行政院國家科學委員會補助專題研究計畫 □ 成 果 報 告 ■期中進度報告

感測資料之收集、處理及探勘技術之研究及其應用 (新編多年期第二年)

The Research and Application of the Techniques for Sensor Data Collection, Processing, and Mining

計畫類別：■ 個別型計畫　　□ 整合型計畫

計畫編號：NSC – 97 – 2221 – E – 004 – 006 – MY3

執行期間：2008 年 08 月 01 日至 2011 年 07 月 31 日

計畫主持人：陳良弼

成果報告類型(依經費核定清單規定繳交)：■精簡報告　□完整報告

本成果報告包括以下應繳交之附件：

□赴國外出差或研習心得報告一份

□赴大陸地區出差或研習心得報告一份

□出席國際學術會議心得報告及發表之論文各一份

□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢

□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立政治大學資訊科學系

中　華　民　國　99　年　5　月　31　日

中文摘要

隨著科技發展，內嵌無線通訊、精密感測、及計算等功能之感測器裝置的使用已日漸普及。在可預見的未來，智慧型感測器系統將大規模融入人們的生活環境，提供大量、即時、且各式各樣的感測器資料。本研究計畫以三年為期，開發一以智慧型商店經營為應用之感測器應用雛型系統，並發展相關技術。在本年度計畫執行過程中，我們已完成預定完成之研究項目，分別為感測器資料聚合處理技術、事件串流之段落比對技術及感測資料結合查詢處理技術，並已發表於國際一流相關期刊及會議。本期中報告茲就本年度所完成的研究成果進行精簡報告。

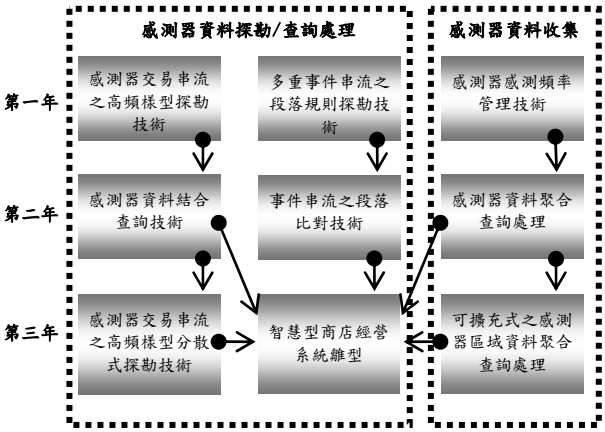
Abstract

Recently, due to the advance of technologies in wireless sensor networks, the applications on sensor networks have received considerable attention. In the near future, sensor network systems will gradually and seamlessly weave into human's living space and provide mass and streaming sensor data. In this project, we make the first attempt to build an intelligent store management system, providing decision supports on businesses and personalized recommendation based on sensor data, and develop the core techniques needed in this application. In this progress report, three research results we achieved in this year are reported, including 1) robust sensor data aggregation technique, 2) episode matching over event streams, and 3) efficient join processing over sensor data streams.

一、前言

隨著無線通訊、超大型積體電路、及微機電等相關技術的蓬勃發展，內嵌通訊、感測、計算等功能之感測器裝置的使用也早已日漸普及。智慧型感測器的相關研究在國內外已經進行了好幾年，特別是在感測器系統硬體設計方面，早已長足進步且日趨成熟。依現有技術要佈建感測器系統並非難事，然而相關應用卻相對匱乏的主因來自於，感測器系統相關資料管理與資料應用技術上的不足。在佈建感測器系統後，如何妥善應用感測器系統資料，正是目前國際領先研究群爭相投入的課題。有鑑於此，本研究計畫將從智慧型感測器資料應用的角度切入，規劃感測器系統可能之應用情境及商業模式，並研發可能應用下之相關資料管理技術。

我們以三年為期，擬研發一個以智慧型商店經營為應用之感測器應用雛型系統，該系統將考量以感測器系統為資料收集平台，利用感測器資料為基礎，進行智慧型商店經營決策協助與顧客購物推薦服務，該應用系統之關鍵技術涵蓋下述兩大範疇：『感測器資料收集技術』及『感測器資料探勘/查詢處理技術』，此二關鍵技術之研究項目與各項目間之關連性及執行進度規劃如圖一所示。對應於此二類關鍵技術，在本年度計畫執行過程中，我們已完成第二年度預定完成之研究項目，分別為感測器資料聚合處理技術、事件串流之段落比對技術及感測資料結合查詢處理技術。各項研究成果將於下一節中進行探討，以下為與本年度三項研究項目相關之國內外研究。



圖一：研究項目及其關連性

國內外相關研究

1、 感測器資料聚合處理技術

在高可靠度無線感測資料聚合計算技術相關研究中，[MF02]首先提出兩個策略，分別為子節點快取(child cache)與分數式傳遞(fractional parent)，來提升聚合計算的準確度。在子節點快取策略中，每個節點皆儲存其子節點先前傳遞過的部份聚合值，當某個感測器節點之子節點資料遺失時，該感測器節點便利用先前儲存之對應部份聚合值進行資料聚合運算。在分數式傳遞，作者提出多路徑式資料傳遞(multi-path routing)的概念，廣播部分聚合值至其上一層的節點，並採用

均分式切割；舉例來說，若某一節點有 5 個母節點則該節點將其部份聚合值除於 5，來解決多路徑式資料聚合所衍生的重複記數問題(double counting problem)。在[NG04]中作者提出利用複本無影響性之速寫結構(duplicate-insensitive sketches)的概念，來解決多路徑式資料聚合計算中重複記數問題。透過利用複本無影響性之速寫結構來表示感測器所將廣播之資料，相同的一筆資料最終將只會被計算一次。基於複本無影響性之速寫結構概念，[CL04]提出使用 FM-Sketches [FM85]，來解決多路徑式資料聚合所產生的重複記數問題。

