行政院國家科學委員會專題研究計畫 期中進度報告

資料串流上連續型查詢處理技術之研究(2/3)
期中進度報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 95-2221-E-004-016-
執 行 期 間 ：95 年 08 月 01 日至 96 年 07 月 31 日
執 行 單 位 ：國立政治大學資訊科學系

計 畫 主 持 人：陳良弼

處 理 方 式 ：期中報告不提供公開查詢

中 華 民 國 96 年 05 月 30 日

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

資料串流上連續型查詢處理技術之研究(2/3)
Research on Continuous Query Processing Techniques over Data streams

計畫類別：■個別型計畫　　□整合型計畫
計畫編號：95-2221-E-004-016-
執行期間：95 年 08 月 01 日至 96 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　96　年　05　月　31　日

資料串流上連續型查詢處理技術之研究(2/3)
Research on Continuous Query Processing Techniques over Data streams

計畫編號：95-2221-E-004-016-
執行期間：95 年 08 月 01 日至 96 年 07 月 31 日
計畫主持人：陳良弼
執行單位：國立政治大學資訊科學系

中文摘要

由於資料串流異於傳統資料庫之特性，加上其應用十分之廣泛，因此，在目前資料工程(data engineering)領域中，是十分熱門的研究範疇，相關的應用系統雛形，及理論研究，皆在知名國際會議及期刊中，大量曝光。本計畫將以研究 DSMS 的核心技術—連續型查詢(continuous query, CQ)之處理為主軸，發展此核心技術所需之關鍵技術，如表格資料之連續型查詢處理(Relational CQ Processing)、查詢與資料串流之監控(Query and Data stream Monitoring)等。於第一年的研究成果中，除了原本僅考慮資料串流環境外，我們更將研究觸角延伸至探討有關感測器網路（sensor network）下之多查詢處理技術。因此，在本年度計畫執行過程中，我們持續發展在感測器網路上之查詢處理技術，提出了實際考慮感測器特性之相關研究報告，並發展一實作系統，將使用者介面與底層技術分離，使得使用者可利用親切介面，透過該系統指揮感測器網路之工作。另外在於查詢與資料串流之監控研究議題上，我們透過計算移動總和(moving sums)來提供資料串流之統計值概算，並提出一新式演算法，在考慮記憶體空間限制下，於資料串流環境上作高頻式樣(frequent itemsets)探勘。

關鍵詞

資料串流、連續型查詢、感測器網路、高頻式樣探勘、統計值概算

Abstract

Progress of high technologies including communication and computation leads to a more convenient life and also brings huge amounts of commercial benefits. However, rapid speed of the communication and powerful capability of the computation generate data as a form of *continuous data streams* rather than static persistent datasets, raising the complexity of data management. A data stream is an unbounded sequence of data continuously generated at a high speed. Such applications as network traffic management, web log analysis, sensor network system and traffic management system may need to handle different categories of data streams. Recently, a new type of data management system, named *data stream management system* (DSMS), has become one of the most popular research areas in data engineering field. One of the kernel technologies in DSMS, named *continuous query processing*, is developed in this project. The continuous query processing technology includes some key techniques such as *relational continuous query processing and query and data stream monitoring*. In the past one year, we have proposed some query processing techniques in the sensor network systems which are important applications on DSMS. Moreover, we also develop a sensor network system, named MAKE DB, used to provide a friendly interface for helping users to access the sensor network system without directly using the detailed underlying techniques. To the research area of "query and data stream monitoring," we propose an approach of calculating moving sums over data

streams to provide the statistics of the data stream. Moreover, a novel method optimizing memory space utilization to find frequent itemsets over data streams is also included.

