R-tree在資料彙總上的應用-
以土壤重金屬污染為例

陳靖國*
朝陽科技大學資訊管理系 助理教授
jkchen@cyut.edu.tw

許智偉
朝陽科技大學資訊管理系 研究生
s10154606@cyut.edu.tw

摘要

階層式空間資料索引最大特色是上層節點的進入項(entry)內的每一個資料項目值，是彙總下層所有節點進入項內的相關資料項目值而得。本論文研究在 R-tree 空間資料索引的節點進入項內，附加幾個預先計算好的空間物件面積的資料彙總項目。查詢某個大範圍區域內的某些彙總資料時，可以直接存取上層代表該範圍的進入項所包含的那些彙總資料，而不需要到下層的各個小範圍區域，再費時計算所有相關節點的個別資料，可以節省許多資料搜尋、計算、和加總時間。最後以土壤重金屬污染面積的相關資料查詢為例，證實本研究的實用性。

關鍵詞：R-tree、空間物件，資料彙總，資料存取

Abstract

The most feature of a hierarchical spatial data index is that each data in an entry of an upper-level R-tree node is obtained by the summation of all the related data in the entries of the lower-level R-tree nodes. This paper studies how to append some pre-calculated data items of spatial object area to the entry of R-tree node. When the sum data of a big region is needed, we can directly access the sum data in the specific entry of an upper-level R-tree node which corresponds to the big region. We do not need to compute and to sum the data in the entries of several lower-level R-tree nodes that correspond to several small regions. This mechanism saves much time of search, calculation, and collection of sum data in some entries. Finally, an example of earth heavy metal pollution is illustrated to prove the practicability of the study.

Keywords: R-tree, Spatial object, Data summation, Data access

1. 前言

尤其地理資訊系統和影像資料處理的蓬勃發展，對空間物件資料的需求更為殷切。彙總資料的處理，有些是屬於點資料，像是人口分佈調查統計；有些則是屬於空間資料，像是土地面積調查統計。各種不同範圍的彙總資料的取得，需要花許多時間計算大量數據。所以要從龐大的資料量中，快速取得有用的彙總資料，是一件相當麻煩費時的事。若能預先將可能被用到的彙總資料計算好並儲存起來，當有需求

時馬上供應，即可加快查詢速度。當然，代價是事先準備的工作要花一點時間。

樹狀索引依據資料型態可以分成兩種類型。第一種是針對點資料建立索引，例如K-D tree [2]、K-D-B tree [13]、Grid file [10]等，這些方法是以一個點的資料代表一個物件，將不同點的資料所隸屬的空間做區分來建立索引。第二種是針對空間資料建立索引，例如R-tree [7]、R^+-tree [14]、R*-tree [1]、X-tree [3]等，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。索引可以加快資料的檢索速度，目前已知的各種應用領域所用到的索引，多數為點資料索引，只有少數為空間資料索引[4, 5, 6, 8, 9, 11, 12, 15]。R-tree為廣泛應用的空間資料索引，因此適合探討在樹狀索引結構上附加預先彙總資料的應用。

環境保護意識抬頭，使得土壤受重金屬污染的問題逐漸受到重視。記錄並查詢受污染的土地相關資料實有必要，所以本研究結合空間資料索引與預先彙總的概念，以快速取得空間物件彙總資料。方法是在R-tree的節點進入項(entry)內，附加幾個預先計算好的有關空間物件面積資料的彙總項目。當需要某個大範圍土地污染資料時，可以在代表某個高階層範圍的土地節點進入項內取得相關資料，而不需要先搜索許多低階層範圍的土地污染資料，再加總成彙總資料，節省找尋、計算、和加總的時間成本。