在[MN05]中，作者提出同時使用多路徑式資料聚合與樹狀式資料聚合來進行高可靠度無線感測資料聚合；在網路通訊失敗率較高的地方使用以 FM-Sketches 為基礎之多路徑式資料聚合，而在網路通訊失敗率較低的地方使用一般的樹狀式資料聚合。然而，使用此方法需主動管理並動態切換資料聚合模式，頻繁地切換各區域所維持的資料聚合模式，將造成額外的能源消耗與能源利用率下降。在[SB04]中，作者則探討較複雜的資料聚合函式計算，如中位數(median)與統計條狀圖(histogram)，此文獻中假設並無網路通訊失敗的情形，並將焦點置於感測器觀測值的統計資料儲存結構之設計。在[CP06]中，作者提出利用重複的隨機資料交換程序(data exchanging process)的方法來計算無線感測器網路中資料聚合值。在一次隨機資料交換程序中，網路中的感測器節點隨機的成為主導節點。成為主導節點的感測器節點利用廣播邀請其鄰居來形成感測器群組，群組中的節點將其所儲存的部分資料聚合值傳遞給主導節點。主導節點接收其群組所傳遞之部分資料聚合值後，計算新的部分資料聚合值，並廣播予群組中之節點。利用如此的隨機資料交換程序，各感測器節點所儲存的部分資料聚合值將逐漸收斂至正確的資料聚合值。然而使用這樣方法的缺點為需要多次的隨機資料交換程序，造成感測器能源的消耗與查詢結果回報的延遲。

2、 事件串流之段落比對技術

事件串流之段落比對的相關研究包括主動式資料庫管理系統(active database management systems)，發行訂閱系統(publish/subscribe systems)，以及在資料串流上的複雜事件處理系統(complex event processing systems)。在主動式資料庫管理系統中，當滿足 ECA 規則(Event-Condition-Action-Rules)的述語(predicate)描述出現時，即觸發即時性回覆。一般而言，ECA 規則的述語為合成事件(composite event) [DB88]，而合成事件又由較簡易的合成事件或是原始事件(primitive events)所組合而成。在過去的數十年間，如 HiPAC [DB88]、Ode [GJ92] [GJ92a]、Snoop [CK94]、SAMOS [GD92] [GD94]和 NAOS [CC96]等計畫已發展出多個主動式資料庫管理系統。

在傳統的發行訂閱系統中[AS99] [DG06] [FJ01]，傳送者會發行一事件、或訊息和新聞給接收者。此事件是由一對屬性值(attribute-value pair)所組合而成。而接收者則可經由給定一對屬性值運算子(attribute-value-operator pair)來訂閱其感興趣之相關內容。[AS99]利用樹狀結構來建立所有使用者的訂閱資訊之索引，在此結構中，一條從根結點(root)至葉結點(leaf)的路徑可代表一使用者訂閱。由不同使用者所發出之相同訂閱，皆會與此路徑相連結。假如一發行事件能夠由此結構的根結點到達某一葉結點，則代表此事件滿足對應於此路徑的訂閱。[FJ01]是在有限的記憶體資源下，建立訂閱群組(subscription cluster)，並利用雜湊函數(hashing function)來降低存取群組的成本。在[DG06]中，訂閱內容可由多個事件來組成，而作者以非決定的有限狀態自動機(nondeterministic finite state automata)為基礎，發展一有效率訂閱索引結構。在此結構之下，可同時處理多個使用者訂閱資料。隨著 XML 的風行，進階的發行定閱機制[AF00] [CF02] [DF03] [PC03]多半應用於擷取使用者感興趣的 XML 文章。在這些文獻中，它們利用 X 路徑表示式(XPaths expression) [CD99]來描述使用者訂閱。當 XML 文章以串流方式進入系統時，它們便根據使用者給定的 X 路徑表示式，對 XML 文章進行比對。在此領域中，大部份的研究[AF00] [DF03] [PC03]都

以有限狀態自動機(finite state automata)為基礎，發展各自的比對演算法，用以解決複雜的 X 路徑表示式。此外，[CF02]和[DF03]則是從使用者的 X 路徑表示式中擷取共同的子表示式(common sub-expression)，用以建立索引來同時處理多個查詢。

在資料串流上的複雜事件處理系統中，[WD06]針對射頻辨識(RFID)的資料串流提出了一事件序列查詢模組，用以處理帶有屬性值的事件。在射頻辨識的環境下，也有其他的資料串流管理系統如[FJ05] [WL06]。在這些研究中，它們各自提出自己的查詢模組和系統架構。至於在以關聯式運算子為主的資料串流管理系統中[AN04] [BM04] [VN02]，多半著眼於有限資源下，如何達到最佳化查詢計劃(query plan)。

3、 感測資料結合查詢處理技術

在資料串流環境中，結合查詢處理為近年來熱門的研究課題。然而，在資料串流環境中處理結合查詢的技術皆使用的集中式處理架構，並不適合直接引用至感測器資料串流的環境。感測器資料的結合查詢處理的相關研究近年來也漸受注目。在[YG03]中，作者將感測器網路視作一個大型分散式資料庫，提出使用 SQL 類似的查詢，來管理感測器資料的蒐集。並使用內網路查詢處理機制(in-network processing)來處理選拔查詢與資料聚合查詢。對於結合查詢，[YG03]則採取將所有資料收集到感測器網路主機，並在該主機上進行資料間的結合操作。[HA04]則進一步探討處理感測器資料結合查詢時，感測器資料由於網路傳遞間的延遲所衍生的查詢處理問題。然而，這兩個方法亦使用集中式處理架構，故消耗大量的感測器能源，造成感測器應用上使用效率的低落。