一、前言

傳統資料庫管理系統(DBMS)皆建構於可恆久存在的資料(persistent data)之上，對於這些資料，DBMS 可以提供穩定的儲存、並可藉由多次來回掃瞄資料庫以達到查詢處理、更新及資料探勘之目的。然而，在許多新興的應用中，例如電腦網路、電話、金融交易或交通網路等，該資料往往會持續而快速地以串流(stream)的形式出現在應用系統內，再加上由與處理速度及儲存空間的受限，使得資料管理的困難度增加，導致傳統資料庫管理技術難以適用，此種特殊的資料型態一般被稱為資料串流(data stream)。由於資料串流異於傳統資料庫之特性，加上其應用十分之廣泛，因此，在目前資料工程(data engineering)領域中，是十分熱門的研究範疇，相關的應用系統雛形及理論研究，皆在知名國際會議及期刊中，大量曝光。

本計畫將以研究 DSMS 的核心技術—連續型查詢(continuous query, CQ)之處理為主軸，發展此核心技術之關鍵技術，包含表格資料之連續型查詢處理、查詢與資料串流之監控等。在本年度計畫報告中，我們將說明在第二年度計畫執行過程中，所發展之重要技術，包含「感測器網路之查詢處理」，及「資料串流之統計值概算與高頻樣型探勘」。

二、研究方法、進行步驟及執行進度報告

在本計畫第一年於表格資料之連續型查詢處理的研究中，除了原本僅考慮資料串流環境外，我們更探討有關感測器網路（sensor network）下之多查詢處理技術，起因於感測器網路乃是一資料串流環境中之重大應用；因此，為了與真正實際應用結合，我們將多查詢處理問題延伸至建構在感測器網路上之資料串流研究。也因為我們的研究真正的涉及實際應用的範疇，使得在實際情形下，感測器網路上的查詢處理技術，會受限於感測器本身之物理限制，如電力(power)、傳送半徑(transmission radius)等，因而加深了資料管理之複雜度，卻也提供了不同的研究議題。

另一方面，在第一年的查詢與資料處理之監控的研究上，我們利用發展查詢串流之樣型探勘技術，來支援可擴充式連續型查詢處理。而本年度我們在查詢與資料處理之監控的研究議題上，主要致力於資料串流之統計值概算與高頻樣型之探勘。資料串流之統計值概算用來分析串流資料中之統計值資訊，支援可調節式連續型查詢處理；而高頻樣型探勘的研究，則可支援查詢串流之樣型探勘技術中，樣型支持度(support)之估計。因此，在本計畫第二年度的報告中，我們將提供感測器網路查詢處理、資料串流之統計值概算與高頻樣型探勘之成果報告。

感測器網路查詢處理之成果報告

1、使用期限下之感測排程技術：

研究目的

在無線感測器網路系統上(Wireless Sensor Network)由於感測器之電力供應是由電池所提供，再加上無線傳輸和感測功能皆須消耗相當大量之電力，使得無線感測器網路的續航力相對顯得不足，而造成實際應用上其實用性的降低。然而目前在感測器網路續航力方面的研究，只著重於如何節省能源消耗量，而無有效方法能確切規劃感測器之使用期限，因此我們從另一個方向來提出一可能之解決方式，以期能達到使用者所預定的監測時限，方便監測計畫的訂定和付出成本的計算。基於上述前提之下，我們希望能藉由發展感測時程技術(Acquisition Rate)及突發事件應對之排程技術，來規劃具有『以達到任務期限為目標』之無線感測器網路，盡可能在不同電力供應下，滿足使用者查詢(Query)所需之特定突發狀況的同時，還能夠將剩餘的電量做有效之分配利用。

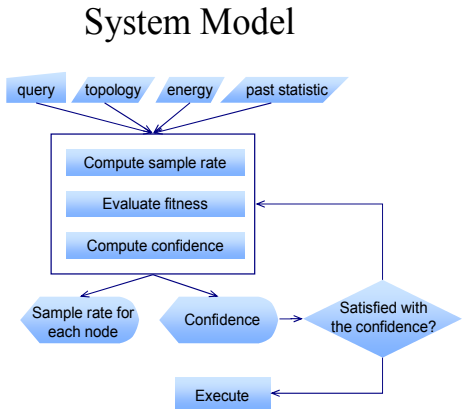
研究方法

在感測器的網路拓樸結構採用樹狀結構

的前提下，越是接近伺服器端的節點，必須負責傳遞越多下端子節點的封包；因此，所需消耗的能量將隨著子節點的數目而增加。其次在於針對使用者所要求的特定條件，其所需消耗能量的多寡若超出原先統計分析的估計值時，如何重新排程以求盡可能達到原先預定使用期限。為此，我們設計了一套能有效管理感測排程的查詢系統，針對各種不同的變因得出最佳的排程，並回覆使用者，根據此排程結果執行的可靠度(Confidence)，來決定此查詢的可行性，或得知應當需要如何修正，該系統架構如圖一所示。

