行政院國家科學委員會專題研究計畫 成果報告

合作式個人化推薦系統之進階技術研究及其應用(第 3 年)
研究成果報告(完整版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 98-2221-E-004-005-MY3
執 行 期 間 ：100 年 08 月 01 日至 101 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

計畫參與人員：碩士班研究生-兼任助理人員：吳柏輝
碩士班研究生-兼任助理人員：林以文
碩士班研究生-兼任助理人員：蔡予欣
碩士班研究生-兼任助理人員：莊坤翰
碩士班研究生-兼任助理人員：鄭瑞賢
碩士班研究生-兼任助理人員：蕭鈺翰
碩士班研究生-兼任助理人員：戴偉恒
博士班研究生-兼任助理人員：林真伊
博士班研究生-兼任助理人員：戴良光
博士後研究：張至維

報 告 附 件 ：出席國際會議研究心得報告及發表論文

公 開 資 訊 ：本計畫可公開查詢

中 華 民 國 101 年 10 月 30 日

中 文 摘 要 ： 隨著科技快速的進步發展，多樣資訊不斷在網際網路上產生與累積，因此資訊過量已成為使用者主要負擔。要如何從各種形式且大量的資料中，精確且有效率地推薦有用的資訊給使用者，成為極具價值的研究課題。本研究計畫從下世代合作式推薦系統的角度切入，發展對應領先技術涵蓋下述兩大範疇：『參考者搜尋技術』及『個人化推薦篩選』。在此三年期計畫執行過程中，我們已完成相關論文共九篇，其中五篇論文已公開發表或被接受於國際知名會議，參與計畫之學生得以參加會議報告研究成果，與國際知名學者交流，對於其研究能量累積有莫大幫助，至於其餘四篇論文亦已投稿靜待審查，本期末報告將完整詳述所有研究成果。

中文關鍵詞： 合作式推薦、索引架構、分群技術、查詢處理、社群分析、天際線查詢

英 文 摘 要 ： With rapid growth of the Internet technology, the information lacking has no longer been a problem. Instead, the information overloading starts to be a challenge. Therefore, an efficient and effective approach to assist users to precisely get the useful information from the massive dataset is needed. Among the existing solutions, the collaborative recommendation mechanism is a popular one to solve this problem. This project is based on the view point of the cooperative recommendation system, which mainly covers two issues: 1) Relevant User Selection and 2) Personalized Recommendation. In the past three years, our research results have been published in some international conferences; each of the results will be detailed in this report.

英文關鍵詞： Collaborative Recommendation, Index Structure, Clustering Technique, Query Processing, Community Analysis, Sky-line Query

中文摘要

隨著科技快速的進步發展，多樣資訊不斷在網際網路上產生與累積，因此資訊過量已成為使用者主要負擔。要如何從各種形式且大量的資料中，精確且有效率地推薦有用的資訊給使用者，成為極具價值的研究課題。本研究計畫從下世代合作式推薦系統的角度切入，發展對應領先技術涵蓋下述兩大範疇：『參考者搜尋技術』及『個人化推薦篩選』。在此三年期計畫執行過程中，我們已完成相關論文共九篇，其中五篇論文已公開發表或被接受於國際知名會議，參與計畫之學生得以參加會議報告研究成果，與國際知名學者交流，對於其研究能量累積有莫大幫助，至於其餘四篇論文亦已投稿靜待審查，本期末報告將完整詳述所有研究成果。

Abstract

With rapid growth of the Internet technology, the information lacking has no longer been a problem. Instead, the information overloading starts to be a challenge. Therefore, an efficient and effective approach to assist users to precisely get the useful information from the massive dataset is needed. Among the existing solutions, the collaborative recommendation mechanism is a popular one to solve this problem. This project is based on the view point of the cooperative recommendation system, which mainly covers two issues: 1) Relevant User Selection and 2) Personalized Recommendation. In the past three years, our research results have been published in some international conferences; each of the results will be detailed in this report.

一、前言

伴隨著資訊科技與網際網路的快速發展，文件、音樂、書籍、影片、照片與部落格等各種形式的物件在網際網路上快速且大量累積，在如此資訊豐富的環境內，資訊過量(information overloading)已成為使用者沉重的負擔。因此，在龐大且雜亂的資料中，如何精確且有效率地取得或推薦有用且使用者感興趣的資訊給使用者，即成為一個極具挑戰性的研究課題。

針對解決資訊過量所帶來的問題，目前以選擇式搜尋(selection based retrieval)技術最為人所熟知。其主要是系統透過使用者下達感興趣之物件的部分資訊，例如關鍵字，尋找並回覆包含該部分資訊之物件作為答案。然而，在某些應用中，要求使用者精確地描述其心中的喜好往往不是一件容易的事情；此外，選擇式搜尋技術更往往囿於使用者所提供的描述，限制了系統提供使用者其他有趣物件的可能。有鑑於此，本研究計畫以下世代合作式推薦系統的角度切入，發展領先技術涵蓋下述兩大範疇：『參考者搜尋技術』及『個人化推薦篩選』。

一、參考者搜尋技術

合作式推薦的概念在於當某一目標使用者需要推薦服務時，系統將參考其他使用者的資訊來進行推薦。因此，因應各種環境差異，提供不同的參考者選拔面向與選拔方式，將有助於滿足各種層面的服務需求。本計畫的在此一研究主軸探討議題包含：以內涵資料為基礎之參考者搜尋與以鏈結資料為基礎之參考者搜尋。

二、個人化推薦篩選

建構在參考者選拔的機制下，當參考者選拔出後，系統將根據參考者來進行推薦。如何妥善地應用選拔出的參考者，進一步地篩選可推薦物件，是推薦系統發展的關鍵與未來趨勢。

本研究在此一研究主軸探討議題包含：具不確定資料環境之子空間天際線查詢處理、具不確定性的串流資料中機率天際線的連續查詢處理、考慮動態天際線之範圍查詢處理和考量反向式天際線之 Top-k 查詢處理。

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。

國內外相關研究

高維度空間之任意子空間 KNN 查詢

現今的分類器(classifier)種類相當的多，像是類神經網路(neural network)、支撐向量分類器(Support Vector Machine, SVM)及受限制的庫倫能量網路(Restricted Coulomb Energy network, RCE-network)。類神經網路的分類器([KM91]、[M97]、[WI01])雖然擁有高精準度的特性，但該分類器的建構必須耗費大量的時間；支持向量分類器([B98]、[CV95])擁有高精準度且建構分類器的時間短,但是受限於不能只使用一種索引方法建構。故為了改進上述分類器的缺點，受限制的庫倫能量網路([DS93]、[MA03]、[RY98])除了擁有高精準度、建構分類器的時間短的特性之外，其僅需要使用一種索引方法建構。

在受限制的庫倫能量網路的建構演算法中，[MA03]對每個訓練用資料(training data)建立一個圓圈，半徑為離此資料點最近的不同類別(class label)的資料點的距離乘上 δ 倍，雖然此演算法精確度很高,但由於需對每個資料點都建立圓圈,使得其所花費的訓練時間相當地高。[RY98]使用預設的半徑對未被圓圈包含的資料點建立圓圈，如果此圓圈包含了不同類別的資料點，則半徑需要收縮，因此導致了其他資料點也脫離了此圓圈，所以需要再重頭為脫離的資料點做處理，造成了資料點的多次掃描。

具高準確率之鏈結樣式社群探索技術

在社會網絡(social networks)中，針對不同的應用環境，社群(community)的定義也會有所不同。大部份社群探勘的研究([CS93]、[DH01]、[FL00]、[IK05]、[SM00])都是將社群定義為在同一社群內的實體(entities)之間是緊密互動的,但是屬於不同社群的實體必須是稀疏互動的。然而，在某些應用環境中，例如：部落格空間的使用者互動關係，那些與相似使用者互動的使用者，其隱含了他們擁有相同的興趣，因此即使他們彼此間可能從未互動過，但是他們仍屬於同一社群，[KR99]及[RK01]因而提出新興社群(emerging communities)的觀念。

為了概括各種不同社群的定義,[LW07]提出鏈結樣式社群(link-pattern based communities)的觀念。鏈結樣式社群是在同一社群內的實體擁有相似的鏈結樣式(link patterns)，也就是，他們具有相似的對內互動行為與對外互動行為。且提出 CLGA 演算法，藉由矩陣相似(matrix approximation)最小化的技術，來尋找社會網絡內的社群結構(community structure)。

尋找影響力最大化的領導者

在社會網絡中，每個使用者與其他使用者的互動關係不盡相同，因此，每個使用者本身所帶有的影響力亦不盡相同。[KK03]根據使用者間的互動關係，定義使用者本身所帶有的影響力，且提出找尋 k 個能達到影響力最大(influence maximization)的使用者之問題。 [KK03]證明該問題屬於 NP-hard 的問題，因此使用貪婪演算法(greedy algorithm)來求得近似解(approximate solution)，而該演算法保證所找到的 k 名使用者所涵蓋的影響力至少會是最佳解的 63%。然而，貪婪演算法每次在選擇一個影響力最大的使用者之後，必須重新計算所有剩

餘使用者的影響力，因此其執行效能仍不盡理想。[LK07]亦採用貪婪演算法解決該問題，不同的是在每一回合中先將所有使用者依據影響力大小進行排序，且由影響力最大的使用者優先考慮，以有效節省那些不可能成為領導者的使用者的計算時間。

[CW10]以期望影響範圍的觀念，重新定義使用者所涵蓋使用者族群的範圍。每個使用者存在一個介於 0 到 1 之間的活躍機率(activation probability)，用來代表這個使用者會受到其朋友至少一人影響的機率，當值越大，代表這個使用者會受影響的機率也越高。根據活躍機率，可以定義每個使用者的影響範圍，稱為增值影響力(incremental influence)。假設當某個使用者的活躍機率從現有的值直接提升至 1 時，這個增加的幅度會影響其周圍使用者的活躍機率提升的數值總和，即為增值影響力。[CW10]利用建立索引(index)的方式，紀錄每個使用者影響範圍內的所有使用者。因此，在每一回合選擇領導者之後，只須更新那些受影響的使用者即可。雖然[CW10]的方法較原先的貪婪演算法[KK03]有效率，但是仍必須花費相當多的時間在索引建立上。