[CG05]探討兩群感測器節點的讀數要進行內網路結合操作時，尋找感測器網路中最佳結合操作之執行地點的問題。文中主要焦點著重於理論上探討，並做了許多強烈的假設，如感測器網路佈置區域中任何地點皆擁有感測器節點等，使得文中所提出的技術在實際應用上並不可行。[YL07]則在感測器網路中利用自我結合(self join)的查詢，進行感測器觀測值趨勢與變化之監測。作者提出查詢重寫技術(query rewriting)，將使用者所下達的自我結合查詢，改寫成兩個選拔查詢以利查詢處理之進行。作者並針對多個查詢連續下達時，設計一個查詢執行排程的機制來避免不必要或重複的查詢處理。[SM05]則在一個特殊階層式感測器網路架構下(越上層的感測器節點擁有越低的查詢處理花費與越低的網路傳輸成本)執行一個選拔查詢的多個選拔語句(selection predicate)。資料由階層式感測器網路最底層的節點產生所產生，經過層層的感測器節點的轉傳與處理，最終完成與傳遞查詢結果於階層式網路之最上層節點。當資料於階層式網路最下層產生時，有兩種選擇來進行查詢處理。第一個選擇為，先將資料在階層式網路較下層的節點進行部分的處理，以減少資料的傳遞成本。第二個選擇為，將資料傳遞到階層式網路較上層節點處理，以減少資料處理的成本。很明顯地，兩個選擇彼此間有衝突，因此如何尋找一個最佳的查詢處理工作節點配置，成為[SM05]查詢執行的效率關鍵所在。

二、研究方法、進行步驟及執行進度報告

我們擬發展之感測器應用雛形系統中所需之兩項關鍵技術：『感測器資料收集技術』及『感測器資料探勘/查詢處理技術』，第二年成果報告如下所示：

感測器資料收集之成果報告

1、 感測器資料聚合處理技術

研究目的

在感測器系統應用中，除單一感測器的資料收集外，針對一群感測器節點資料之聚合查詢，如平均、總和、記數、最大值等，也廣為使用，如應用在平均雨量回報及活動中之感測器節點數回報等。現階段常見的資料聚合方式乃是採用樹狀式資料聚合計算。首先，建構一以主機為根節點(root)的擴張樹(spanning tree)，用以連接各感測器節點。而資料聚合則由葉節點(leaf)開始層層進行：各節點接收其子節點所傳送之部份聚合值，結合本身所觀測之資料，計算出新的部份聚合值，再往母節點傳送，如此一來最終便可於根節點計算出完整聚合值。此聚合方式的缺點，主要為通訊容錯能力不佳；造成許多節

點的部份聚合值可能因通訊失敗而遺失。若該通訊失敗之節點位於根節點附近，將造成大量資料遺失，使得最終計算所得之聚合值遠遠偏離實際聚合值。

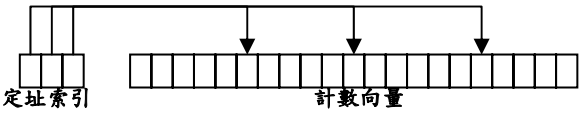
欲提升感測器資料聚合計算上的通訊容錯能力，可選擇採用高可靠率的通訊協定，但卻因此增加感測器能源上的消耗。因此，在使用簡易通訊協定前提下，以多路徑式資料傳遞為基礎來提升其通訊容錯力，廣為相關研究所使用。在多路徑式資料聚合計算中，感測器系統以有向非循環圖，來連結各感測器節點，而感測器節點則經由廣播將其部分聚合值傳至其上層節點。由於一筆部分聚合值有多筆複本在網路中傳遞，唯有在所有複本皆遺失的狀況下，才會造成部分聚合值的遺失，因此提升了聚合計算之通訊容錯能力。但也因此造成相同資料可能被多次接收，衍生出重複計數的問題。重覆計數在某些查詢下，並不會影響查詢結果，例如求取擁有最大溫度值的感測器編號；但對於某些聚合查詢，例如回報活動中之感測器節點數，將造成錯誤的查詢結果。有鑒於此，我們以多路徑式資料遞送方式為基礎，設計一具高可靠度、高準確率與高能源效率之感測器資料聚合計算方法。

研究方法

我們延伸線性計算速寫技術(linear-counting sketches)[WV90]來避免多路徑式資料聚合中重複計數問題。此技術主要包含一隨機雜湊資料結構。給定一多重集合，其使用方法如下：首先，配置一長度為 m，初始值為 0 的位元陣列(bit array)。同時，使用一均勻雜湊函式(hash function)將多重集合中的所有元素對應至位元陣列，並將所對應到的位址值設定為 1。最後計算位元陣列中，所有位址值非 0 的位址數目(V_n)，透過 $\hat{n} = -m \times \ln(V_n)$ 公式即可估算多重集合中相異元素數量 $\hat{n}$。透過此技術，在多重集合中之相同元素會被對應至相同位址，避免重複計算。基於線性計算速寫技術我們可設計一資料結構，根據使用者的允許誤差與誤差值變異數，來設定資料結構長度；可以想見，此資料結構的長度應與真正的資料聚合值有關。然而，最終的資料聚合值在決定資料結構長度時是未知的，因此過去常使用真正資料聚合值的上限來決定資料結構長度。此法之副作用有二：1、資料結構長度過長及 2、離網路主機較遠之節點所傳遞之位元陣列大部分位址值為 0，造成大量能源消耗。因此我們提出一個新式演算法，動態調整資料結構長度，以避免上述二缺點。我們的方法進行步驟如下:

步驟一: 使用動態計數速寫結構表示感測器資料

首先，令所有參與資料聚合的節點 u_i，根據其感測資料值 v_i，配置一長度為 m_i 之動態計數速寫資料結構。一個節點的動態計數速寫資料結構 DC(BI, CV)包含一個定址索引元件 BI (border Index)與一個計數向量 CV (counting Vector)。圖二為我們所設計之動態計數速寫資料結構之概念圖。定址索引提供不同節點之計數向量長度索引位置，而計數向量則為一初始值為 0 之位元陣列。接著，所有節點 u_i 將其計數向量隨機均勻地設定 v_i 個元素為 1，並將定址索引指向 m_i 的位置。



圖二：動態計數速寫資料結構

步驟二: 內網路動態計數速寫資料結構聚合計算

接著，由網路最底層的節點開始，將其資料結構廣播到其上層節點。收到動態計數速寫資料結構之節點，將其本身之資料結構與接收到之資料結構進行內網路資料聚合(in-network aggregation)。此聚合動作定義如下：給定兩個動態計數速寫資料結構 $DC_1(BI_1, CV_1)$與 $DC_2(BI_2, CV_2)$，若$|CV_2| \geq |CV_1|$，則 DC_1 與 DC_2 之總和 $DC_3(BI_3, CV_3)$滿足 1、$BI_3 = BI_1 \cup BI_2$ 與 2、$CV_3[i] = CV_1[i] \vee CV_2[i], \forall\, i = 0, \ldots, |CV_1| - 1$，且 $CV_3[i] = CV_2[i], \forall\, i = |CV_1|, \ldots, |CV_2| - 1$。接著該節點將聚合過後的資料結構以此類推地往其上層節點傳送，直至所有動態計數速寫資料結構集合於網路主機節點。

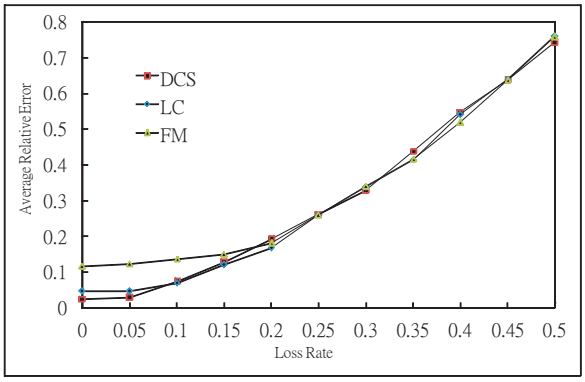
步驟三: 近似資料聚合結果計算

待所有動態計數速寫資料結構集合於網路主機節點後，主機會將所有動態計數速寫資料結構聚合，產生一最終動態計數速寫資

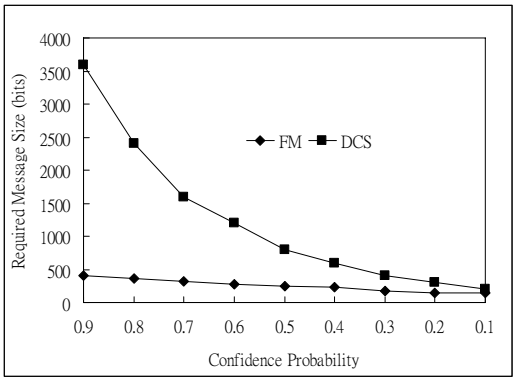
料結構，並根據此結構之定址索引與計數向量所提供之資訊，估算近似資料聚合結果並回傳予使用者。

實驗結果

在本研究中我們實作所提出之方法(DCS)並與 LC[FC08]及 FM[CL04]進行比較。圖三為近似資料聚合結果準確率的比較圖表，在使用相同空間時，動態計數速寫資料結構提供較高準確率，同時也兼顧能源效率。同時，由圖四可得知， DCS 使用較少的傳輸量來達成使用者所給定的誤差需求。



圖三：準確率實驗比較



圖四：所需空間實驗比較

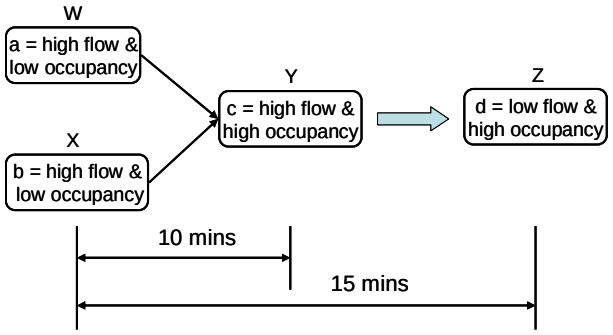
感測器資料探勘/查詢處理之成果報告

1、事件串流之段落比對

研究目的

在多數應用中，資料多半被表示成事件(events)，例如：股價的波動、電信網路中的警報訊息、使用者瀏覽網頁的紀錄、以及在交通管理系統中道路的狀況等等。通常特殊事件的發生可能意味著未來某些特定事件即將發生，因此我們可將此一特性應用在預測(prediction)功能上。明確而言，在段落規則(episode rule)已知的前提下，我們可發展段落比對(episode matching)技術來預測事件發生。