本方法假設網路分布狀態和密度皆為已知，且節點均勻分佈於監測區域，原因在於均勻分佈下能平均回收各地區的資料，使得資料的代表性較高。首先我們依據過去的統計值資訊，假設資料為常態分佈(normal distribution)，對突發事件的機率和平均耗電量做一估計，接著保留估計期望值加上 X 倍的容許偏差範圍內之電量後，將剩餘電量依據節點數和階層數做平均分配。

假設各節點在初始時，電力皆相同的情況下，因為越上層的節點其負責傳遞的資料量相對越多，因此，上層的節點則成為整個系統執行時間的瓶頸(bottleneck)。在期望取得的資料有區域代表性的前提下，我們希望能平均取得各個單位區域的資料，故最佳結果應是各節點回收的資料筆數差異度最小的情形，而又同時能將電量盡量耗盡，以達到最高的資料總量。因此，在保留突發事件處理的電量之後，剩餘電量會依據上列最佳解的定義來分配使用，並求出各階層節點的任務週期(duty cycle)。而後根據統計之機率和標準差使用 Chebyshev's Inequality 來估計此一系統成功執行的可靠度，並視使用者對此可靠度滿意與否，來決定是否要增加保留給突發事件之電量，已達到更高的可靠度，或是犧牲部分可靠度來換取較高的偵測頻率(Sample rate)，來得到較多的資料筆數。



圖一：無線感測器感測排程技術之系統架構

2、以趨勢分析為基礎之感測省電技術：

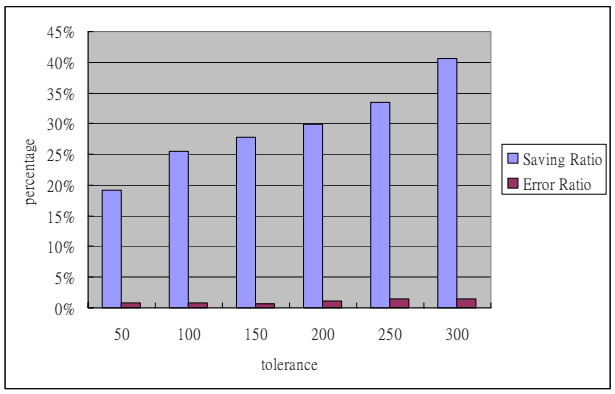
研究目的

近來由於微機電系統及通訊技術的發達，使得無線感測器的外觀越趨迷你、製造成本降低、感測功能增強，這也使得它的運用相對廣泛，舉凡軍事、醫學、環境監測、居家照顧等都是無線感測器網路的相關應用。通常無線感測器是使用電池作為電量來源，由於受限於所處環境之因素，因此，通常一旦電量耗盡，即失去功能無法再行充電，當在無法開源(增加電源供應)的狀況下，如何節流(節省電量消耗)，是無線感測器網路研究範疇中，相當重要的課題之一。

先前的感測器資料管理系統，如 TinyDB 與 TAG，其收集資料的方式為定期地偵測環境變數，並且回傳到資料管理系統，而其偵測週期是由管理者直接下達指定。若偵測時間設定較長，則較節省能源，但相對得到較少的環境資訊；若偵測時間設定過短，則相對應得到較多資訊，但卻十分耗電。在一般的感測應用中，如溫度監測等，一段時間內，如一天內的溫度變化，常具有週期性，那麼過份密集的偵測，其實並無意義，既耗電且得到的資訊皆是我們所能預期。有鑑於此，我們將利用一週期探勘的演算法，針對感測器所偵測之歷史資料進行分析，探勘這些資料的週期資訊，並利用所找出的週期特性，來動態調整偵測時間，以期達到省電的目的。