具不確定資料環境之子空間天際線查詢處理

在 2001 年，[BK01]提出天際線運算(skyline operation)，將資料根據多個挑選準則(multi-criteria)，作適當的篩選，其廣泛應用於推薦式系統內。然而，天際線運算必須花費龐大的執行時間及記憶體空間，因此，許多改進計算天際線的相關研究也接踵而來，有些利用建立索引(index)加快速度，如：[KRR02]、[LZ07]，或是利用非索引(non-index)，將資料先做簡單計算，儲存在資料結構中，以利計算，如：[CG03]、[GS05]。亦有些研究是針對在子空間中處理天際線查詢[PF07]、[PJE05]、[TX06]，依據使用者所挑選出來的屬性，進行天際線的計算。

在一些實際應用中，如：無線網路感測器或是物件式的資料型態，資料本身是具有不確定性的性質。由於資料本身是不確定性，舊有的天際線方法無法明確地判斷該類型的資料是否屬於天際線，因此[PJL07]針對不確定性的物件，定義不確定物件彼此之間支配的關係，及每個不確定物件成為天際線機率的數值，所有天際線機率值大於使用者給定的門檻值的不確定物件即為機率天際線。然而，在實際應用中，使用者通常對於資料本身沒有預先的知識，故不容易設定一適當的門檻值。

具不確定性的串流資料中機率天際線的連續查詢處理

大部份天際線運算的研究仍侷限於靜態(static)的資料，在即時應用環境中，資料隨著時間不斷地進入系統，處理天際線查詢的問題就面臨如何在動態資料改變下，能有效率地回答天際線的考驗。在[LY05]和[TP06]研究中，為了減少針對串流資料去處理天際線查詢的計算成本，利用緩衝儲存器(buffer)來儲存確定是或有機會是天際線的資料，將不可能為答案的資料捨棄。然而這些研究方法都是針對確定性的資料(certain data)進行天際線運算。

[LS08]和[ZL09]是針對在不確定的串流資料下，提出處理機率天際線查詢的方法，而此二研究中所定義的不確定資料是資料本身存在一個機率值。然而，在某些應用環境中，如：在物件式的資料型態，一個物件是由至少一個實例(instance)所構成，當該類型的資料隨著時間不間斷地進入系統，如何快速地處理機率天際線的查詢是一極大挑戰。

考慮動態天際線之範圍查詢處理

天際線查詢[BK01]的結果包含至少在一屬性中其值相對於其他物件而言是最佳值的那些

物件。然而，[PT05]發現，在實際應用中，使用者較希望能以自身的需求當作查詢的條件，且回傳的答案必須盡可能接近該需求，因而，提出動態天際線查詢(dynamic skyline)的觀念。[PT05]將物件儲存於 R-tree，且根據使用者的需求，尋找在 R-tree 內哪些物件與該使用者需求較為接近，進而回傳動態天際線。[SS06]提出空間(spatial)動態天際線查詢，依據物件與使用者需求的尤拉距離(Euclidean distance)來決定其距離，[CL08]則是提供一較廣泛的應用環境，只要物件與使用者需求是在度量空間(metric space)內，即可快速地計算動態天際線。上述的研究都必須將每一物件與使用者需求計算其在各屬性的距離，才能決定哪些物件是屬於動態天際線。為了解決該問題，[SB08]只需計算部份可能是答案的物件，以加快回傳動態天際線的時間，然而，該方法沒有事先對資料建立索引，故處理時間仍過長。

考量反向式天際線之 Top-k 查詢處理

動態天際線(Dynamic skyline)[PTF05]站在使用者的角度，分別考慮各個規格是否較符合使用者的需求，利用優勢策略(dominance strategy)排除使用者不可能選出的產品，而使用者則只會從這些不被優勢策略排除的產品中挑出喜歡的產品，動態天際線的好處是無論使用者對這個規格的重視度不同，使用者仍只會從這些產品中挑出。而反向式天際線(Reverse skyline)[LC08]則是站在產品的角度找出產品有被那些使用者視為動態天際線，並且研究出了較有效率的演算法去計算出一個產品的反向式天際線。[WT09]提出了 BRS 演算法改進了[LC08]的做法並加速反向式天際線計算。

二、研究方法、進行步驟及執行進度報告

我們於過去三年中所發展之兩項關鍵技術：『參考者搜尋技術』及『個人化推薦篩選』，成果報告如下所示：

參考者搜尋技術之成果報告

1、以內涵資料為基礎之參考者搜尋：高維度空間之任意子空間 KNN 查詢

研究目的

在推薦式系統中，可利用現有的資料建構一分類器，將使用者的需求資訊精準的分類到一類別中，並將同類別中符合使用者需求的資料推薦給使用者。然而，就多媒體資料而言，其特徵空間通常擁有相當高的維度，且在高維度的空間中，建構一分類器是相當耗時的工作。為了避免使用者等待過長的時間，如何在高維度空間中有效率的建構一個精準的分類器已成為一重要的問題。

與其他種類的分類器相比，受限制的庫倫能量網路已被證實具有高精準度、建構分類器的時間短、且可用一種索引方法建構的優點。然而，受限制庫倫能量網路使用圓圈來框住訓練資料，在同一圓圈內的訓練資料必須是屬於同一類別，且每一資料必須被一個圓圈框住；該些限制會使得其必須花費相當高的訓練時間。因此，我們利用受限制庫倫能量網路分類器的特性，針對高維度空間內的資料，設計一演算法，以有效率的建構一個高精準度的分類器。

研究方法

假設訓練用資料的物件存在某存取的順序，而我們建構受限制的庫倫能量網路的演算法則是依該順序一一存取訓練用資料內的物件。每個物件 p 搜尋每個圓圈 C，如果圓圈的中心點 b 與 p 是相同類別，則測試 b 與不同類別但為最近物件 nb 的距離，若 b 與 p 的距離小於此

距離，則 p 可被此圓圈 C 包含，且圓圈 C 的半徑改設為 b 與 p 的距離。如果 p 對每個圓圈 C 都找不到相同類別，或 p 對找到的相同類別圓圈 C 的中心點 b 的距離大於此中心點到不同類別但最近物件 nb 的距離的話，則建立一個以 p 為中心點的圓圈，半徑為 0，並計算 p 與最近但不同類別的物件 nb 的距離，將此距離紀錄起來。當每個物件都能被一個圓圈所包含，則完成受限制的庫倫能量網路的建構。

我們發現在建構受限制的庫倫能量網路的過程中，必須對每一物件找到與它屬於不同類別但為最近的物件，其是最花時間的工作，所以我們採用多個參考點方法([BF96]、[BS97]、[YO01])的觀念，提出一索引方法，以快速地找出查詢點的最近物件。一開始取出多個參考點，對每個參考點建立 B+-tree。若資料點 p 越接近查詢點 q，則資料點 p 與參考點 r 的距離跟查詢點 q 與參考點 r 的距離越相近，故我們選擇查詢點 q 在每個 B+-tree 的鄰近葉節點(leaf node)的資料點當作候選的最接近最近鄰居(candidate approximate nearest neighbors)，一直循環的對每個 B+-樹取出資料點，直到取出的資料點滿足使用者給定一個門檻值 λ。在這 λ 個資料點中，找出距離最接近查詢點 q 的資料點 a 當作最接近最近鄰居。利用此資料點 a 與查詢點 q 的距離，快速的刪除不可能是最近鄰居的資料點。

另外，為了更進一步地降低建構受限制的庫倫能量網路所需花費的處理時間，若資料點 p 與查詢點 q 在前 k 個維度的平方和大於或等於最接近鄰居 a 與查詢點 q 的距離平方，則此資料點 p 不可能是最近鄰居的資料點。而維度計算的先後次序，是以每個維度的變異數(variance)來當作依據，變異數越大的維度優先計算。

實驗結果

我們所提出的方法(RE algorithm)與一支撐向量分類器演算法(LIBSVM)及兩個受限制的庫倫網路演算法(Mu's algorithm[MA03]、Rajan's algorithm[RY98])做比較。實驗使用合成資料和真實資料來比較各個演算法的訓練時間和精準度。

- 使用合成資料(synthetic data)的實驗

合成資料的產生是先隨機對每個類別產生圓圈的中心點，然後對每個中心點計算與其他中心點最近的距離，換句話說，計算與最近的不同類別中心點的距離，然後以此距離的 0.49 倍當作半徑，在這個半徑範圍內產生資料點。實驗所使用的參數及描述如表 1 所示。

當我們固定 D 與 C，而對 N 作變化，即增加資料量，實驗結果如圖一所示。我們可以看到除了 Mu 演算法之外，其他三個演算法的效能都跟資料量成線性成長，而我們的演算法(RE)效能最好。

- 使用真實資料(real data)的實驗

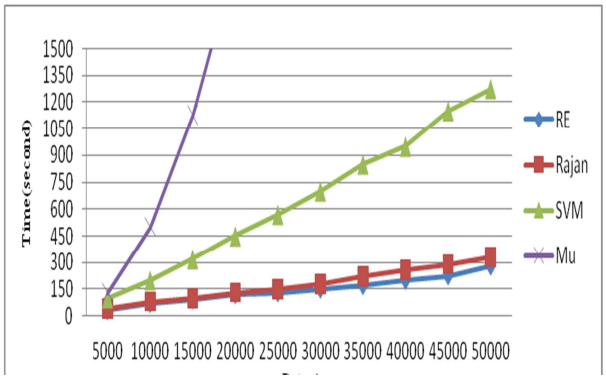
使用四個真實資料庫來比較各個演算法的訓練時間和精準度，表 2 為各資料庫的相關設定。

由表 3 得知，三個受限制的庫倫網路建構演算法的精準度均比支持向量分類器好。另外，由於 Mu 跟 Rajan 演算法所使用的圓圈數量較多，因此其精準度均比 RE 演算法好，但其差距約在 3%以內。另外，由表 4 得知，RE 演算法比其他演算法的訓練時間少得多。