一段落規則(episode rule)可用$\alpha \rightarrow \beta$形式來表示之，α為此段落規則的述語(predicate)，β為其後項(consequent)，且此一述語及後項皆屬段落(episode)。段落是由一群事件組合而成，可以一有向非循環圖(directed acyclic graph)表示之。此外，我們設定一述語框架(predicate window)用以規範述語中的事件必須在一限定之時間區間內發生；同時亦設定另一規則框架(rule window)，用以規範整個段落規則中的所有事件必須在另一限定的時間區間內發生。圖五即為一段落規則之實例，此規則表示若在 10 個時間單位內，述語內的所有事件皆按其特定的發生次序發生，那麼，在後項內的事件，將有很高的機率會發生於 15 個時間單位內。由於前述多數的應用所涉及之事件多半以串流且不間斷的型態進入系統，因此在此研究中，我們希望能持續監控串流事件，並利用已知的段落規則在事件串流中進行快速比對，以達到線上預測(online-prediction)功能。



圖五：一段落規則之實例

研究方法

為達線上預測的功能，我們將已知段落規則的述語，跟串流資料進行持續性比對，當發現在某個時間區段內，此述語中的全部事件已按照規範於述語內的順序發生，則此時便會提出警報(alarm)，告知使用者此段落規則的後項極可能會在某一時間範圍內出現。為了有效率地回報所有可能會發生之後項事件的預測區間，因此我們記錄了述語的最小發生(minimal occurrence)。對於一段落規則而言，在事件串流中的某段時間區間內，有可能會有多個述語發生；在這之中，不為其他述語的時間區間所包含的一述語發生，即稱之此述語之最小發生。我們只需擷取出述語的最小發生之開始和結束時間，即可估

算出警報所需包含的時間區段，故本研究提出四個演算法來偵測述語的最小發生，分述如下：

DirectMatch：為一非常直接且簡單的方法，透過預存(buffering)方式，我們將新近的事件儲存在一緩衝器(buffer)中，透過反向掃瞄(backward scanning)緩衝器的方式來找到最新的述語最小發生

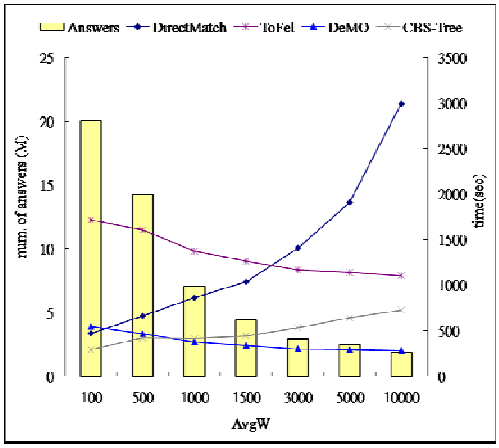
ToFel：為一往前搜尋(forward retrieval)的方法，對於述語中所描述的每個事件都建立一佇列(queue)，用已儲存相對應之事件的發生時間；隨著進入系統的事件，我們不斷更新事件相對應的佇列。佇列中所記錄的事件發生時間會根據述語中對事件順序的描述，彼此用鍊結相連，一旦當述語中所描述的最後事件發生時，我們便可透過佇列間的鍊結往前搜尋其他佇列，檢查是否能產生一完整的述語最小發生。

CBS-Tree：將串流中新近的事件以完全二元樹(complete binary tree)的索引方式記錄，若記錄在完全二元樹中的事件有 L 筆，則此二元樹則有 L 個葉節點，每個葉節點存放一事件和相對發生時間，每個內部節點(internal node)的資料則是兩個子節點資料的集合。隨著串流資料進入系統，此結構以由下往上的方式將所有樹節點資料進行更新。CBS-Tree 是個往後搜尋(backward retrieval)的方法，藉由二元樹狀資料結構所記錄的事件發生時間，能有效率找出述語最近的最小發生，避免檢查全部儲存在樹狀結構中事件資訊。

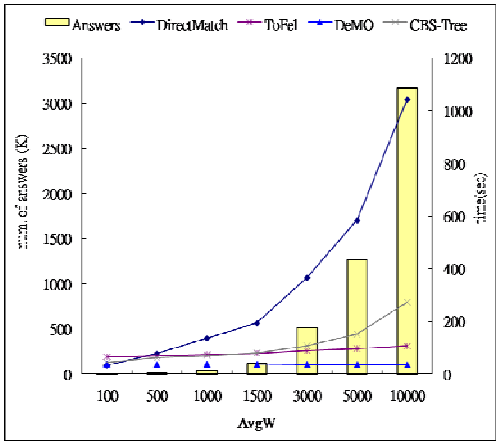
DeMo：亦為往前搜尋(forward retrieval)的一種演算法，此演算法是記錄最近發生(latest occurrence)的開始時間。此法主要是將每個串流中的事件之開始時間記錄於一述語中的最後一事件上，若目前串流中的事件恰好是一述語中的最後事件，則此時便會檢查此述語的最後事件上所記錄的所有時間點，判斷能否產生一合法的最小發生。

實驗結果

我們將四個方法實作，進行多項測試比較，僅列出較具代表性的數據如下：圖六的數據是利用真實交通資料庫測試而得，顯示不同框架大小的情況下，四個方法執行時間之比較。由於在此資料庫中，事件大都發生在相近的時間點內，因此往後搜尋的方法較為耗時。圖七的數據則是利用人造資料庫測試而得，在此數據中，四個方法執行時間皆隨框架增大而遞增，但由於往前搜尋的方法不需再檢查過去的事件，因此對於框架增大影響相對較小。