研究方法

在本研究中，其方法主要分成兩個部分：1、歷史資料趨勢分析：系統先固定一個密集性的感測時間，當收集一定資料量之後，便開始計算其週期；在得到週期之後，我們利用此週期將資料離散的切割，並把相對應時間點的資料收集在一起。我們假設週期內所有相對應時間點的資料分佈呈現常態分佈，並利用此分析得來之資料趨勢當作未來估計資料的依據。2、未來資料估計與新進資料更新：當歷史資料趨勢模型建立之後，我們將各資料模型分兩種情況來討論：a、某些過去時間點的資料分佈太過分散(資料的變異程度太大)，由於該時間點上的資料將無法有效的估算未來相對應時間點的資料，所以對應到其未來時間點時，必須要實際偵測(無法省電)。b、其他過去時間點的資料分佈較為集中(資料的變異程度較小)，對應到其未來時間點，我們有 p 的機率實際去偵測該筆資料(無法省電)，$1-p$ 的機率利用資料趨勢模型的期望值來回報(達成省電目的)。以上的兩種狀況，當有實際偵測資料時，該資料會加入資料趨勢模型中，達到更新資料趨勢的效果。利用 Chebyshev's Inequality 我們可以保證回傳的估計資料，在一定的信心水準之下，落在可容許的誤差範圍之內。實驗結果如圖二所示，我們所提出的方法在將近 50% 的省電率中，資料錯誤率僅不到 5%，証明了我們的系統在省電率與錯誤率之間提供了非常良好的平衡。



圖二、能源節省比率及資料錯誤率

3、感測器網路之近似聚合查詢技術：

研究目的

現行的無線感測器網路由於成本的考量與應用環境的限制，感測器電池的更換或充電多被視為不可行，因此無線感測器能源的節省是目前感測器相關研究中最主要且熱門的課題。在感測器節點的所有操作中，又以無線通訊最為耗電，因此目前的感測器節點採用的是 Zigbee 通訊協定，該通訊協定主要著眼點為低成本與低功耗的無線通訊。但採用 Zigbee 架構，需付出的代價則為高達約 30%的封包流失率。為提升聚合查詢的容錯能力與可靠度，我們提出一個近似聚合查詢技術，並利用統計機率分析的方式，為查詢結果提供一個近似保證。

研究方法

現階段無線感測器網路中主要有兩種資訊傳遞方式：樹狀式資料遞送(Tree-Based Routing)與多路徑式資料遞送(Multi-Path Routing)。樹狀式資料遞送方式，主要針對整個感測器網路，建構一個以主機(Host)為根節點的擴張樹(Spanning Tree)，網路節點所感測或接收之資料，皆往其母節點傳送。但如前文所述，在 Zigbee 的架構下，節點間通訊封包流失率約為約百分之三十，許多節點的感測資料將因此而遺失。另一方面，多路徑式資料遞送，則將整個網路拓樸建構為以主機為終節點的 DAG (Directed Acyclic Graph)，網路節點所感測或接收之資料，皆往其上層節點傳送，相同的資料可能會被接收多次。這樣的傳遞方式，會導致相同資料重複計數(Double Counting)。重覆計數在某些查詢下，並不會對查詢結果產生影響，例如求取網路中擁有最大溫度值的感測器編號。但是對於某些查詢，例如計算個數 count(*)，則造成錯誤的查詢結果。有趣的是，多路徑式資料遞送的方式，由於單一筆資料被接收多次，反而在 Zigbee 的架構下，擁有較佳的容錯能力(Fault Tolerance)。

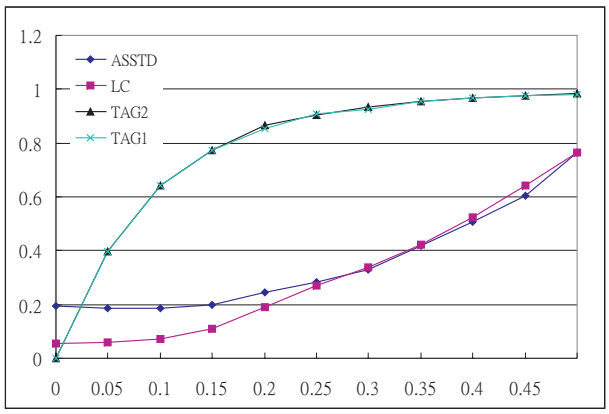
樹狀式資料遞送與多路徑式資料遞送各有其優缺點。在沒有網路錯誤發生的狀況

下，樹狀式資料遞送擁有精確的查詢結果，多路徑式資料遞送則僅能提供近似(或錯誤)的結果。而在網路狀況不佳的條件下，樹狀式資料遞送將會遺失大部分的感測結果，而多路徑式資料遞送則保有較佳的容錯能力，並提供一個較佳的查詢結果。有鑒於此，我們將提出一個在多路徑式資料遞送方式下，有效避免聚合查詢重複計算的資料摘要結構與演算法，並提供一個有誤差保證的近似查詢結果。