整體而言，RE 演算法在精準度及訓練時間皆比 SVM 來的好，而與同為受限制的庫倫網路分類器的兩種演算法(Mu 和 Rajan)作比較，RE 演算法在精準度上與 Mu 和 Rajan 差距約在 3%以內，但訓練時間卻遠小於 Mu 和 Rajan，最大差距甚至達到 78.24 倍。

表 1、參數設定

參數	描述
D	維度數
C	類別數
N	每個類別的資料數



圖一、資料量變化對各演算法的影響

表 2、真實資料庫的相關設定

資料庫	類別數	訓練資料量	測試資料量	維度數
SHUTTLE	7	43,500	14,500	9
NBUST	10	60,000	10,000	780
ML FACE	20	564	60	15,360
UM FACE	20	912	100	32,400

表 3、各個演算法的精準度

DB	SHUTTLE	MNIST	ML FACE	UM FACE
RE	99.16%	95.43%	98.33%	99%
Rajan	99.74%	96.40%	98.33%	99%
Mu	99.90%	97.04%	98.33%	100%
LIBSVM	97.61%	94.46%	98.33%	94%

表 4、各個演算法的訓練時間(單位：秒)

DB / Algo.	SHUTTLE	MNIST	ML FACE	UM FACE
RE	1.937	1377.531	10.328	37.406
Rajan	5.938	3753.938	24.828	89.861
Mu	151.547	7893.406	81.75	387.782
LIBSVM	96.172	1380.063	65.375	366.86

2、以內涵資料為基礎之參考者搜尋：考慮最近和最遠鄰居為基礎之天際線查詢

研究目的

假設系統想推薦給使用者一間房子，他們會考慮房子附近的公共設施，而公共設施有好有壞，有些是好的公共設施，像是學校和圖書館，甚至是工作場所。而有些是不好的設施，像是舞廳和垃圾場。如何讓系統能夠在一些房子當中找到遠離不好的設施又靠近許多好的設施是本項目的研究目的。

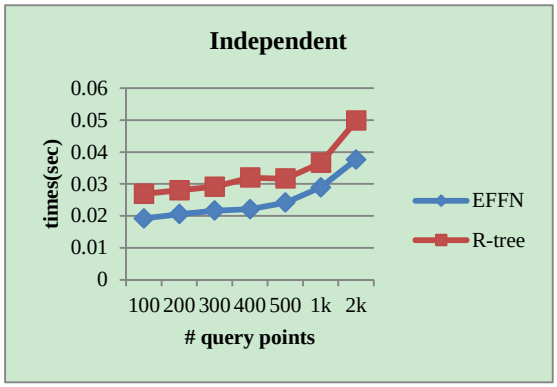
問題：一般而言，若是有若干間想買的房子，而且都想靠近好的設施且遠離不好的設施的前提下，本篇研究用距離最近不好的設施的距離當作衡量標準，其意義是代表其他不好的設施都在這個範圍以外。另一方面，這邊會找最遠的好的設施，其意義也是相反，代表其他的好的設施都在這個範圍以內。再用天際線查詢，以這兩種距離維度當作衡量標準，回報天際線查詢給使用者。

研究方法

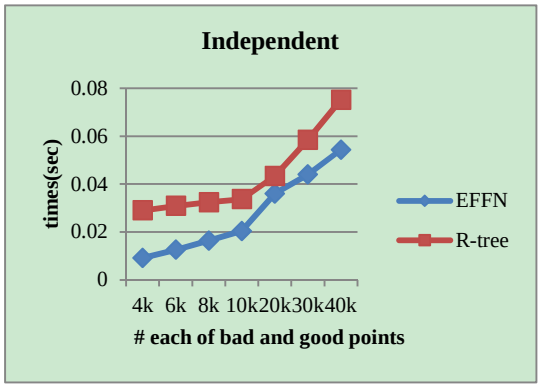
方法主要是提出變形的 Quad-tree，稱為 Grid Quad-tree，能夠利用同一種資料結構找出最近和最遠鄰居。而我們也利用將每個格子編號，再提出一個方法能夠讓我們知道附近格子的編號，進一步去搜尋附近格子中的點。在最近鄰居的方法中我們利用上述的方法，在查詢點附近搜尋點，一步一步縮小範圍，最後得知最近鄰居的距離。接著我們在尋找最遠鄰居時，我們提出兩個 Lemma 能夠利用天際線的性質，讓我們能夠利用上述的特性找出最遠鄰居，同樣的，我們在最遠鄰居當中也是慢慢搜尋有可能的格子而找出最遠鄰居的距離。最後，我們提出 EFFN 演算法，在尋找天際線過程中，先刪除一些不可能成為天際線的查詢點，最後回傳天際線查詢給使用者。

實驗結果

在執行時間上優於對手，如圖二與圖三實驗數據圖。接著在未來，我們希望能夠讓使用者能夠選擇自己喜愛的公共設施，甚至能夠給一些權重，讓這個應用能夠更貼近使用者需求。



圖二



圖三

3、以鏈結資料為基礎之參考者搜尋：具高準確率之鏈結樣式社群探索技術

研究目的

在社會網絡中，針對不同的應用環境，社群的定義也會有所不同。大部份社群探勘的研究是考慮在同一社群內的實體之間是緊密互動的，但是屬於不同社群的實體必須是稀疏互動的。然而，在某些應用環境中，與相似使用者互動的使用者，其隱含了他們擁有相同的興趣，因此即使他們彼此間沒有任何的互動，但是他們仍然屬於同一社群。為了概括各種不同社群的定義，[LW07]提出鏈結樣式社群的觀念。

鏈結樣式社群是同一社群內的實體擁有相似的鏈結樣式。[LW07]另提出一社群雛型圖(community prototype graph)的觀念，來表示社群結構，且設計一個迭代演算法 CLGA，藉由矩陣相似(matrix approximation)最小化的技術，來尋找一個與社會網絡最為相似的社群雛型圖。

該演算法必須將社會網絡與其社群雛型圖建構成相似度矩陣(affinity matrices)，且在每一次迭代中，每一實體必須使用窮舉搜尋(exhaustive search)，以決定其屬於哪一社群。然而，社群網絡通常是由數十萬、數百萬個實體所組成，因此 CLGA 演算法必須要求龐大的記憶體空間。另外，每一實體為了決定所屬的社群，必須採用窮舉搜尋的方式，因此 CLGA 演算法要求龐大的執行時間。

研究方法

根據鏈結樣式社群的定義，也就是在同一社群內的實體必須具有相似的對內互動行為與對外互動行為，我們重新設計一目標方程式(objective function)。為了能夠達到在同一社群內的實體，必須盡可能地擁有相似的對內互動頻率以及對外互動頻率的目的，因此，該目標方程式可以分成二個部份來看：第一、同一社群內的實體間的互動頻率必須與整體內部平均互動頻率差異愈小愈好。第二、對某二個社群間實體的互動頻率亦必須與整體彼此間平均互動頻率差異愈小愈好。另外，在我們研究中發現，利用該目標方程式所找到的最佳社群(optimal solution)與藉由 CLGA 演算法所找到的最佳社群是相同的。

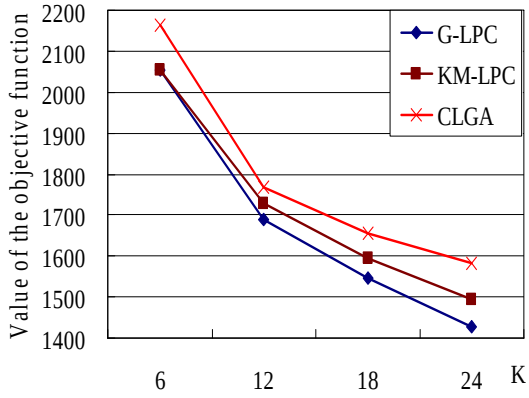
由於我們所設計的目標方程式是一個優化的問題(optimization problem)，利用窮舉搜尋的方式找到全域的最佳解(global optimum solution)不是一個可行的方法，因此，我們先從社會網絡中挑選出數個取樣實體，再採用凝聚式的階層式分群法(agglomerative hierarchical clustering)得到 k 個初始的社群質量中心(initial community centroids)。接著，因每一實體與其他的實體間的互動頻率可視為它的特徵向量(feature vector)，故實體的特徵向量與初始的社群質量中心的距離則可當作它與該社群的相似度之依據。因此，將每一實體加入至與自己本身差距最小的社群質量中心。經由該流程，在同一社群質量中心的實體為同一社群的成員，此時已完成社群的初始化。

根據該初始的社群，我們提出以貪婪為基礎的演算法(G-LPC)及以 K-Means 為基礎的演算法(KM-LPC)。G-LPC 演算法是採用迭代的方式，在每一迭代中，每一實體會改加入至可以使得目標方程式的值最小化之社群。而 KM-LPC 演算法亦是採用迭代的方式，在每一迭代中，每一實體的特徵向量與每一社群的質量中心計算其距離，且將該實體改加入至與質量中心距離最小的社群內。當每一實體都已決定所要加入的社群後，該二演算法都會依據社群內的實體，更新其質量中心。不斷地重覆上述流程，直到沒有一實體改變其社群為止。

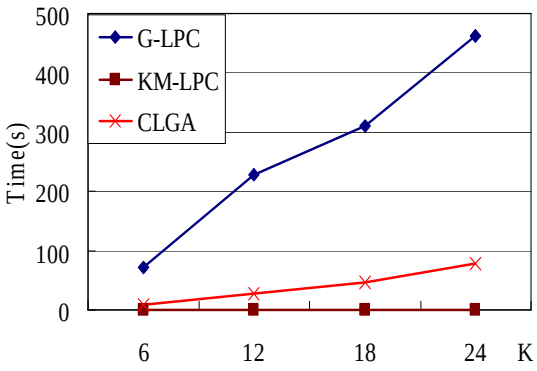
實驗結果

我們使用 Enron 郵件資料庫，在 Enron 資料庫內，有 151 員工互通電子郵件的資訊。我們將 151 員工當作社會網絡中的實體，且員工如果曾經互相傳送電子郵件的話，則相對應的節點會用一條鍊結(link)相互連接，因而產生一個 0/1 的圖形(graph)。