2. 文獻探討

2.1 R-tree

R-tree [7]是階層式多維度動態資料索引，應用於多維度空間物件之存取。R-tree由葉節點(leaf node)以及非-葉節點(non-leaf node)所組成，每一個節點包含若干個進入項(entry)，一個進入項代表某個物件或某個子節點相關資訊，包括一個最小涵蓋矩形(MBR)與一個指標(pointer)。葉節點的進入項內容為(*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；非-葉節點的進入項內容為(*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的索引結構中，根節點會涵蓋所有子節點所涵蓋的範圍，而葉節點則包含實際物件的資訊。若*M*為每一個節點所能擁有的最大進入項數量，而*m*為每一個節點至少擁有的最小進入項數量，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)每一個非-根節點的進入項數量必須介於*m*和*M*之間；(3)所有葉節點都出現在同一層。圖1為11個物件散佈情形，黑色粗體實線的矩形：A~K代表物件。圖2為物件對應的R-tree索引結構，其中*m*為2，*M*為3。

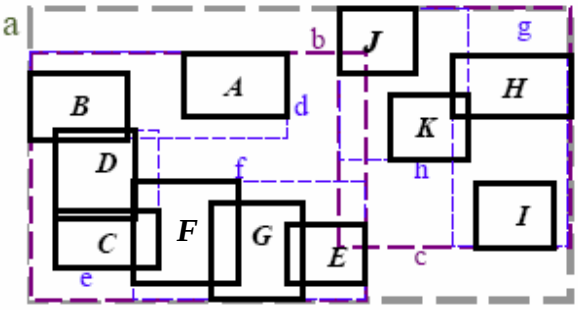

圖1、物件散佈情形

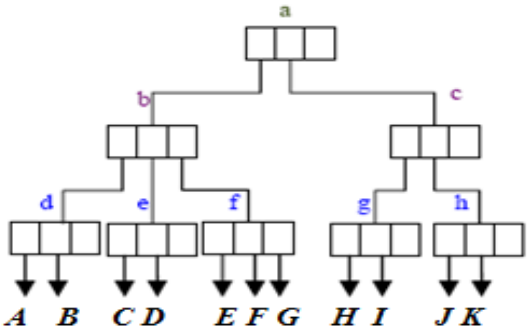

圖2、對應的R-tree索引

2.2 STING

STatistical **IN**formation **G**rid (STING)為階層式彙總資料方法[15]，應用於點資料上。STING是以階層式架構方式由上而下規律地將資料空間切割成許多單元(cell)，切割方式如下。首先將最高階層(1st level)的資料空間切割成k個單元，其中$k=2^i*2^i$，$1\leqq i\leqq n$，因此下一階層(2nd level)的資料空間將被分成k個單元，接著將該階層的k個單元，分別再切割成k個單元，則下一階層(3rd level)的資料空間將被分成k^2個單元，依此類推，架構如圖3所示[15]。STING在每一個單元內記錄彙總資料，包含物件個數(n)、平均值(m)、標準差(s)、最小值(min)、最大值(max)等。由於每一個單元所記錄的彙總資料，皆是由下一階層的若干個子單元所記錄的彙總資料，加以計算並儲存而得的，因此高階層單元會彙總低階層單元的總和資料。

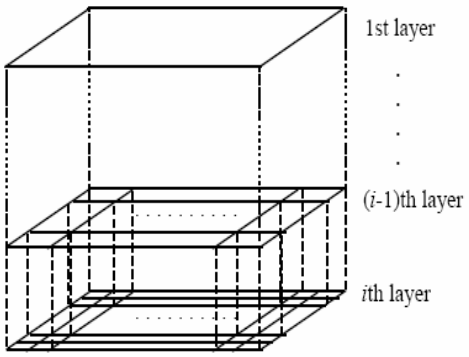

圖3、STING架構

3. 空間物件彙總資料

空間物件和點物件最明顯的差別在於空間物件有面積或體積的存在，但是點物件沒有。R-tree 節點的每一個進入項，可以描述空間物件的某個特定範圍，如果事先彙總好資料，將小範圍的基本資料，逐層彙總成大範圍的資料，再將不同範圍的彙總資料儲存在各個代表該範圍的進入項上，可加快資料取得速度。

