O_{10}三個節點，分別代表O_9和O_{10}農藥污染區，***g***鄉鎮的農藥污染區包含O_{11}和O_{12}二個節點，分別代表O_{11}和O_{12}農藥污染區。

應用於此案例，則R-tree節點將記錄著特定區域裡的農藥污染區相關資訊，*a*記錄污染區的面積值、m_a記錄污染區平均面積值、s_a記錄污染區面積的標準差、min_a記錄面積最小的污染區名稱、max_a記錄面積最大的污染區名稱、*MRCTOR*記錄涵蓋所有多重污染區域的最小範圍、*LAOR*記錄面積最大的多重污染區

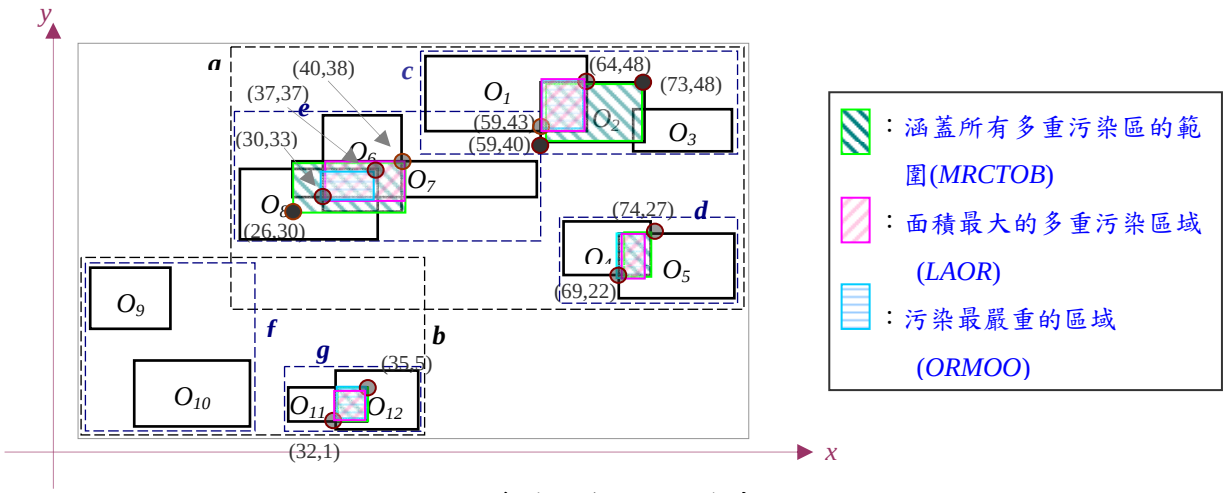

圖 7、多重污染區的分佈情形

域、*ORMOO* 記錄污染最嚴重的區域。

葉節點O_1、O_2、O_3、…、O_{12}分別代表各個農藥污染區，我們從葉節點得到資料來源(如表 1 所示)，取得來源後，即可進行運算判斷動作，並記錄在父節點上，逐層由底層往上記錄，反覆運算判斷動作直到記錄到根節點為止(如表 2 和表 3 所示)，使得每一個節點皆具有彙總性資訊和空間資訊，進一步衍生出表 1、表 2、表 3 和圖 7 得知相對應的資料探勘項目資訊，其中我們以null表示沒有被重覆污染的農藥污染區。

在R-tree節點上附加資料探勘項目，除了使每一個節點皆包含許多豐富的探勘資訊，並且能依據空間位置給予對應的空間資訊，例如：***c***節點記錄著***c***鄉鎮包含O_1、O_2和O_3這 3 個農藥污染區、總面積為 430 平方公尺、平均面積為 143.3 平方公尺，總面積的標準差為 69.8，表示面積的各個資料的分散程度為 69.8，O_1是面積最大的農藥污染區、O_3是面積最小的農藥污染區，***c***鄉鎮的多重污染區域分佈在空間座標*x*軸為(59,40)與*y*軸為(73,48)的範圍裡，其中在空間座標*x*軸為(59,43)與*y*軸

表 1、Level 2 各個空間物件的彙總性統計資訊

Level 2	*a*	m_a	s_a	min_a	max_a	*MRCTOR*	*LAOR*	*ORMOO*
O_1	240	240	0	(O_1, 240, 0)	(O_1, 240, 0)	null	null	null
O_2	112	112	0	(O_2, 112, 0,)	(O_2, 112, 0,)	null	null	null
O_3	78	78	0	(O_3, 78, 0)	(O_3, 78, 0)	null	null	null
O_4	84	84	0	(O_4, 84, 0)	(O_4, 84, 0)	null	null	null
O_5	135	135	0	(O_5, 135, 0)	(O_5, 135, 0)	null	null	null
O_6	130	130	0	(O_6, 130, 0)	(O_6, 130, 0)	null	null	null
O_7	165	165	0	(O_7, 165, 0)	(O_7, 165, 0)	null	null	null
O_8	180	180	0	(O_8, 180, 0)	(O_8, 180, 0)	null	null	null
O_9	88	88	0	(O_9, 88, 0)	(O_9, 88, 0)	null	null	null
O_{10}	135	135	0	(O_{10}, 135, 0)	(O_{10}, 135, 0)	null	null	null
O_{11}	40	40	0	(O_{11}, 40, 0)	(O_{11}, 40, 0)	null	null	null
O_{12}	88	88	0	(O_{12}, 88, 0)	(O_{12}, 88, 0)	null	null	null