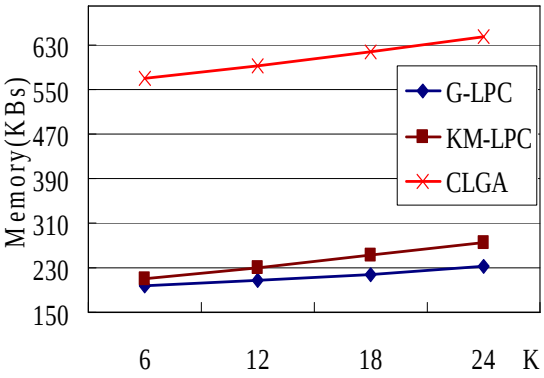
在圖四中，我們所提出的二個演算法均比 CLGA 演算法所得到的分群結果佳，其中，G-LPC 演算法可得到最佳的分群結果。由圖五得知，KM-LPC 演算法所需要的處理時間遠比其他二個演算法少得多。最後，由圖六得知，G-LPC 演算法與 KM-LPC 演算法所需要的記憶體空間遠比 CLGA 演算法來得少。總體而言，雖然 G-LPC 演算法需要花費最多的執行時間，但其所得到的分群結果卻是最好的。另外，不論是在分群結果、執行時間或記憶體要求，KM-LPC 演算法都比 CLGA 演算法佳。



圖四、不同的社群個數對分群結果的影響



圖五、不同的社群個數對處理時間的影響



圖六、不同社群個數對記憶體要求的影響

4、以鏈結資料為基礎之參考者搜尋：尋找影響力最大化的領導者

研究目的

社會網絡可視為個體(individual)間之互動情形所形成的圖形(graph)，圖形中的每個節點(vertex)代表一位使用者，而連接兩個節點的邊(edge)則代表兩個使用者之間存在著某種互動關係(interaction)，最常見的即為朋友關係。當兩個使用者存在互動關係時，則我們稱其中一個使用者對另一個使用者擁有影響力(influence)，這個影響力可能是單向也有可能是雙向的。當一個領導者所涵蓋的使用者族群越大時，其所擁有的影響力也越大。

社會網絡通常呈現 Power-law 的分佈，也就是只有少部份的使用者擁有很大的影響力，絕大部分的使用者都僅擁有很小的影響力。因此，我們希望能藉由該特性將大部分影響力很小的使用者排除，以提升演算法的整體執行效能。

研究方法

在本研究中，我們提出 ECE 演算法(Efficient Candidates Elimination algorithm)，其藉由一門檻值以有效地過濾所有不可能被選為領導者的使用者，使能大幅減少所需要檢查的資料數量，進而增進貪婪演算法的執行效率。

根據增值影響力[CW10]的定義，每個使用者的影響範圍會與其被選為領導者的順序有關。對某個使用者來說，會影響其增值影響力的使用者有三種，分別為向外路徑鄰居(out-path neighbor)、向內路徑鄰居(in-path neighbor)及競爭(competition)使用者。向外路徑鄰居即為在整個社會網絡上，能夠由該使用者出發藉由圖形的邊連線到的使用者；向內路徑鄰居則是相反。而競爭使用者則是在整個社會網絡上，該使用者所有向外路徑鄰居的向內路徑鄰居。只要這三種使用者任何一個比該使用者更早被選為領導者，皆會造成該使用者的增值影響力的數值下降。

由於貪婪演算法每次都會挑選增值影響力最大的使用者成為領導者，因此對某個使用者 a 來說，我們考慮上述三種會影響其增值影響力的所有使用者集合 B，只要使用者 a 的增值影響力比在 B 集合中所有使用者的增值影響力都大的話，我們就可以保證在這些使用者之中，使用者 a 被選為領導者的順序一定最優先。反之，如果我們無法保證 a 與某個在 B 集合中的使用者 b 的增值影響力何者較大時，則我們會去計算該使用者 b 比使用者 a 更早被選為領導者的情形下，使用者 a 的增值影響力。經過不斷的重複檢查，直到所有剩餘在 B 集合中使用者的增值影響力都較使用者 a 的值小的時候，即代表在這些使用者之中，使用者 a 被選為領導者順序最差的情形，此時使用者 a 的增值影響力即為在貪婪演算法中的下限值。

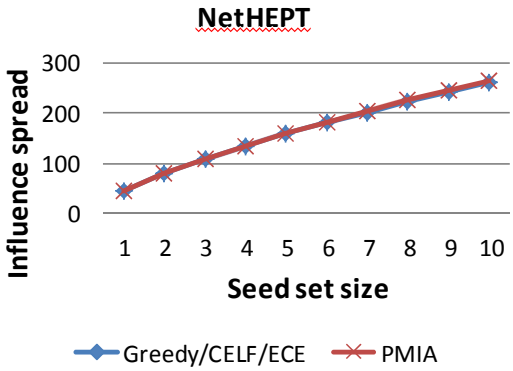
由於我們希望推薦 k 個使用者，因此挑選增值影響力的下限值排名第 k 名的使用者，其增值影響力的下限值即為門檻值。當某個使用者的增值影響力小於此門檻值時，即代表這名使用者不可能被選為領導者，因為至少有 k 個使用者大於該門檻值。因此，我們只須對所有增值影響力大於該門檻值的使用者進行計算即可。此削減策略(pruning strategy) 過濾掉所有不可能被選為領導者的使用者，以達到減少資料檢查數量之目的，進而增進整體演算法的效能。

實驗結果

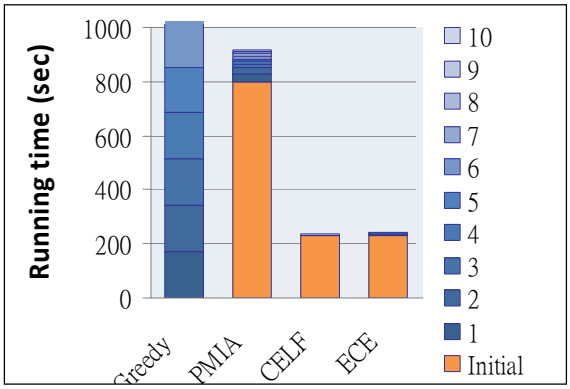
以 NetHEPT 學術協作網路的資料作為實驗資料集，其中，每一節點代表一名作者，若某兩名作者曾一同發表論文，則會有一條邊連接代表該兩名作者的節點，而該資料集內共包含 15,233 個作者及 32,235 條邊。在實驗部份，我們根據[CW10]所提出來的模式來定義每個使用者的影響範圍，且以 Greedy 演算法[KK03]、CELF 演算法[LK07]、以及 PMIA 演算法[CW10]

作為本研究的比較對象。

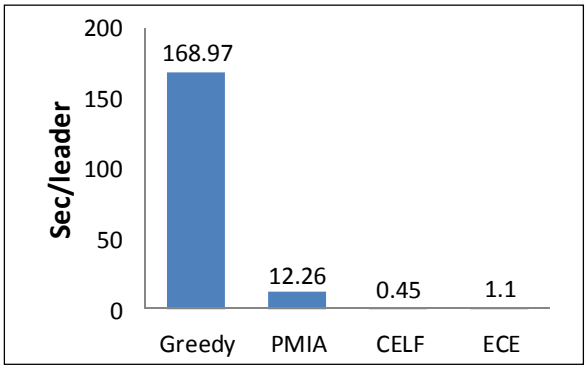
在本實驗中，領導者集合的大小設定為 10，而每條路徑(path)最低機率的門檻值設定為 0.1。由圖七得知，該四種演算法所找出來的領導者集合之影響範圍是不相上下的，其表示該四種演算法尋找領導者的能力是相同的。



圖七、領導者集合的影響範圍



圖八、各演算法的執行時間



圖九、各演算法找出一個領導者的平均時間

在圖八中，以橘色呈現的初始時間(initial time)表示建立索引所需花費的時間。從圖八得知，該四種演算法在執行時間上呈現了三種不同的等級，PMIA 演算法僅僅用了 Greedy 演算法一半的時間，但是其必須花費許多時間在建立索引上；CELF 演算法與 ECE 演算法相較之下又比 PMIA 演算法更加快速。在找尋每個領導者所耗費的平均時間亦呈現相同的三種不同等級，其結果如圖九所示。

在我們所提出的 ECE 演算法中，經由削減策略過濾後，可將 NetHEPT 內原有的 15,233 位作者刪除至只剩下 15 個可能成為答案的作者，其削減比例(pruning rate)為 99.90%。此外，在計算削減策略所需要的資訊所額外付出的時間只佔總執行時間的 1%。因此，根據實驗顯示，我們所提出來的方法僅需額外耗費非常少的時間，卻能大幅增進整體貪婪演算法的效能。

個人化推薦篩選之成果報告

1、具不確定資料環境之子空間天際線查詢處理

研究目的

[BK01]提出針對具確定性的資料作天際線查詢的處理，且該查詢已廣泛應用於推薦式系統。然而資料在多數應用中是具有不確定的性質存在，故[PJL07]針對具不確定的資料，定義每一資料是屬於天際線的機率。另外，[PJL07]根據使用者給定的門檻值，回傳天際線機率高於門檻值的資料給使用者。然而，使用者通常對於資料本身沒有預先的知識，故門檻值是不易設定的。為了避免設定門檻值高低所造成回傳結果可能太多或太少的缺失，使用者只需要指定一個數值 k，系統會自動回傳前 k 個天際線機率最高的資料給使用者。此外，系統可依據使用者所挑選的屬性，找出前 k 個天際線機率最高的資料為推薦結果。

研究方法

一個不確定物件(uncertain object)是由一個以上所屬此物件的實例所構成。為了要回傳前 k 個天際線機率最高的物件，最直覺的方法即是計算出所有不確定物件的天際線機率，而前 k 個天際線機率最高的物件即為結果。然而，精確地計算出所有不確定物件的天際線機率的成本是相當高，因此，必須盡可能地將不可能為結果的物件刪除。