多路徑式資料遞送方式產生近似結果的原因，來自於對相同的一筆感測值進行多次的重複計數(Double Counting)。這樣的問題類似於傳統資料庫中，針對表格(Relation)中某個屬性(Attribute)進行相異值數量(Distinct Value)的估算問題。因此，我們引用線性計算速寫技術 (Linear Counting Technique)來避免相同感測器被重覆讀取的問題。

線性計算速寫技術主要包含一個隨機的(Randomized)雜湊資料結構。線性計算速算技術主要用來估算一個多重集合(multi-set)中，相異值的數量。給定一個多重集合，線性計算速算技術的使用方法如下。首先，產生一個長度為 m，初始值為零的位元陣列。同時引入一個均勻散佈且獨立的雜湊函式，該函式將給定之多重集合中的元素對應至所產生的位元陣列，並將所對應到的位址設定為一。接下來，將所有多重集合中的元素，對應至位元陣列。最後，計算位元陣列中，所有非零的位址數目。並利用非零的位址數目(V_n)，進行相異值的估算。相異值$\hat{n}$，可利用下列估算子(estimator)計算: $\hat{n} = -m * \ln(V_n)$。我們的估算子，可利用統計機率分析的方式，證明查詢結果擁有極高的近似保證。除此之外，我們的估算子可根據使用者所給定的允許誤差進行雜湊資料結構空間的調整，兼具查詢的準確率與節省資料結構空間使用的。

在模擬實驗中，我們使用與[CLK04]相同的實驗設定，同時實作我們所提出的方法(LC)，並以 TAG[MFH02]與 ASSTD[CLK04]為實驗的比較對象。在模擬的環境中我們下達 Count(*)的聚合查詢，比較調整通訊錯誤(communication link failure rate)來觀察各方法的容錯能力。圖三為我們實驗結果，圖中橫軸為通訊錯誤率，而縱軸為平方根標準錯誤誤差(RMS)。從圖中可以看到我們的方法明顯的擁有錯誤容錯率。並且比現階段的方法擁有更高的準確度。



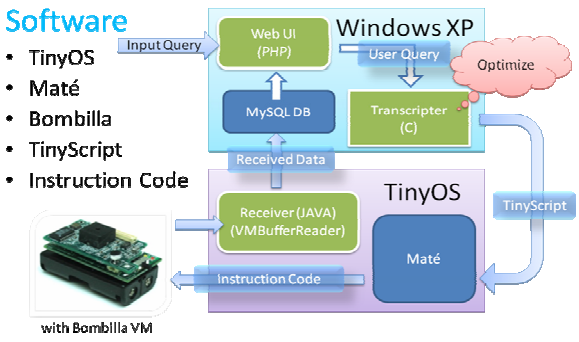
圖三、實驗結果

4、MAKE DB 感測查詢系統：

研究目的

為了提供使用者親切的web-based介面，利用和SQL相似的(SQL-like)的查詢語言，來對無線感測器網路下查詢，作資料蒐集，我們擬設計一系統將底層的低階語言與上層的使用者介面分開，讓使用者不需要接觸底層架構即可利用友善介面下查詢，因而開發了MAKE DB感測查詢系統。同時，此一系統更可提供我們在無線感測器網路相關研究上的資料蒐集。