本研究提出幾項與空間物件面積相關的彙總資料項目：物件面積總和(*object area sum*,簡稱 a)、面積總和平均值(*mean of object area sum*, 簡稱 m_a)、面積總和標準差(*standard deviation of object area sum*, 簡稱 s_a)、面積總和最少值(*the least value of object area sum*, 簡稱 lst_a)、面積總和最多值(*the most value of object area sum*, 簡稱 mst_a)。其中面積總和最少值(lst_a)/最多值(mst_a)是針對某個特定範圍的下一階層所有子範圍裡，找出某個子範圍內，所有物件面積總和最少/最多的面積值。以下依序說明這些彙總項目的計算公式和變數說明。

(1) 空間物件的面積總和 (a)

計算公式為 $a=\sum_i a_i$ ，其中 a_i 為含 a 的節點 p 的下一階層某個子節點 i 所包含的物件總面積。亦即 a 是由節點 p 的下一階層每一個子節點 i，所包含的物件面積 a_i 相加而得。每一個 a_i 又是由節點 i 的下一階層每一個子節點 j，所包含的物件面積 a_j 相加而得，依此遞迴而得到各個相關面積值數據。

(2) 空間物件的面積總和平均值 (m_a)

計算公式為 $m_a=\frac{\sum_i m_{a_i} n_i}{EEN}$ ，其中 m_{ai} 為含 m_a 的節點 p 的下一階層某個子節點 i，所包含的物件面積平均值；n_i 為子節點

i 所包含的物件個數值；EEN 為節點 p 所包含的有效進入項個數值。亦即 m_a 是由節點 p 的下一階層每一個子節點 i，所包含的物件個數值 n_i 與物件面積平均值 m_{ai} 相乘的總和，再除以節點 p 所包含的有效進入項個數值 EEN 而得。每一個 m_{ai} 又是由節點 i 的下一階層每一個子節點 j，所包含的物件個數值 n_j 與物件面積平均值 m_{aj} 相乘的總和，再除以節點 i 所包含的有效進入項個數值 EEN_i 而得，依此遞迴而得到各個相關面積平均值數據。

(3) 空間物件的面積總和標準差 (s_a)

計算公式為 $s_a = \sqrt{\frac{\sum_i (s_{ai}^2 + m_{ai}^2) n_i}{EEN} - m_a^{\,2}}$，其中 s_{ai} 為含 s_a 的節點 p 的下一階層某個子節點 i，所包含的物件面積值標準差；m_{ai} 為子節點 i 所包含的物件面積平均值；n_i 為子節點 i 所包含的物件個數值；m_a 為節點 p 所包含的物件面積平均值；EEN 為節點 p 所包含的有效進入項個數值。亦即 s_a 是由節點 p 的下一階層每一個子節點 i，所包含的物件面積值標準差 s_{ai} 平方與物件面積平均值 m_{ai} 平方相加，再乘以物件個數 n_i 的總和，再除以節點 p 所包含的有效進入項個數值 EEN，減掉節點 p 所包含的物件面積平均值 m_a 平方後，再開根號而得。每一個 s_{ai} 又是由節點 i 的下一階層每一個子節點 j，所包含的物件面積值標準差 s_{aj} 平方與物件面積平均值 m_{aj} 平方相加，再乘以物件個數值 n_j 的總和，再除以節點 i 所包含的有效進入項個數值 EEN_i 之後，減掉節點 i 所包含的物件面積平均值 m_{ai} 平方後結果再開根號而得，依此遞迴而得到各個相關的面積值標準差數據。

(4) 空間物件的面積總和最少值 (lst_a)

計算公式為 $lst_a = \min_i(lst_{ai})$，其中 lst_{ai} 為含 lst_a 的節點 p 的下一階層某個子節點 i，所包含物件面積的最少值。亦即 lst_a 是比較節點 p 的下一階層每一個子節點 i 所包含的 lst_{ai}，找出最小的 lst_{ai}。每一個 lst_{ai} 又是從節點 i 下一階層每一個子節點 j 中找出最小的 lst_{aj}，依此遞迴而得到各個相關物件總和面積最少值的數據。

(5) 空間物件的面積總和最多值 (mst_a)