表 2、Level 1 各個空間物件的彙總性統計資訊

Level 1	*a*	m_a	s_a	min_a	max_a	*MRCTOR*	*LAOR*	*ORMOO*
c	430	143.3	69.8	(O_3, 78, 1)	(O_1, 240, 1)	((59,40),(73,48))	(((59,43),(64,48)) , 2)	(((59,43),(64,48)) , 2)
d	219	109.5	25.5	(O_4, 84, 1)	(O_5, 135, 1)	((69,22),(74,27))	(((69,22),(74,27)) , 2)	(((69,22),(74,27)) , 2)
e	475	158.3	21.2	(O_6, 130, 2)	(O_8, 180, 2)	((26,30),(40,38))	(((33,33),(40,38)) , 2)	(((30,33),(37,37)) , 3)
f	223	111.5	23.5	(O_9, 88, 1)	(O_{10}, 135, 1)	null	null	null
g	128	64.0	24.0	(O_{11}, 40, 0)	(O_{12}, 88, 0)	((32,1),(35,5))	(((32,1),(35,5)) , 2)	(((32,1),(35,5)) , 2)

表 3、Level 0 各個空間物件的彙總性統計資訊

Level 0	*a*	m_a	s_a	min_a	max_a	*MRCTOR*	*LAOR*	*ORMOO*
a	1124	140.5	50.1	(O_3, 78, 1)	(O_1, 240, 1)	((26,22),(74,48))	(((33,33),(40,38)), 2)	(((30,33),(37,37)) , 3)
b	351	87.8	33.4	(O_{11}, 40, 0)	(O_{10}, 135, 1)	((32,1),(35,5))	(((32,1),(35,5)), 2)	(((32,1),(35,5)), 2)

為(64,48)的區域為面積最大的多重污染區域，同時也是污染最嚴重的區域。依此類推，***d***節點則記錄著***d***鄉鎮包含O_4和O_5這 2 個農藥污染區、總面積為 219 平方公尺、平均面積為 109.5 平方公尺，各個面積的分散程度為 25.5，面積最大的農藥污染區是O_5、面積最小的農藥污染區是O_4，***d***鄉鎮的多重污染區域分佈在空間座標*x*軸為(69,22)與*y*軸為(74,27)的範圍裡，該範圍同時也是***d***鄉鎮面積最大的多重污染區域和污染最嚴重的區域。***e***節點則記錄著***e***鄉鎮包含O_6、O_7和O_8這 3 個農藥污染區、總面積為 475 平方公尺、平均面積為 158.3 平方公尺，各個面積的分散程度為 21.2，面積最大的農藥污染區是O_8、面積最小的農藥污染區是O_6，***e***鄉鎮的多重污染區域分佈在空間座標*x*軸為(26,30)與*y*軸為(40,38)的範圍裡，其中在空間座標*x*軸為(30,33)與*y*軸為(40,38)的區域為面積最大的多重污染區域，在空間座標*x*軸為(30,33)與*y*軸為(37,37)的區域為污染最嚴重的區域。已得知***c***鄉鎮、***d***鄉鎮和***e***鄉鎮農藥污染區的相關資訊後，便可進一步彙總***a***縣農藥污染區的相關資訊，***a***縣涵蓋***c***鄉鎮、***d***鄉鎮和***e***鄉鎮這 3 個的農藥污染區，面積相加即為總面積 1124 平方公尺、平均面積為 140.5 平方公尺，各個面積的分散程度為 50.1，O_1是面積最大的農藥污染區、O_3是面積最小的農藥污染區，多重污染區域分佈在空間座標*x*軸為(26,22)與*y*軸為(74,48)的範圍裡，其中在空間座標*x*軸為(30,33)與*y*軸為(40,38)的區域為面積最大的多重污染區域，在空間座標*x*軸為(30,33)與*y*軸為(37,37)的區域為污染最嚴重的區域。

我們採用 R-tree 的階層性架構，由最底層開始逐層往上累積資訊，讓每一個節點包含彙總性統計資訊和空間資訊，同時高階層節點具備低階層節點的階層式資訊，因此只需透過特定節點，即能快速取得對應的階層式空間資訊，例如：想得知 ***a*** 縣的農藥污染區的相關資料，則只需尋找對應的 ***a*** 節點就可快速得知彙總性統計資訊和空間資訊。若欲知 ***a*** 縣細部資料，便只需往下針對 ***c*** 鄉鎮、***d*** 鄉鎮或 ***e*** 鄉鎮的農藥污染區執行查詢即可。而不需要由最底層找出 ***a*** 縣所有的農藥污染區，重新反覆的執行運算判斷動作，大幅減少運算時間並快速得到彙總性統計資訊和空間資訊。