研究方法



圖四、MAKE DB 系統架構圖

MAKE DB 的系統架構如圖四所示，其中所使用之相關元件分述如下：

NesC：由柏克萊大學所設計的一種專門用來開發感測器應用程式的程式語言(類似 C 語言)。這種程式語言是採取元件導向的結構，以元件來表達抽象的系統函式和硬體，而感測器程式就是組合各功能的元件來達到整體程式的目標。

http://www.tinyos.net/dist-1.1.0/tinyos/windows/nesc-1.1.2a-1.cygwin.i386

TinyOS：是專為無線嵌入式感測器系統所設計的作業系統，採用元件導向的結構。可依據嵌入應用，輕易增減控制執行的功能，且程式碼佔量極少，有助於記憶體空間的硬體資源精省，且能夠同時執行多個要求快速回應的控制運作。TinyOS 內部是由 nesC 所寫成的各種元件集合而成。不需行程(Process)管理，不需虛擬記憶體，不需記憶體管理，採用靜態配置記憶體技術。

http://tinyos.net/windows-1_1_0.html

Maté：Maté 是 TinyOS 的一個元件，是利用 NesC 所撰寫出來的。可以想像它是一個物件，用來提供 VM(Virtual Machine) 開發。此外 Maté 還包括了一個 Java toolchain 的部分，提供我們使用一種較簡單、較高階的 scripts language 來對 TinyOS networks 進行程式的撰寫。

http://www.cs.berkeley.edu/~pal/mate-web/rpms/mate-asvm-2.2-1.noarch.rpm

Bombilla：Bombilla 正是用 Maté 所開發出來的一個 VM 實例。而我們目前系統所使用的 Maté，在它上面執行的正是 Bombilla 這套 VM。

TinyScript：由 Maté 所支援的一種簡單且高階的 scripts language。

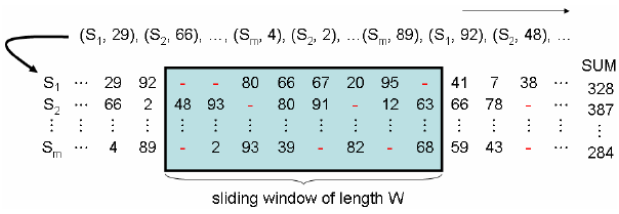
Instruction Code：一種類似組合語言的 code。我們所寫的 TinyScript 程式會被 Maté 編譯成 Bombilla 能看懂的 Instruction Code。這種 Instruction Code 會被散佈(Broadcast)到感測器網路上去執行。

<u>資料串流之統計值概算與高頻樣型探勘之成果報告</u>

1、多資料串流之統計值概算

研究目的

資料串流與傳統靜態資料庫(static database)環境之最大差異在於，在資料串流中，資料量無上限，且必須利用有限的空間得到串流資料之概要，才能進一步分析該串流資料。近年來，在資料串流環境中，監控連續資料的變化已有大量之研究成果。而在本研究中，主要考慮在 m 個不同資料來源(data sources)的環境中，收集各來源所流進來的資料，並且隨著時間演進，監控一段固定時間，得到各來源在該時間內的資料總和。以圖一為例，我們假設每筆資料皆是以(來源, 數值)的配對(pair)來表示，我們希望在長度為 W 的移動視窗(sliding window)中，維護每一資料來源在該移動視窗內的移動總和(moving sum)之統計值資訊。例如：在來源 S_1 中，S_1 在圖一中移動視窗內移動總和為 328。我們可以利用此移動總和之統計量資訊，可廣泛應用於資料串流之統計值概算，例如，可用來得知全球與特定時段中的平均溫度值，進而預測未來溫度之變化。

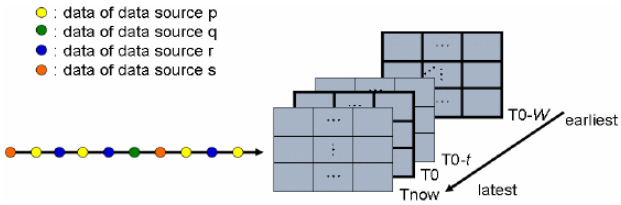


圖五、m 個來源資料中之移動總和計算

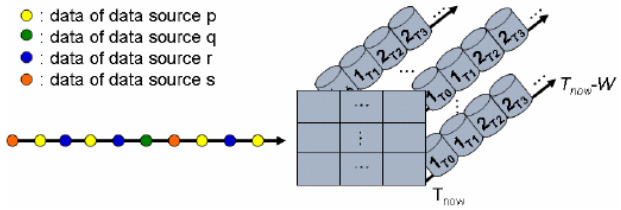
研究方法

為了處理資料量龐大的串流資料，我們以 CountMin (CM) sketch [CM05]的結構，來儲存大量累積資料之摘要（summary），以期只需依據 CM sketch 的內容便可概算滑動視窗內的所有統計值資訊，且能保證這些結果都可落在使用者所訂立的誤差範圍內。在本研究中，我們基於 CountMin sketch 的結構，以不同的觀點，提出二種方法來來計算移動總和，分述如下：一、離散方法(The discrete