計算公式為 $mst_a = \max_i(mst_{ai})$，其中 mst_{ai} 為含 mst_a 的節點 p 的下一階層某個子節點 i 所包含物件面積的最多值。亦即 mst_a 是比較節點 p 的下一階層每一個子節點 i 所包含的 mst_{ai}，找出最多的 mst_{ai}。每一個 mst_{ai} 又是從節點 i 的下一階層每一個子節點 j 中找出最大的 mst_{aj}，依此遞迴而得到各個相關的物件總和面積最多值的數據。

4. 實例應用

本研究修改 R-tree 節點原來的進入項資料結構來包含上述所提到的彙總資料項目，修改後進入項的結構為(MBR, $pointer$, a, m_a, s_a, lst_a, mst_a)。不同節點的每一個進入項內的彙總資料設定如下。首先設定最底層 level 葉節點各個進入項的彙總資料，再往上一層的非-葉節點內，計算每一個進入項的各項彙總資料，如此逐層往上處理，直到將根節點的各個進入項處理完為止。由於彙總資料是由 R-tree 最底層 level 開始往上加總計算而得到的，因此高層 level 節點的進入項的彙總資料，會彙集低層 level 節點的進入項的彙總資料。

假設台灣中部某兩縣，代號 ***a***、***b***，的土壤重金屬污染區調查結果，如圖 4 所示。圖 4 呈現土壤重金屬污染區的分佈情形，每一個 O_1~O_{12} 代表不同的重金屬污染區，而圖 5 表示污染區對應的 R-tree 索引結構。透過預先彙總運算處理，可得不同範圍污染區的彙總資料，如表 1、2、和 3 所示。每一個進入項除了可以描述某個特定範圍污染區的地理位置之外，也記錄著代表該範圍污染區許多相關的彙總資料。以 ***c*** 鄉鎮的污染區資料為例，經由存取

R-tree 中表示 ***c*** 鄉鎮的節點進入項 ***c***，可以取得 ***c*** 鄉鎮污染區的面積相關彙總資料。例如 ***c*** 鄉鎮的污染區總面積為 430 平方公尺、污染區平均面積為 143.3 平方公尺、污染區面積標準差為 69.7，***c*** 鄉鎮的污染區面積總和最多為 240 平方公尺、最少為 78 平方公尺。依此類推，可得知 ***d*** 和 ***e*** 鄉鎮污染區相關的彙總資料。在已知 ***c***、***d***、***e*** 鄉鎮污染區相關的彙總資料後，可進一步彙總得到 ***a*** 縣污染區相關的彙總資料，透過存取表示 ***a*** 縣的根節點進入項 ***a***，可以取得 ***a*** 縣污染區面積相關的彙總資料。例如 ***a*** 縣的污染區總面積為 1124 平方公尺、污染區平均面積為 374.7 平方公尺、污染區面積標準差為 111.6，***a*** 縣的污染區面積總和最多為 475 平方公尺、污染區面積總和最少為 219 平方公尺。

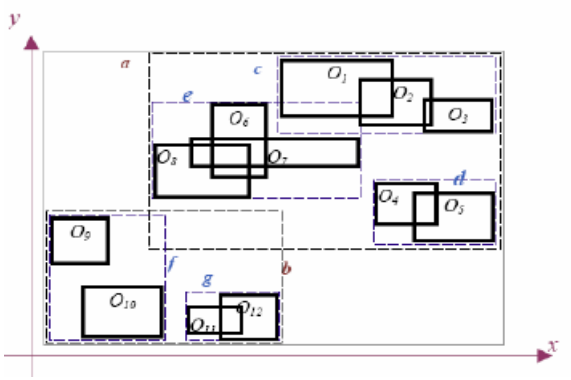

圖 4、污染區的分佈情形

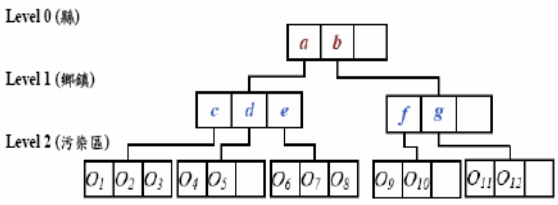

圖 5、對應的 R-tree 索引

表 1、各個土壤重金屬污染區的彙總資料

Level 2	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}	O_{12}
a	240	112	78	84	135	130	165	180	88	135	40	88
m_a	240	112	78	84	135	130	165	180	88	135	40	88
s_a	0	0	0	0	0	0	0	0	0	0	0	0
lst_a	240	112	78	84	135	130	165	180	88	135	40	88
mst_a	240	112	78	84	135	130	165	180	88	135	40	88