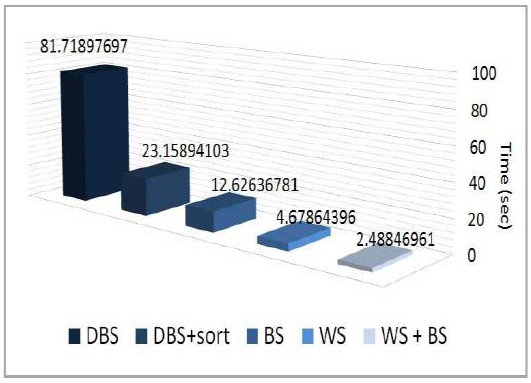
計算出所有不確定物件的天際線機率之上限值(upper bound)，再取出第 k 個天際線機率上限值，若某不確定物件的天際線機率上限值低於第 k 個天際線機率上限值，則該物件一定不是答案，因此，可直接刪除該物件。由於第 k 個天際線機率上限值會決定刪除能力，因此如何加速計算第 k 個天際線機率上限值是非常重要。

最直覺的作法為隨機挑出 k 個不確定物件，精確計算出天際線機率，以最小的天際線機率為第 k 個天際線機率，接著，一一取出其他不確定物件且計算其天際線機率上限值，若此值小於目前第 k 個天際線機率，則刪除此不確定物件，否則，則繼續計算，將天際線機率上限值更將精確(refine)，直到求出精確的天際線機率，再判斷是否為結果之一，若是的話，則此時將第 k 個天際線機率更新，反之，繼續處理下一個不確定物件。

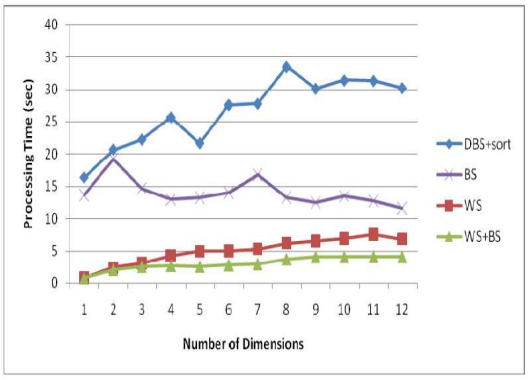
由於隨機挑選 k 個不確定物件，若挑選不當，很可能會將第 k 個天際線機率的刪除能力降低，為了能有更佳的刪除能力，因此先找出每個不確定物件的最佳實例(best instance)，將所有最佳實例採用 SFS 演算法[CG03]計算出排序數值，再挑選 k 個排序數值最小的作為一開始的 k 個不確定物件。另外，我們提出 WorstSky 及 BestSky 策略以求出更精確的天際線機率上限值。WorstSky 策略是先求出每個不確定物件的最差實例(worst instance)和最佳實例，若不確定物件的最佳實例或是不確定物件的實例，可以被任一其他不確定物件的最差實例支配(dominate)掉，則可確定此不確定物件或不確定物件實例的天際線機率則為零。利用 WorstSky 策略可將沒有機會成為天際線的物件刪除，並得到剩下未刪除的物件的天際線機率上限值。而 BestSky 策略是為了求出更精確的上限值，得到更佳的刪除能力。其方法為將每個不確定物件的最佳實例求出，再將這些最佳實例依照是否被其他最佳實例支配，分成許多階層，位在第一階層的最佳實例即為沒有被任何其他最佳實例所支配，位於第二階層的最佳實例是被第一層的最佳實例所支配，依此類推。當挑選第 k 個天際線機率時，可先由第一階層的不確定物件中挑選，若第一階層物件的總數小於 k，此時再考慮第二階層的物件，依此推算，即可找出最佳的第 k 個天際線機率。

實驗結果

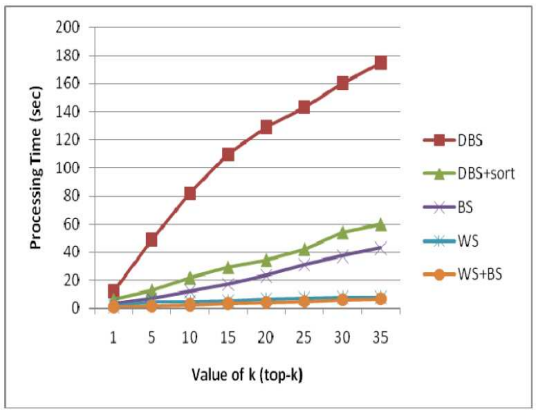
以 NBA 球員的數據統計資料為資料庫，經過篩選之後，共有 9,963 筆資料，每筆資料有 12 個屬性。相關參數初始設定為：屬性為 6 個，k 值為 10。由圖十可以看出當同時運用 WorstSky 與 BestSky 策略來刪除物件和找出更佳的天際線機率上限值是表現最佳，而採用隨機挑選第 k 個天際線機率上限值，其效果最差。由圖十一得知，隨著屬性個數的增加，處理時間也會隨之增加，同時運用 WorstSky 與 BestSky 策略仍可獲得最佳的效能，而 DBS+sort 和 BS 隨著屬性個數的增加，處理時間是不穩定成長，其原因為刪除能力是根據一直不斷地計算不確定物件的天際線機率而變動的關係。由圖十二得知隨著 k 值的增加，計算時間也隨之增加，效果仍是同時運用 WorstSky 與 BestSky 策略最佳。



圖十、動態上限值策略的比較



圖十一、屬性個數對處理時間的影響



圖十二、k 值的挑選對處理時間的影響

2、具不確定性的串流資料中機率天際線的連續查詢處理

研究目的

[PJL07]針對不確定性的資料，提出機率天際線(probabilistic skyline)的觀念。然而，機率天際線的相關研究仍侷限於靜態的不確定性資料上，但在即時應用環境中，資料以連續性且不間斷的方式進入系統，隨著日月累積，資料量越來越多，舊有的資料相對可參考的價值降低，因此使用者指定一個滑動視窗(sliding window)的長度 w，該視窗隨著時間保存最近 w 個時間單位的資料，且移除不在最近的 w 個時間單位內所產生的資料。如何在隨時間資料快速更新的情況下，計算出在滑動視窗內資料的機率天際線，此將是一困難點。

研究方法

在串流資料環境下，連續查詢機率天際線的基本方法為每當滑動視窗內的資料更新時，就執行一次機率天際線運算，即可求出答案，但是每次重新計算的成本是非常高的，且效率不佳。為了降低機率天際線的計算時間，必須避免計算每個不確定性物件精確的天際線機率，藉由計算物件的天際線機率之上限值(upper bound)與下限值(lower bound)，以利用上、下限值達到快速判斷不確定性物件是否為答案的效果。

第一，若不確定性物件的天際線機率上限值小於使用者設定的門檻值，則刪除此不確定物件；第二，若不確定性物件的天際線機率下限值大於或等於門檻值，此物件必為回傳結果之一，此兩種情況都不需要再去計算精確的天際線機率值，即可判斷不確定性物件是否為答案。若無法藉由上、下限值判斷時，則須更進一步計算，求得更精確的上下限值，直到真正計算出精確的天際線機率為止。

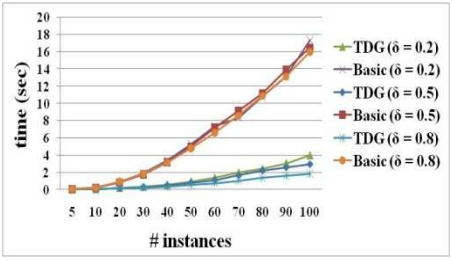
在資料串流的環境下，我們設計一資料結構，以儲存在滑動視窗內的所有資料，且為了加快計算的速度，該資料結構會對每個不確定性物件建立各自的最小堆疊(min heap)，且每個不確定性物件的最小堆疊內的所有資料是根據 SFS 方法[CG03]排序。該資料序列存在任一物件絕對不會支配排在它前面的物件之特性，利用該特性，以達到比較停止條件(stop condition)，減少物件比較次數與加快執行時間。其為 Basic 方法。

由於滑動視窗內的資料眾多，為了希望達到盡可能減少比較次數，快速判斷不確定性物件是否為答案的效果，故設計一時間支配圖(Time Dominant Graph - TDG)資料結構，藉由先比較一些資料，獲得天際線機率上、下限值，以達到刪除效果。TDG 記錄在目前滑動視窗中的天際線實例與目前不是但將來有機會成為天際線實例的實例。由於被記錄在 TDG 中的實例若是天際線實例，則此實例的天際線機率為 1，是具有最高的支配力，其具有較佳的刪除能力。因此計算每個不確定性物件的天際線上、下限機率時，可先與記錄在 TDG 的實例比較，計算出上、下限值，判斷物件是否為答案，若仍無法判斷，再採用 Basic 方法，取出實例和其他不確定性物件的實例一一比較。其為 TDG 方法。

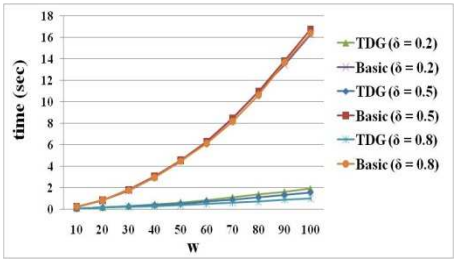
實驗結果

以 NBA 球員的數據統計資料為資料庫，共包含 19,112 筆資料，每筆資料有 10 個屬性。計算 100 次滑動視窗的天際線查詢處理時間，平均時間作為評估計算效能的指標。實驗參數預設值為：(1)單位時間點內的資料筆數：10，(2)滑動視窗長度：10 個時間單位，(3)屬性個數：5，(4)門檻值(δ)：分別為 0.2、0.5 及 0.8。