method)，在本方法中，每經過 t 個時間，我們就會計算此時間段中各資料來源流入的資料，因此，在長度為 W 的移動視窗中，共有 W/t 個不同的資料片段，如圖六所示。當視窗移動的時候，只需把最舊的 t 時間內的總和資料移除，再記錄新進的總和資料即可。二、連續方法(The continuous method)，在本方法中，使用指數統計圖(exponential histogram)的技術，記錄目前最近的移動視窗中，所有流入資料的時間戳記(timestamp)與其數值，如圖七所示。並藉由指數統計圖內資料的時間戳記，刪除已離開目前移動視窗的資料，來確保指數統計圖內的資料一定是最近 W 時間內的資料，再利用記錄之數值來計算移動總和。



圖六、每 t 個時間點的 CM sketches



圖七、CM sketches 記錄指數統計圖資訊

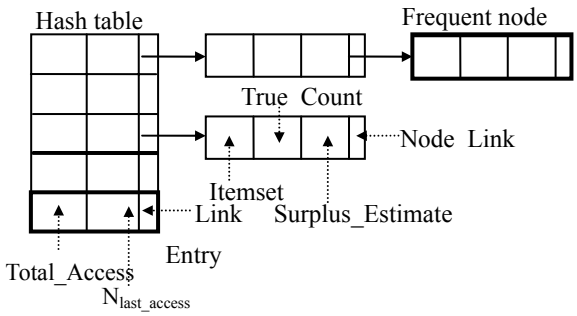
2、高頻式樣探勘

研究目的

在許多的應用上，如網路監控，無線感測器網路之資料收集等，都必須處理串流型態的資料。同時，如同以往在傳統資料庫上的資料探勘一樣，在串流型態的資料中作資料探勘，也可能發現許多有用的知識；因此，在資料串流中的資料探勘研究，儼然成為資料探勘領域中，非常熱門的研究之一。在本研究中，我們將在資料串流上作資料探勘的議題，縮小至高頻式樣探勘(frequent itemsets mining)的議題。在資料串流上探勘高頻式樣的研究中，大部分的方法，都是將原始資料串流，建立成為一個概要(synopsis)，並假設此概要能儲存於系統中，而忽略了非高頻式樣(non-frequent itemset)所佔用的空間，可能會造成系統資源大幅度地降低，因此我們在此研究中，結合了 Lossy Counting [MM02]和 hCount [JQS03]的原理，提出了一個新式的概要結構，使得原始的資料串流可以儲存在此一固定空間的概要結構中。

研究方法

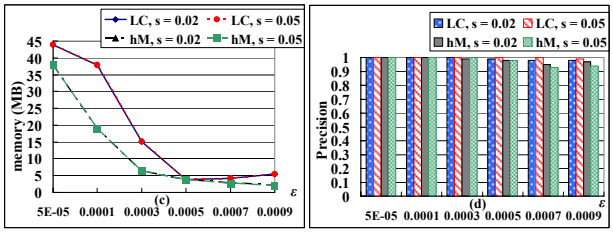
在 Lossy Counting 的方法中，其概要必須將支持度(support)大於使用者定義之錯誤容許參數(error parameter)的所有式樣儲存下來，以達到式樣的估計支持度跟真正支持度的差異，能保證在小於錯誤容許參數之內。但因為錯誤容許參數遠小於最小支持度(minimum support)，使得真正支持度落於錯誤容許參數和最小支持度之間的式樣數量，太過龐大，進而造成概要所需儲存空間過大。相對而言，在 hCount 的方法中，提出一略圖(sketch)，來將原始資料串流的資訊壓縮於一個固定空間內，並在回答高頻項目(frequent items)時，將所有項目一一檢查。因此，若將 hCount 的方法直接延伸來處理高頻式樣探勘的問題，在利用該略圖檢查每個式樣的支持度是否大於最小支持度時，會耗費相當多的時間。因此針對上述兩個方法的缺點，我們設計一新式概要，如圖八所示，此概要由兩個部分組成：