圖六：交通資料庫測試



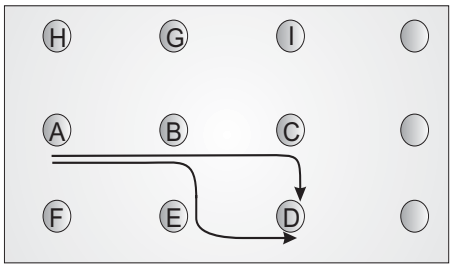
圖七：人造資料庫測試

2、 感測資料結合查詢處理技術

研究目的

過去感測器資料查詢處理技術，大部分皆著眼於選拔查詢(selection query)或資料聚合查詢(aggregation query)。然而，隨著感測器網路技術的成長與應用的普遍，在許多感測器網路的高階應用中，如物件追蹤、感測器控制、事件偵測等應用，通常需對多個感測器節點觀測值，進行彼此間之資料關連處理。以物件追蹤為例，考量一賣場中佈置的感測器系統，如圖八所示：假設感測器網路中的感測器節點承載被動式無線射頻感應模組 (passive radio frequency identification reader)，且進入賣場中的顧客購物車上皆配有

無線射頻辨識標籤，當顧客經過感應器時，感應器感應到購物車上的無線射頻辨識標籤時，會產生一筆包含顧客辨識編號的感測資料，表示該顧客曾經過該感測器所佈置的位置。在這樣的無線射頻感測器環境下，有這樣的感測資料，我們可利用不同感測器節點所感測出來之資料間關連，進一步得知使用者在賣場中的行走路徑。



圖八：物件追蹤應用

在上述的應用中，我們需對多個感測器節點觀測值彼此間進行資料關連處理。在過去我們稱這樣不同資料間關連上的處理為結合操作(join processing)；然而，現階段感測器查詢處理技術，並無法有效率地支援與處理感測器資料間結合操作。此外，在過去分散式資料庫環境中，雖有許多設計良善的資料結合處理技術，然而，有別於過去靜態的資料環境，感測器系統上的資料處理面臨的是具高即時性的動態資料串流環境，也使得過去的技術無法直接使用。為避免利用集中處理方式，將所有感測值集中蒐集於感測器網路主機，並在主機上直接進行感測器資料結合的處理，造成許多不必要的感測器能源消耗，在本研究中，我們針對感測器資料串流環境的特性，量身設計一套感測器資料結合查詢策略與最佳化技術。

研究方法

我們的感測器資料結合查詢策略與最佳化技術分為幾個部分，分述如下：

步驟一：單一結合查詢處理機制與最佳化演算法設計

當使用者下達查詢時，我們需要一套感測器結合查詢處理機制來處理使用者查詢。一個結合查詢通常會有許多等價的執行規劃，不同的執行排程順序，會造成不同的處理花費、執行效率，及資料處理延遲。因此，我們設計一感測器結合查詢之最佳化演算法，該演算法根據現下的感測器資料串流統計資料，分析所有可能的查詢規劃，產生一個最佳的結合查詢規劃。例如，如圖八中所示，針對路徑查詢 D→C→B→A，我們可選擇先將於 D 節點所新產生出來的資訊，送至 C 節點進行比對，看是否有結合結果產生，若於 C 節點並無結合結果產生，則該資料之查詢處理將可中止。但若於 C 節點有結合結果產生，則我們將所產生之結合結果再送至節點 B 而後節點 A 進行相同運算。尤以上觀察可知，針對圖八中之路徑查詢 D→C→B→A，我們將有(D→C→B→A)、(D→C→A→B)、(D→B→C→A)、(D→B→A→C)、(D→A→C→B)與(D→A→B→C)共六種可行之查詢規劃。如何從中挑選或避免最差之查詢規劃，成為單一結合查詢處理機制之關鍵所在。我們提出一個花費模型來模擬各種查詢規劃之預期花費，並且近一步地以該花費模型設計一最佳查詢規劃挑選演算法，該演算法根據貪婪法則(greedy method)於每步驟挑選結合對象時，挑選最不可能產生結果之對象進行進行結合比對，以求若最終無查詢結果產生時，能儘早將查詢處理給終止，減少不必要的感測器能源消耗。

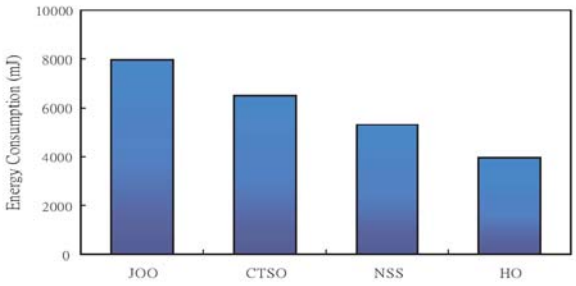
步驟二：多個結合查詢規劃之合併執行策略設計

當感測器網路朝向大規模建置發展時，可以想見，感測器網路面臨的將是一個同一時刻有多個不同查詢下達的應用環境。在這樣的環境下，除了有效率的執行單一查詢外，若我們能從眾多查詢中找出可共用的子查詢，依此規劃查詢執行策略，並分享這些子查詢所得到的結果，將有助於提升感測器網路資源的整體使用率。再以圖八中之查詢為例。於圖八中，我們給定兩個路徑查詢分別為 D→C→B→A 與 D→E→B→A。由於兩個查詢間有相同的子查詢，獨立執行的方法除了無法分享相同子查詢結果外，由於感測器節點間的通訊依賴無線傳輸協定，鄰近區域的節點間會面臨彼此間的通訊干擾與衝突，造成感測器能源利用率的下降與查詢處理效率的低落。有鑑於此，我們系統化歸納最佳的多個結合查詢之合併執行模式，利用