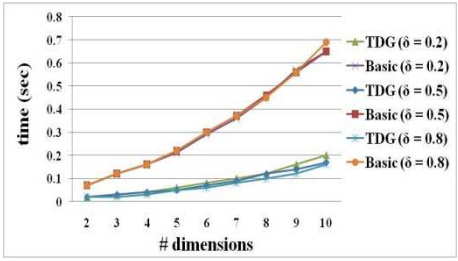
由圖十三得知，兩種方法均隨著資料筆數的增加，兩者的執行時間都是增加，但是在不同門檻值的設定下，TDG 方法的效能大幅優於 Basic 方法，在資料筆數增加的情況下，優越表現更佳顯著。從圖十四得知，隨著滑動視窗長度的增加，TDG 方法仍是優於 Basic 方法。由圖十五得知，在不同的屬性個數中，TDG 方法的平均查詢計算時間仍是優於 Basic 方法。



圖十三、不同的資料筆數對處理時間的影響



圖十四、滑動視窗長度對處理時間的影響



圖十五、物件屬性的個數對處理時間的影響

3、考慮動態天際線之範圍查詢處理

研究目的

[BK01]提出天際線查詢的觀念，天際線是由至少在一屬性中其值是最佳值的物件所組成；然而，[PT05]認為在真實應用中，使用者希望能以自身的需求當作查詢的條件，且回傳的答案必須盡可能接近該需求，因而提出一動態天際線查詢(dynamic skyline query)的觀念。

針對動態天際線的查詢，有些研究提出一些有效率的方法以縮減動態天際線查詢的處理時間。然而，那些研究均侷限於使用者自身的需求必須是精確的，但是在實際生活中，使用者通常不會很精確地知道自己的需求，而是希望給定一個具範圍的查詢(range query)。因此，我們的研究著重於設計一有效率演算法，根據使用者給定的一具範圍的查詢，在無須轉換每一物件與查詢本身的關係之情況下，能夠快速地回傳推薦物件給使用者。

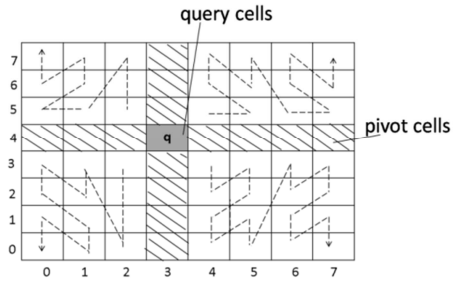
研究方法

當使用者給定一具範圍的查詢時，為了快速地回傳推薦物件給使用者，我們利用格子索引(grid index)及多方向的 Z 順序曲線(multidirectional Z-order curve)的特性，建立一資料結構，以有效地判別哪些物件絕對不屬於動態天際線且刪除該些物件，以節省更多不必要的計算，進而加快回傳動態天際線的處理時間。

物件的每一屬性可以當作一座標軸，而我們可依據物件本身的屬性值，將其視為在空間中的一點(point)。接著，我們將空間切割成數個沒有重疊的格子(cells)。當使用者給定一具範

圍的查詢時，該查詢所涵蓋的範圍可能同時橫跨數個格子，而這些格子稱為查詢格子(query cells)。根據查詢格子，其他格子的座標可以做一個轉換，使其符合 Z 順序曲線(Z-order curve)的特性，接著，我們將轉換過後的座標做一排序，得到一個曲線，其稱為多方向的 Z 順序曲線。以圖十六為例，使用者給定的範圍查詢為灰色格子，故該灰色格子為查詢格子，而在該格子的左上方、右上方、左下方及右下方的格子可以依據 Z 順序曲線的特性，決定各格子的執行順序，也就是虛線所標示的順序。

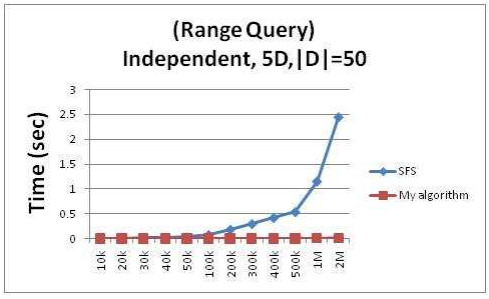
在尋找動態天際線的過程中，我們採用三種刪除策略，以加快其處理時間：(1)依據查詢格子的位置，決定出標竿格子(pivot cells)，如圖四斜線的部分，標竿格子是那些只有在一個維度不等於但在其餘維度都等於查詢格子的格子。若有個物件落在某個標竿格子內，則我們可以利用該物件刪除那些距離比標竿格子還遠的物件。(2)依據多方向的 Z 順序曲線決定格子存取的順序，保證若物件 A 比物件 B 先存取，那麼物件 B 必不會支配物件 A。(3)根據查詢格子轉換過後的座標，與查詢範圍最近的那一個物件，稱之為最好的點(best point)，我們將每個格子內最好的點做比較，若最好的點都已經被其他格子內最好的點所支配，那麼該格子不可能有物件會屬於動態天際線。



圖十六、格子索引和多方向的 Z 順序曲線

實驗結果

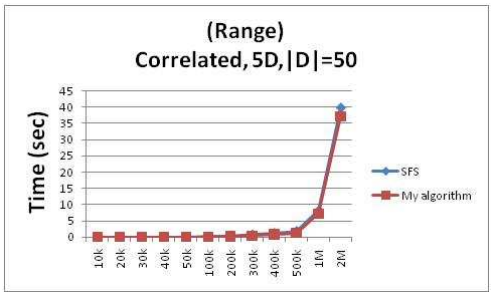
我們使用[BK01]所提供的二種不同資料分佈的合成資料庫(synthetic databases)：獨立分佈(independent distribution)及相依分佈(correlated distribution)作為實驗對象。SFS 演算法[CG03]為本研究的比較對象，因其為一靜態天際線查詢的演算法，每一物件必須根據使用者給定的範圍查詢做一個座標轉換，接著利用該演算法計算出天際線。



圖十七、獨立性資料分佈

圖十七及圖十八分別採用物件呈現獨立分佈及相依分佈的資料庫，其中 5D 代表每個物件都有 5 個屬性，|D|=50 代表每個屬性的最大值為 50。由圖五得知，我們的方法所需要的處理時間明顯的比 SFS 演算法少。其原因為當物件個數愈多時，SFS 演算法必須花愈多的時間去求得每一物件在經由座標轉換處理後的位置，然而，我們的方法無須做這樣的轉換處理，

因此較不會因物件個數的多寡而影響處理時間。從圖十八可以發現，當物件本身是呈現相依性資料分佈時，我們的方法與 SFS 演算法均會因物件個數的多寡而對動態天際線的處理時間有顯著的影響。不過，我們的方法仍比 SFS 演算法有效率。



圖十八、相依性資料分佈

4、考量反向式天際線之 Top-k 查詢處理

研究目的

[PT05]提出動態天際線查詢(dynamic skyline query)的觀念，考慮單一使用者，動態天際線是由至少在一屬性中其值是最符合此使用者需求的產品所組成；然而，[LC08]站在公司角度考慮，公司更希望能夠知道在動態天際線的觀念下回傳有哪些使用者把指定產品當作動態天際線，因而提出一反向式天際線查詢(reverse skyline query)的觀念。並且提出一演算法在不需要計算出每個單一使用者的動態天際線有哪些產品的情況下仍然可以找到正確的答案。

我們的推薦系統站在公司的角度去考慮，我們不但希望可以找出有哪些使用者視產品為動態天際線，我們更希望能夠找出前 top-k 個產品具有最多使用者把此產品視為天際線的性質，我們希望推薦這 k 個產品使我們有最好的銷售率。[WT09]的 BRS 演算法除了利用[LC08]的演算法且更進一步利用 R-tree 的索引結構增加判斷的策略減少了計算反向式天際線的處理時間。然而，對於我們想找出前 k 個有最多人視產品為動態天際線的目的之下，BRS 必須針對每個產品套用 BRS 取得所有產品的相對應的反向式天際線。因此，我們的研究著重於設計一有效率演算法使我們更快的找出此 k 個產品，我們利用了 R-tree 的結構使我們得以判斷一群點的反向式天際線個數之上限，使我們可以依此上限盡量避免去計算每個產品的反式天際線。最終，我們仍然可以得到正確的前 k 個最受歡迎產品。

研究方法

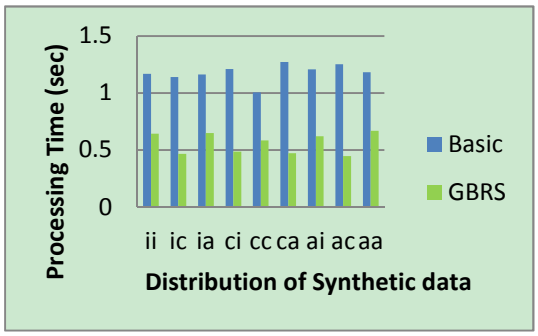
先將使用者及產品分別用 R-tree 做索引，以方便之後利用兩個 R-tree 中的 MBR 來取得反向式天際線(reverse skyline)數量的上下限估計值。

在考慮和其他產品競爭下的產品 R-tree 中的一個 MBR，排除不可能將此產品的 MBR 底下產品視為動態天際線查詢(dynamic skyline query)的使用者後，可以算出一個 MBR 的上限值，此上限值表示在此產品的 MBR 中的任一產品都不可能超過此反向式天際線數量。之後利用產品 MBR 的反向式天際線上限值作為判斷我們要推薦產品答案在哪一個產品 MBR 底下。

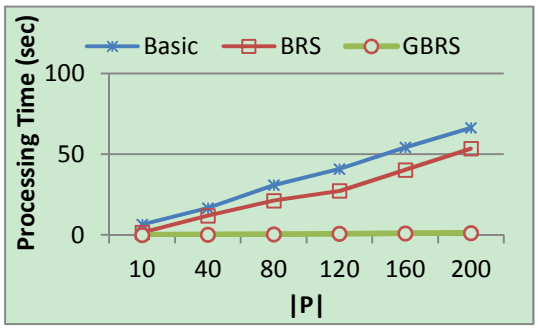
從產品 R-tree 的根部(root)開始尋找，每次都優先選擇擁有最高的反向式天際線上限值的 MBR 尋找，直至找到前 k 個具有最高反向式天際線數量的產品才完成動作。除此之外，我們在方法中利用[WT09]的引理，雖然犧牲上限值的準確度但提昇了計算的速度，更運用已計算的資訊去加快執行的時間。