4.結論

STING方法用在Grid file的資料探勘項目為個數(*n*)、平均數(*m*)、標準差(*s*)、最小值(*min*)、最大值(*max*)。我們以這樣的資料探勘項目為概念，針對空間物件面積提出相關的資料探勘項目，分別為面積值(*a*)、平均面積值(m_a)、面積值的標準差(s_a)、面積最小的物件名稱(min_a)、面積最大的物件名稱(max_a)、涵蓋所有重疊區域的最小矩形(*MRCTOR*)、面積最大的重疊區域(*LAOR*)和同時被最多物件重疊的區域(*ORMOO*)。透過a、m_a和s_a明確得知特定範圍內空間物件的面積值、平均面積值和面積值的分散程度，由min_a和max_a清楚得知面積最小和面積最大的空間物件是哪一個，透過*MRCTOR*、*LAOR*和*ORMOO*探勘項目，可以得知特定範圍內包含全部重疊區域的最小矩形範圍、所有重疊區域中面積最大的區域和被最多物件同時重疊的區域。

由 R-tree 葉節點取得資料探勘項目的資料來源後，反覆執行運算判斷動作，從底層開始由下往上逐層記錄彙總性統計資訊和空間資訊，持續記錄到根節點為止，使得 R-tree 每一個節點皆包含許多豐富的資訊，同時高階層節點皆具備低階層節點的階層性統計資訊，節省搜尋和計算下層所有資料的時間。將資料探勘應用於空間資料索引，不但能增加資料查詢廣度，並且能依據空間位置快速得知對應的彙總性資訊和空間資訊。

5.參考文獻

[1] J. L. Bentley, "Multidimensional Binary Search Trees Used for Associative Searching", *Communications of the ACM*, Vol. 18, pp.509-517, 1975.

[2] R. Cooley, B. Mobasher, and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", *IEEE International Conference of Tools with Artificial Intelligence*, pp.558-567, 1997.

[3] V. Estivill-Castro and I. Lee, "AMOEBA: Hierarchical Clustering Based on Spatial Proximity Using Delaunay Diagram", *In Proc. 9th Int. Spatial Data Handling (SDH2000)*, pp.10-12, 2000.

[4] C. Faloutsos and K. Lin, "Fastmap: A Fast Algorithm for Indexing, Data Mining and Visualization of Traditional and Multimedia Datasets", *ACM SIGMODE Conf.*, pp. 163-174, 1995.

[5] V. Gaede and O. Gunther, "Multidimensional Access Methods", *ACM Computing Surveys*, pp.170-231, 1997.

[6] S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attribute", *In Proc. 1999 Int. Conf. Data Engineering (ICDE'99)*, pp.512-521, 1999.

[7] A. Guttman, "R-Tree: A Dynamic Index Structure for Spatial Searching", *Proc. ACM SIGMODE*, pp.47-57, 1984.

[8] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", *Morgan Kaufmann Publisher*, 2001.

[9] C. Kleissner, "Data Mining for the Enterprise", *Proceedings of the Thirty-First Hawaii International Conference on Vol. 7*, pp.295-304, 1998.

[10]R. Kosla and H. Blockeel, "Web Mining Research: A Survey", *ACM SIGKDD Explorations*, Vol. 2, pp.1-15, 2000.

[11]C. D. Lloyd, "Local Models for Spatial Analysis", *Boca Raton: Taylor & Francis*, pp. 28-31, 2007.

[12]R. McGill, J. Q. Tukey, and W. A. Larsen, "Variations of box plots", *The American Statistician*, pp.12-16, 1978.

[13]J. Nievergelt, H. Hinterberger, and K.C. Sevcik, "The Grid File: An Adaptable, Symmetric Multikey File Structure", *ACM Trans. Database System* Vol. 9, No. 1, pp.38-71, 1984.

[14]J. T. Robinson, "The K-D-B-tree: A Search Structure for Large Multidimensional Dynamic Indexes", *In Proc. of the ACM SIGMODE International Conf. on Management of Data*, pp.10-18, 1981.

[15]W. Wang, J. Yang, and R. Muntz, "STING: A Statistical Information Grid Approach to Spatial Data Mining", *In Proc. 23rd Intl. Conf. on VLBD*, pp.186-195, 1997.

[16]J. Xing, C. Yingkun, X. Kunqing, M. Xiujun, S. Yuxiang, and C. Cuo, "A Novel Method to Integrate Spatial Data Mining and Geographic Information System," *IEEE International Vol. 2*, pp. 795-798, 2005.

[17]D. Zhang and L. Zhou, "Discovering Golden Nuggets: Data Mining in Financial Application", *IEEE Transactions* on Vol. 34, pp.513-522, 2004.

[18]O. R. Zaiane, "Web Usage Mining for a Better Web-based Learning Environment," *In proc. of Conference on Advanced Technology for Education*, pp. 60-64, 2001.

[19]T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases", *In Proc. 1996 ACM-SIGMODE Int. Conf. Management of Data (SIGMODE'96)*, pp.103-114, 1996.