多個結合查詢彼此間的相同子查詢，來提升感測器資料結合的操作效率與感測器能源的利用率。我們提出一搜尋空間設定演算法，將所有有可能之共同查詢執行系統化地歸納與整理。此外，我們進一步地提出多種最佳共同查詢執行策略來有效率且準確地於所建立之搜尋空間找尋最佳共同查詢執行規劃。

實驗結果

如圖九中所示，我們將提出之共同查詢執行策略與單獨執行各查詢之策略相比，明顯可見到我們所提出的共同查詢執行方法，有效率的增進查詢處理之處理效率並降低查詢處理所需之能源消耗。其中，我們的最佳查詢搜尋演算法，將可提供約一倍之查詢效率增進。



圖九：實驗結果

三、未來工作

在未來一年的計畫執行中，就感測器資料收集研究主題方面，我們將設計一套查詢執行方法，有效率地執行多個具部分查詢區域重疊之區域聚合查詢。另外，在感測器資料探勘/查詢處理技術研究主題方面，我們亦將擴展第一年所研發的成果：單一交易串流之高頻樣型探勘技術，用以開發多重交易串流之分散式高頻樣型探勘技術。同時，結合第一、二年已發展之技術，及第三年即將開發之新技術，我們將進行以感測器網路為基礎之智慧型商店經營決策系統雛形實作。

四、成果自評

本計畫為三年期之計畫，在本年度的計畫執行過程中，我們已順利完成了原本預定於第二年完成之感測器資料聚合處理技術、事件串流之段落比對技術、及感測資料結合查詢處理技術。研究成果包含相關研究論文三篇，一篇論文以公開發表於國際知名之資料庫相關會議，一篇論文則是已被國際一流之平行及分散式系統相關期刊所接受，另一篇論文也已投稿至資料庫一流國際期刊，目前狀態為 Minor Revision。在第二年度的計畫完成之際，第三年度的相關研究包含可擴充式之感測器區域資料聚合查詢處理技術及多重交易串流之高頻樣型分散式探勘技術，幾乎都以緊鑼密鼓的展開，透過去年度及本年度計畫的執行使我們累積了大量的相關研究經驗，相信對於第三年的計畫執行將有莫大幫助。

已發表之論文

Y. C. Fan and A. L. P. Chen. An Approximation Algorithm for Optimizing Multiple Path Tracking Queries over Sensor Data Streams. DEXA2009: 532-546 (EI).

已接受之論文

Y. C. Fan and A. L. P. Chen. Efficient and Robust Schemes for Sensor Data Aggregation Based on Linear Counting. Accepted to appear in IEEE Transactions on Parallel and Distributed Systems, Doi: 10.1109/TPDS.2010.33 (SCI/EI).

審查中之論文

C. W. Cho, Y. H. Wu, S. J. Yen, Y. Zheng, and A. L. P. Chen. On-Line Rule Matching for Event Prediction. Under Minor Revision in The VLDB Journal (SCI/EI).

參考文獻

[AF00] M. Altinel and M. J. Franklin, “Efficient Filtering of XML Documents for Selective Dissemination of Information,” In *Proc. of Intl. Conf. on Very Large Data Bases*, pp. 53-64, 2000.

[AN04]A. M. Ayad and J.F. Naughton, “Static Optimization of Conjunctive Queries with Sliding Windows over Infinite Streams,” In *Proc. of the ACM SIGMOD Conf. on Management of Data*, pp. 419-430, 2004.

[AS99] M. K. Aguilera, R. E. Strom, D. C. Sturman, M. Astley, T. D. Chandra, “Matching Events in a Content-Based Subscription System,” In *Proc. of the ACM Symposium on Principles of Distributed Computing*, pp. 53-61, 1999.

[BM04] S. Babu, R. Motwani, K. Munagala, I. Nishizawa, and J. Widom, “Adaptive Ordering of Pipelined Stream Filters,” In *Proc. of the ACM SIGMOD Conf. on Management of Data*, pp. 407-418, 2004.

[CC96] C. Collet and Coupaye T., “Composite Events in NAOS,” In *Proc. of Intl. Conf. on Database and Expert Systems Applications*, pp. 244-253, 1996.

[CF02] C. Y. Chan, P. Felber, M. N. Garofalakis, R. Rastogi, “Efficient filtering of XML documents with

XPath expressions," *VLDB Journal* 11(4), pp. 354-379, 2002.

[CG05] V. Chowdhary and H. Gupta. Communication-efficient implementation of join in sensor networks. In *Proc. of Intl. Conf. on Database System for Advanced Applications*, 2005.

[CK94] S. Chakravarthy, V. Krishnaprasad, A. Eman, and S. K. Kim, "Composite Events for Active Databases: Semantics, Contexts and Detection," In *Proc. of Intl. Conf. on Very Large Data Bases*, pp. 606-617, 1994.

[CL04] J. Considine, F. Li, G. Kollios, and J. Byers. Approximate aggregation techniques for sensor databases. In *Proc. of the IEEE Conf. on Data Engineering*, pp. 449-460, 2004.

[CP06] J. Y. Chen, G. Pandurangan, and D. Xu. Robust computation of aggregates in wireless sensor networks: distributed randomized algorithms and analysis. *IEEE Trans. on Parallel and Distributed System*, vol. 17, no. 9, pp. 987-1000, 2006.

[DB88] U. Dayal, B. T. Blaustein, A. P. Buchmann, U. S. Chakravarthy, M. Hsu, R. Ledin, D. R. McCarthy, A. Rosenthal, S. K. Sarin, M. J. Carey, M. Livny, R. Jauhari, "The HiPAC Project: Combining Active Databases and Timing Constraints," *SIGMOD Record* 17(1), pp. 51-70, 1988.