實驗結果

考慮了 Basic Method 以及 BRS[WT09]來比較我們所提出的 GBRS 演算法，圖十九我們可以看出無論是哪種人造資料集都有很好的效果。因為 BRS 在人造資料太差不列出。圖二十則是實驗了真實的資料及我們可以看出 GBRS 的效果非常好。



圖十九、人造資料比較結果



圖二十、真實資料比較結果

三、成果自評

本計畫之研究成果包含相關研究論文九篇，其中五篇已公開發表或被接受於國際知名會議，其餘四篇也已投稿靜待審查。參與計畫之學生得以參加會議並報告研究成果，與國際知名學者交流，對於其研究能量累積亦有莫大幫助。

已接受或發表之論文

[LKA10] C. Y. Lin, J. L. Koh, and A. L. P. Chen, "A Better Strategy of Discovering Link-Pattern Based Communities by Classical Clustering Methods," the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2010)

[SWC10] H. Z. Su, E. T. Wang, and A. L. P. Chen, "Continuous Probabilistic Skyline Queries over Uncertain Data Streams," the 21st International Conference on Database and Expert Systems Applications (DEXA 2010)

[TTC12] M. F. Tsai, C. W. Tzeng, and A. L. P. Chen, "Discovering leaders from social network by action cascade," EuroSys 2012 5th Workshop on Social Network Systems (SNS2012)

[WWC11] W. C. Wang, E. T. Wang, and A. L. P. Chen, "Dynamic Skylines Considering Range Queries," the 16th International Conference on Database Systems for Advanced Applications (DASFAA2011)

[YWC12] T. A. Yeh, E. T. Wang, and A. L. P. Chen, "Finding Leaders with Maximum Spread of Influence through Social Networks" accepted by International Computer Symposium (ICS2012)

已投稿之論文

[CC] D. Y. Chiu, and A. L. P. Chen, "An Online Classifier for Enhancing the Accuracy of Multimedia Data Retrieval," Submitted for publication.
[HC] M. W. Huang, D. Y. Chiu, and A. L. P. Chen, "Efficient Computation of Subspace Top-K Probabilistic Skylines on Uncertain Data," Submitted for publication.
[LC] Y. W. Lin and A. L. P. Chen, "Skyline Query Processing Considering Nearest and Farthest Neighbors," Submitted for publication.
[WC] P. H. Wu and A. L. P. Chen "Top-k Query Processing Considering Reverse Skyline Retrieval," Submitted for publication.

參考文獻

[B98] C. J. C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition,” Data Mining and Knowledge Discovery, 1998.

[BF96] J. Barros, J. French, W. Martin, P. Kelly, and M. Cannon, “Using the Triangle Inequality to Reduce the Number of Comparisons Required for Similarity-based Retrieval,” International Conference on Storage and Retrieval for Image and Video Databases, 1996.

[BK01] S. Börzsönyi, D. Kossmann, and K. Stocker, “The skyline operator,” International Conference on Data Engineering, 2001.

[BS97] A. Berman and L. Shapiro, “Efficient Image Retrieval with Multiple Distance Measures,” Interna-tional Conference on Storage and Retrieval for Image and Video Databases, 1997.

[CG03] J. Chomicki, P. Godfrey, J. Gryz, and D. Liang, “Skyline with presorting,” International Conference on Data Engineering, 2003.

[CL08] L. Chen and X. Lian, “Dynamic Skyline Que-ries in Metric Spaces,” Advances in Database Tech-nology, 2008.

[CS93] P. K. Chan, M. D. F. Schlag, and J. Y. Zien, “Spectral K-Way Ratio-Cut Partitioning and Cluster-ing,” Design Automation Conference, 1993.

[CV95] C. Cortes and V. Vapnik, “Support-Vector Networks,” Machine Learning, 20(3), 1995.

[CW10] W. Chen, C. Wang, and Y. Wang, “Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks,” the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2010.

[DH01] C. H. Q. Ding, X. He, H. Zha, M. Gu, and H. D. Simon, “A Min-Max Cut Algorithm for Graph Parti-tioning and Data Clustering,” International Conference on Data Mining, 2001.

[DS93] A. Drug, W.V. Stoecker, J.P. Cookson, S.E. Umbaugh, and R.H. Moss, “Identification of Variegat-ed Coloring in Skin Tumors: Neural Network vs. Rule-based Induction Methods,” IEEE Engineering in Medicine and Biology Magazine, 1993.

[FL00] G. Flake, S. Lawrence, and C. Giles, “Efficient Identification of Web Communities,” International Conference on Knowledge Discovery and Data Mining , 2000.

[GS05] P. Godfrey, R. Shipley, and J. Gryz, “Maximal Vector Computation in Large Data Sets,” International Conference on Very Large Data Bases, 2005.

[IK05] H. Ino, M. Kudo, and A. Nakamura, “Partition-ing of Web Graphs by Community Topology,” Interna-tional Conference on World Wide Web, 2005.

[KK03] D. Kempe, J. Kleinberg and É. Tardos, “ Maximizing the Spread of Influence through a Social Network,” the

9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2003.

[KM91] J. K. Kruschke and J. R. Movellan, “Benefits of Gain: Speed Learning and Minimal Hidden Layers in Back-propagation networks,” IEEE Transaction on systems, Man and Cybernetics, 1991.

[KR99] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, “Trawling the Web for Emerging Cyber-Communities,” Journal of Computer Networks, 1999.

[KR02] D. Kossmann, F. Ramsak, and S. Rost, “Shoot-ing Starts in the Sky: An Online Algorithm for Skyline Queries,” International Conference on Very Large Data Bases, 2002.

[LC08] X. Lian and Lei. Chen. Monochromatic and bichromatic reverse skyline search over uncertain databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD2008, Vancouver, BC, Canada, pp. 213-226.

[LK07] J. Leskovec, A. Krause, C. Guestrin, C. Fa-loutsos, J. VanBriesen, and N. Glance, “Cost-effective Outbreak Detection in Networks,” the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2007.

[LS08] J. J. Li, S. L. Sun, and Y. Y. Zhu, “Efficient Maintaining of Skyline over Probabilistic Data Stream,” International Conference on Natural Compu-tation, 2008.

[LW07] B. Long, X. Wu, Z. M. Zhang, and P. S. Yu, “Community Learning by Graph Approximation,” In-ternational Conference on Data Mining, 2007.

[LY05] X. Lin, Y. Yuan, W. Wang, and H. Lu, “Stab-bing the Sky: Efficient Skyline Computation over Slid-ing Windows,” International Conference on Data En-gineering, 2005.

[LZ07] K. C. K. Lee, B. Zheng, H. Li, and W. C. Lee, “Approaching the Skyline in Z Order,” International Conference on Very Large Data Bases, 2007.

[M97] T.M. Mitchell, “Artificial Neural Networks,” Machine Learning, 1997.

[MA03] X. Mu, M. Artiklar, M.H. Hassoun, and P. Watta, “An RCE based Associative Memory with Ap-plication to Human Face Recognition,” INNS-IEEE In-ternational Joint Conference on Neural Networks, 2003.

[PF07] J. Pei, A. W. C. Fu, X. Lin, and H. Wang, “Computing Compressed Multidimensional Skyline Cubes Efficiently,” International Conference on Data Engineering, 2007.

[PJE05] J. Pei, W. Jin, M. Ester, and Y. Tao, “Catching the Best Views of Skyline: A Semantic Approach based on Decisive Subspaces,” International Conference on Very Large Data Bases, 2005.

[PJL07] J. Pei, B. Jiang, X. Lin, and Y. Yuan, “Probabil-istic Skylines on Uncertain Data,” International Con-ference on Very Large Data Bases, 2007.

[PT05] D. Papadias, Y. Tao, G. Fu, and B. Seeger, “Progressive Skyline Computation in database sys-tems,” ACM Transactions on Database Systems, 2005.

[RK01] P. Reddy and M. Kitsuregawa, “Inferring Web Communities through Relaxed Cocitation and Dense Bipartite Graphs,” International Conference on Web Information Systems Engineering, 2001.

[RY98] V. Rajan, J. Ying, S. Chakrabarty, and K. Pat-tipati, “Machine Learning Algorithms for Large Multi-dimensional Dynamic Indexes,” IEEE Conference on Systems, 1998.

[SB08] D. Sacharidis, P. Bouros, and T. Sellis, “Caching Dynamic Skyline Queries,” Statistical and Scientific Database Management, 2008.

[SM00] J. Shi and J. Malik, “Normalized Cuts and Im-age Segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000.

[SS06] M. Sharifzadeh and C. Shahabi, “The spatial skyline queries,”International Conference on Very Large Databases,

2006.

[TP06] Y. Tao and D. Papadias, “Maintaining Sliding Window Skylines on Data Streams,” IEEE Transac-tions on Knowledge and Data Engineering, 2006.

[TX06] Y. Tao, X. Xiao, and J. Pei. “SUBSKY: Efficient Computation of Skylines in Subspaces,” International Conference on Data Engineering, 2006.

[WI01] B. M. Wilamowski, S. Iplikci, O. Kaynak, and M. Onder Efe, “An Algorithm for Fast Convergence in Training Neural,” INNS-IEEE international Joint Con-ference on Neural Networks, 2001.

[WT09] X. Wu, Y. Tao, R. C. W. Wong, L. Ding and J. X. Yu. Finding the influence set through skylines. In Proceedings of the 12nd International Conference on Extending Database Technology, EDBT2009, Saint Petersburg, Russia, pp. 1030-1041.

[YO01] C. Yu, B. C. Ooi, K. L. Tan, and H. V. Jagadish, “Indexing the Distance: An Efficient Method to KNN Processing,” International Conf. on Very Large Data Bases, 2001.

[ZL09] W. Zhang, X. Lin, Y. Zhang, W. Wang, and J. X. Yu. “Probabilistic Skyline Operator over Sliding Windows” International Conference on Data Engi-neering, 2009.