表 2、各鄉鎮土壤重金屬污染區的彙總資料

Level 1	***c***	***d***	***e***	***f***	***g***
a	430	219	475	223	128
m_a	143.3	73.0	172.5	111.5	64.0
s_a	69.7	49.0	7.5	23.5	24.0
lst_a	78	84	130	88	40
mst_a	240	135	180	135	88

表 3、各縣土壤重金屬污染區的彙總資料

Level 0	***a***	***b***
a	1124	351
m_a	374.7	175.5
s_a	111.6	47.5
lst_a	219	128
mst_a	475	223

5. 結論

為了快速取得空間物件面積的相關彙總資料，本文提出資料索引結合預先彙總的概念，在 R-tree 節點的進入項內，嵌入幾個實用的彙總資料。預先計算並儲存好這些彙總資料，當需要特定大範圍區域的彙總資料時，直接讀取代表該範圍的 R-tree 上層節點進入項所包含的彙總資料，就可以快速得到答案，例如 ***a*** 縣特定範圍的污染區面積總和、面積總和平均值、面積總和最多值等，而不需要到下層的小範圍區域，實際加總所有節點進入項內的個別資料，大幅省去許多下層節點進入項的資訊搜尋、計算、和加總時間。

參考文獻

[1] N. Beckmann, H. P. Kriegel, R. Schneider, and B. Seeger, “The R*-tree: An Efficient and Robust Access Method for Points and Rectangles,” In Proc. of ACM SIGMOD Int. Conf. Management of Data, pp. 322-331, 1990.

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

[3] S. Berchtold, D. A. Keim, and H. P. Kriegel, “The X-tree: An Index Structure for High-Dimensional Data,” In Proc. 22th Int. Conf. on VLDB, pp. 28-39, 1996.

[4] Y. Chang, C. Liao, and H. Chen, “NA-Trees: A Dynamic Index for Spatial Data,” Journal of Information Science and Engineering, Vol. 19, No. 1, pp. 103-139, 2003.

[5] P. Ferragina and R. Grossi, “The String B-Tree: A New Data Structure for String Search in External Memory and Its Applications,” Journal of the ACM, Vol. 46, pp. 236-280, 1998.

[6] M. Greenspan and M. Yurick, “Approximate K-D Tree Search for Efficient ICP,” In Proc. of the 4th IEEE Int. Conf., pp. 442-448, 2003.

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

[8] C. Jensen, D. Tiesyte, and N. Tradisauskas, “Robust B+-Tree Based Indexing of Moving Objects,” In Proc. 7th Int. Conf., pp. 12-21, 2006.

[9] A. Mondal, Yilifu, and M. Kitsuregawa, “P2PR–tree: An R-tree Based Spatial Index for Peer-to-Peer Environments,” In Proc. of Int. Workshop on Peer-to-Peer Computing and Databases, 2004.

[10] 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.

[11] P. O'Neil and D. Quass, "Improved Query Performance with Variant Indexes," In Proc. of the ACM SIGMOD Conf. on the Management of Data, pp. 38-49, 1997.

[12] C. Procopiuc, P. Agarwal, and S. Peled, "STAR-Tree: An Efficient Self-Adjusting Index for Moving Objects," In Proc. of the Workshop on Alg. Eng. and Experimentation, pp. 178-193, 2002.

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

[14] T. Sellis, N. Roussopoulos, and C. Faloutsos, "The R^+-Tree: A Dynamic Index for Multi-Dimensional Objects," In Proc. of the13th VLDB Conf., pp. 507-518, 1987.

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