[DG06] A. J. Demers, J. Gehrke, M. S. Hong, M. Riedewald, and W. M. White, "Towards Expressive Publish/Subscribe Systems," In *Proc. of Intl. Conf. on Extending Database Technology*, pp. 627-644, 2006.

[DF03] Y. Diao and M. J. Franklin, "High-Performance XML Filtering: An Overview of YFilter," *IEEE Data Engineering Bulletin* 26(1), pp. 41-48, 2003.

[FC08] Y. C. Fan and A. L. P. Chen, Efficient and robust sensor data aggregation using linear counting sketches. In *Proc. of the IEEE Symps. on Parallel and Distributed Processing*, pages 1-12, 2008.

[FJ01] F. Fabret, H. A. Jacobsen, F. Llirbat, J. Pereira, K. A. Ross, and D. Shasha, "Filtering Algorithms and Implementation for Very Fast Publish/Subscribe," In *Proc. of the ACM SIGMOD Conf. on Management of Data*, pp. 115-126, 2001.

[FJ05] M. J. Franklin, S. R. Jeffery, S. Krishnamurthy, F. Reiss, S. Rizvi, E. Wu, O. Cooper, A. Edakkunni, and W. Hong, "Design Considerations for High Fan-In Systems: The HiFi Approach," In *Proc. of Biennial Conf. on Innovative Data Systems Research*, pp. 290-304, 2005.

[FM85] P. Flajolet and G. N. Martin. "Probabilistic counting algorithms for database applications. *Journal of Computer and System Science*," pp. 31, 1985.

[GD92] S. Gatziu and K. R. Dittrich, "SAMOS: an Active Object-Oriented Database System," *IEEE Database Engineering Bulletin*, 15(1-4), pp. 23-26, 1992.

[GD94] S. Gatziu and K. R. Dittrich, "Detecting Composite Events in Active Database Systems Using Petri Nets," In *Proc. of Workshop on Research Issues in Data Engineering: Active Database Systems*, pp. 2-9, 1994.

[CD99] J. Clark and S. DeRose. "XML Path Language (XPath) Version 1.0", W3C Recommendation, http://www.w3.org/TR/xpath, 1999.

[GJ92] N. H. Gehani, H. V. Jagadish, and O. Shmueli, "Composite Event Specification in Active Databases: Model & Implementation," In *Proc. of Intl. Conf. on Very Large Data Bases*, pp. 327-338, 1992.

[GJ92a] N. H. Gehani, H. V. Jagadish, and O. Shmueli, "Event Specification in an Active Object-Oriented Database," In *Proc. of the ACM SIGMOD Conf. on Management of Data* , pp. 81-90, 1992.

[HA04] M. A. Hammad, W. G. Aref, and A. K. Elmagarmid. "Stream window join: tracking moving objects in sensor network databases." In *Proc. of Intl. Conf. on Scientific and Statistical Data Base Management*, 2004.

[MF02] S. Madden, M. J. Franklin, and J. M. Hellerstein, and W. Hong. "TAG: a tiny aggregation service for ad-hoc sensor networks." In *Proc. of the Symps. on Operating System Design and Implementation*, pages 131-146, 2002.

[MN05] A. Manjhi, S. Nath, and P. B. Gibbons. Tributaries and Deltas: efficient and robust aggregation in sensor network streams. In *Proc. of the ACM SIGMOD Conf. on Management of Data*, pp. 287-298, 2005.

[NG04] S. Nath, P. B. Gibbons, S. Seshan, and Z. R. Anderson. Synopsis diffusion for robust aggregation in sensor network. In *Proc. of the ACM Conf. on Embedded Networked Sensor System*, pp. 250-262, 2004.

[PC03] F. Peng, and S. S. Chawathe, "XPath Queries on Streaming Data," In *Proc. of the ACM SIGMOD Conf. on Management of Data*, pp. 431-442, 2003.

[SB04] N. Shrivastava, C. Buragohain, D. Agrawal, and S. Suri. Medians and beyond: new aggregation techniques for sensor networks. In *Proc. of the ACM Conf. on Embedded Networked Sensor System*, pp. 239-249, 2004.

[SM05] U. Srovastava, K. Munagala, and J. Windom. Operator placement for In-network stream processing. In *Proc. of ACM Symposium on Principles of Database Systems*, 2005.

[VN02] S. Viglas and J. F. Naughton. Rate-Based Query Optimization for Streaming Information Sources. In *Proc.of the ACM SIGMOD Conf. on Management of Data*, pp.37-48, 2002.

[WV90] K. Y. Whang, B. T. Vander-Zanden, and H. M. Taylor. A linear time probabilistic counting algorithm for database applications. *ACM Trans. on Database Systems*, vol. 15, issue2, pp. 208-229, 1990.

[WD06] E. Wu, Y. Diao, and S. Rizvi, "High-performance complex event processing over streams," In *Proc. of the ACM SIGMOD Conf. on Management of Data*, pp. 407-418, 2006.

[WL06] F Wang and P. Liu, "Temporal Management of RFID Data," In *Proc. of Intl. Conf. on Very Large Data Bases*, pp. 1128-1139, 2006.

[YL07] X. Yang, H. B. Lim, M. T. Ozsu, K. L. Tan, "In-network Execution of Monitoring Queries in Sensor Networks," In *Proc. of the ACM SIGMOD Conf. on Management of Data*, 2007.

[YG03] Y. Yao and J. Gehrke, "Query processing in sensor networks," In *Proc. of Intl. Conf. on Innovative Data Systems Research*, 2003.