圖八：新式概要結構

1、雜湊表格(hash table)：利用一雜湊函數(hash function)將原始串流資料中每個式樣的資訊，完整記錄在此雜湊表格中，2、高頻節點(frequent node)：將所有高頻式樣的資訊，

另外儲存在高頻節點中，以期加速高頻式樣的探勘。就某一特定式樣而言，若該式樣儲存於高頻節點中，則我們可根據高頻節點上的資訊來決定該式樣的支持度；但若此式樣沒有紀錄於任何高頻節點中，我們可以根據雜湊表格中所記載之資訊，回估此特定式樣的支持度。根據實驗結果，如圖九所示，在幾乎完全相同的準確率前提下，我們所提出的新式概要在記憶體需求上，表現明顯優於 Lossy Counting。



圖九：記憶體空間及準確率之實驗結果

三、未來工作

在未來一年的計畫執行中，表格資料之連續型查詢處理方面，會探討以服務品質為考量之負荷卸載技術，同時我們亦會持續以感測器網路之查詢處理為實例，以達到跟實際應用結合之效。在串流監控技術方面，我們將探討激變偵測的相關課題，降低負荷卸載機制的負面影響，並積極地促進系統資源的重新分配，以提高系統整體利用率。另外，對於序列資料，我們將整合多數值串流上的內容篩選技術及多屬性串流上的內容篩選技術，進一步發展跨串流的內容篩選技術，用以滿足同時查詢多個異質串流的需求。

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，我們開發了多項感測器網路查詢處理技術，同時在查詢與資料串流之監控的研究議題上，亦發展了資料串流之統計值概算及高頻樣型探勘之技術，研究成果可謂相當豐富，包含了相關研究論文共五篇，及一完整之實作系統。其中，一篇論文已公開發表於國際知名會議，兩篇論文已經投稿，另外兩篇論文為本年度碩士生之畢業論文，同時該實作系統亦提供我們於相關研究上，真實資料之收集。在接近第二年度的計畫完成之際，第三年度相關之研究皆已展開，透過第一及第二年度計畫的執行過程中所累積的大量相關研究經驗，可望於第三年度中對計畫執行發揮相當大之功效。

已發表之論文

[WC06] T. C. Wu and Arbee L.P. Chen, "Maintaining Moving Sums over Data Streams," In Proceedings of the Second International Conference on Advanced Data Mining and Applications (ADMA2006) pp. 1077-1084.

已投稿之論文

[FC] Y. C. Fan and Arbee L.P. Chen, "Efficient and Robust Sensor Data Aggregations using Linear Counting Sketches," *Submitted for publication*.

[WC] E. T. Wang and Arbee L.P. Chen, "A Novel Hash-based Approach for Mining Frequent Itemsets over Data Streams Optimizing Memory Space Utilization," *Submitted for publication*.

碩士生之畢業論文

G. R. Lin. "Adaptive Power-Saving Techniques for Wireless Sensor Networks based on Incremental Data Trend Analysis," National Tsing Hua University, 2007.

C. S. Chen. "Lifetime-based Acquisition Scheduling for Wireless Sensor Networks," National Tsing Hua University, 2007.

參考文獻

[CLK04] J. Considine, F. Li, G. Kollios, and J. Byers. Approximate aggregation techniques for sensor database. In Proc. of ICDE, pages 449-460, 2004.

[CM05] G. Cormode and S. Muthukrishnan, "An improved data stream summary: the count-min sketch and its applications," J. Algorithms 55(1): 58-75

[JQS03] C. Jin, W. Qian, C. Sha, J. X. Yu and A. Zhou, "Dynamically Maintaining Frequent Items Over A Data Stream", In Proceedings of the 12th ACM International Conference on Information and Knowledge Management, 2003, pp. 287-294.

[MFH02] 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 Annual Symps. On Operating System Design and Implementation, pages

131-146, 2002.
[MM02] G. S. Manku and R. Motwani, "Approximate Frequency Counts over data Streams", In Proceedings of the 28th International Conference on Very Large Databases, 2002, pp. 346-357.