報告人姓名	陳良弼	服務機構及職稱	國立政治大學講座教授兼理學院院長
會議/訪問時間 地點	4/1~5/2012 Washington DC, USA		
會議名稱	The 28th IEEE International Conference on Data Engineering		

我於 3/31 前往美國 Washington DC 參加第 28 屆 IEEE International Conference on Data Engineering。這個會議共五天，除了 4/2-4/4 三天為大會外，4/1 與 4/5 各為一天的 workshops。4/1 的 workshops 包括 Data Driven Decision Guidance and Support Systems、Self-Managing Database Systems、Spatial Temporal Data Integration and Retrieval、以及 Data Engineering Meets the Semantic Web，晚上則為 Conference Reception。

會議第一天的 keynote speech 由 INRIA 的 Serge Abiteboul 所講的 Viewing the Web as a distributed knowledge base。Serge 是我在南加大讀博士時的學長，研究領域在 database theory，也一直活躍在資料庫研究中。首先他提出一個問題：Alice 與 David 即將結婚，他們的朋友想幫他們準備一本 album of photos；那麼，從那麼多異質性的資料來源裡，他們要如何進行這件事呢？

Serge的做法是把Web當成一個distributed knowledge base，把各式資料都當成 logical facts及rules，然後在這個distributed knowledge base裡做reasoning以找出答案。Serge講解了他們延伸Datalog所設計的Webdamlog的各種features，並以例子逐項說明。

之後我參加了Web 2.0 Applications的討論。第一篇論文GeoFeed: A Location-Aware News Feed同時考慮user location(spatial relevance)及friend list(social relevance)來推薦新聞。第三篇論文CI-Rank: Ranking Keyword Search Results Based on Collective Importance藉著users所下的keywords之間其他keywords所展示的重要性來排序搜尋結果，非常有趣。第四篇論文Temporal Analytics on Big Data for Web Advertising獲得大會最佳論文獎。這篇論文的作者來自Microsoft與IBM，以Monitor、Mine、與 Manage的方式對巨量資料做時序分析。

下午我參加了Spatio-Temporal Data Management的session。共有SWST: A Disk Based Index for Sliding Window Spatio-Temporal Data、Querying Uncertain Spatio-Temporal Data、The Min-dist Location Selection Query、Bi-level Locality Sensitive Hashing for K-Nearest Neighbor Computation等四篇論文發表。其中第三篇論文與我目前Reverse KNN的研究具有相似目的，值得參考。

會議第二天的keynote speech為Microsoft的Surajit Chaudhuri所講的How different is big data? Big data與cloud computing被視為目前資訊科技的兩大挑戰，Surajit提到過去在relational database裡對big data相關議題的探討以及其不足，還有現今cloud computing以及data rich web services所帶來的新趨勢的影響。

之後我主持了Industrial Session 1: Support for Large Scale Data Analytics。這個session參加者眾，並對三篇論文的發表討論熱烈。第一篇論文Exploiting Common Subexpressions for Cloud Query Processing來自Microsoft與Arizona State

University，提出先處理subexpression的最小子集以避免重複計算的想法。第二篇論文Vectorwise: a Vectorized Analytical DBMS為學界研究推廣為商品的一例，其performance比現今產品都要好，引起來自Microsoft、IBM、HP等業界研究員的多方質詢。第三篇論文Scalable and Numerically Stable Descriptive Statistics in SystemML來自IBM，著眼於數值計算中小誤差的不能忽視及其作法。

午餐時安排的座談會為Funders Session，由National Science Foundation的Le Gruenwald、Department of Energy的Ceren Sust、及National Institutes of Health的Olga Brazhnik報告其單位在資料工程方面的研究經費機會；其中最好的消息應該是NSF剛由白宮通過一個Big Data的新研究議題。雖然對美國學界來講此座談會堪稱重要，但由聽眾稀少的情形看來，參加會議的應該大部分來自美國境外或者是美國的業界(大會統計，此次會議論文被接受最多的三大單位為Hong Kong、Singapore、及Microsoft)。

下午參加了Top-K Processing以及poster sessions。這個會議延續去年方式，所有的論文作者都要再以poster來說明他們的論文內涵，這樣的作法可讓與會者對所有有興趣的論文與其作者討論。晚上參加cruise與晚宴。由於4/6有事必須回到台灣，我搭乘4/4的航班，於4/5晚回到台灣，結束此行。以下附上會議時的一些觀察與心得：

- Met senior database researchers Erich Neuhold, David Lomet, Mike Stonebraker, Philip Berstein, Umesh Dayal, Paul Larson, David Maier, Masaru Kitsuregawa, and Guy Lohman
  - They persistently attended ICDE
  - Interesting to see a gap between senior and young researchers
  - The age of the attendees ranges from 20+ to 70+
- Talked to Lei Chen (for a visit), Hong Cheng, Reynold Cheng, Xueming Lin, Li-Zhu Zhou, Kyuseok Shim, Wei-Shinn Ku (about jobs in Taiwan), Yu Zheng (about internship and collaboration at MSRA)
- Collaboration becomes popular (from 千人計畫, MSRA internship, etc.)
- Talked to Mi-Yen about AS; can freely leave for visit for some period of time (visited MSRA for three months, and will visit Jian Pei for two months)
- Chaired industrial session; Paul, Guy, Masaru, Xiaofang, Qiming, Kyu-Young were all there
- Talked to Eric Lo (now at The Hong Kong Polytechnic University)
- Talked to Elisa Bertino; lots of activities and travels for the data security research
- Talked to Umesh long about current database research activities at HP Labs
- Talked to Sean; 千青計畫 (under 35 years old)
- Talked to Kyuseok about SNU and KAIST

此次會議從台灣來參加的另有中研院資訊所兩位成員。

攜回大會論文集。

國科會補助計畫衍生研發成果推廣資料表

日期:2012/10/27

國科會補助計畫	計畫名稱: 合作式個人化推薦系統之進階技術研究及其應用
	計畫主持人: 陳良弼
	計畫編號: 98-2221-E-004-005-MY3　　學門領域: 資料庫系統及資料工程

無研發成果推廣資料

98 年度專題研究計畫研究成果彙整表

計畫主持人：陳良弼	計畫編號：98-2221-E-004-005-MY3
計畫名稱：合作式個人化推薦系統之進階技術研究及其應用	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	7	7	100%	人次	
		博士生	2	2	100%		
		博士後研究員	1	1	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	5	5	100%	篇	[LKA10] C. Y. Lin, J. L. Koh, and A. L. P. Chen, ’’’’’’’’’’ A Better Strategy of Discovering Link-Pattern Based Communities by Classical Clustering Methods,’’’’’’’’’’ the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2010) [SWC10] H. Z. Su, E. T. Wang, and A. L. P. Chen, ’’’’’’’’’’ Continuous Probabilistic Skyline Queries over Uncertain Data Streams,’’’’’’’’’’ the 21st

							action cascade,’’’’’’’’’ EuroSys 2012 5th Workshop on Social Network Systems (SNS2012) [WWC11] W. C. Wang, E. T. Wang, and A. L. P. Chen, ’’’’’’’’’ Dynamic Skylines Considering Range Queries,’’’’’’’’’ the 16th International Conference on Database Systems for Advanced Applications (DASFAA2011) [YWC12] T. A. Yeh, E. T. Wang, and A. L. P. Chen, ’’’’’’’’’ Finding Leaders with Maximum Spread of Influence through Social Networks’’’’’’’’’ accepted by International Computer Symposium (ICS2012)
		研究報告/技術報告	4	4	100%		[CC] D. Y. Chiu, and A. L. P. Chen, ’’’’’’’’’ An Online Classifier for Enhancing the Accuracy of Multimedia Data Retrieval,’’’’’’’’’ Submitted for publication. [HC] M. W. Huang, D. Y. Chiu, and A. L. P. Chen, ’’’’’’’’’ Efficient Computation of Subspace Top-K Probabilistic Skylines on Uncertain Data,’’’’’’’’’ Submitted for publication. [LC] Y. W. Lin and A. L. P. Chen, ’’’’’’’’’ Skyline Query Processing Considering Nearest and Farthest Neighbors,’’’’’’’’’ Submitted for publication. [WC] P. H. Wu and A. L. P. Chen ’’’’’’’’’ Top-k Query Processing Considering Reverse Skyline Retrieval,’’’’’’’’’ Submitted for publication.
		研討會論文	0	0	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	

		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
其他成果(無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)		范耀中同學榮獲台灣電機電子工程學會（TIEEE）最佳博士論文優等獎（指導教授：陳良弼教授） 蘇映晨同學榮獲中國電機工程學會 99 青年論文獎第三名（指導教授：陳良弼教授）					

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：■已發表 □未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：(以 100 字為限)

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）(以 500 字為限)

伴隨著資訊科技之發展，資訊缺乏已不再是問題，我們想要瀏覽的任何東西都變得捶手可得。在這樣的環境中，使用者面臨到的問題是如何在龐大的資料庫中快速地找到感興趣的，因為我們絕對無法瀏覽所有的物件。因此本研究計畫為未來合作式推薦系統，著眼於這類新興應用環境的成長與需求，為了提供更有效率、更有品質的推薦結果，也就因為合作式推薦的應用範圍廣泛，而且可以有效地縮減盲目搜尋的時間，所以不論從學術研究或商業發展的觀點，都是值得投入人力深耕的研究領域。而伴隨著本應用系統下所研發的資料管理技術，也將成為未來推薦系統相關資料分析與應用的基石，進而提昇推薦服務的效率與品質，也引領國內外相關研究朝更多元化之智慧型推薦系統應用的方向前進。因此本計畫之研究成果，對於提升我國在相關領域的技術水準與研究地位，將可達成影響深遠的正面貢獻。而本計畫之研究成果包含相關研究論文九篇，其中五篇已公開發表或被接受於國際知名會議，其餘四篇也已投稿靜待審查。參與計畫之學生得以參加會議並報告研究成果，與國際知名學者交流，對於其研究能量累積亦有莫大幫助。