出國報告（出國類別：國際會議）

2012 年 DaWaK
「資料倉儲與知識探勘」國際會議
出國報告

服務機關：行政院主計總處主計資訊處

姓名職稱：王設計師興娟

派赴國家：奧地利

出國期間：101/9/1~101/9/7

報告日期：101/10/30

摘　　要

DEXA（Database and Expert Systems Applications）協會是致力於資訊科技領域研究的非營利組織，每年於各國舉辦會議，其中「資料倉儲與知識探勘」（International Conference on Data Warehousing and Knowledge Discovery；簡稱 DaWaK）會議研討的相關議題與理論，已被企業組織廣泛認可為提升資料分析、決策支援、自動萃取數據知識等能力的關鍵技術，DaWaK 2012 希望可以在面對「資料倉儲」、「知識探勘」、「資料挖掘應用」，甚至是新興的「智能雲」等領域發展過程，能有創新的理論、方法論、演算法或是解決方案。

會議期間邀請各國學術專家、產業實務專家、研究員及學生，探討有關資料倉儲與知識探勘的發展與實際應用，分享彼此經驗及研究成果，並如何應用執行於產業上。

目　　次

壹、目的

為了汲取各國「資料倉儲」、「知識探勘」的研究新知與技術經驗，參加DEXA（Database and Expert Systems Applications）協會在奧地利維也納舉辦的第 23 屆年會 DEXA 2012，其中的第 14 屆「資料倉儲與知識探勘」（International Conference on Data Warehousing and Knowledge Discovery；簡稱 DaWaK）會議，是該年會的 8 個主題會議之一，會期自 9 月 3 日至 6 日為期 4 天。

希望藉由本次會議研討、經驗交流，可以學習並應用於統計調查及資料處理業務，以期提升資料品質，加強統計調查資料應用層面，進而提供有效決策支援資訊。

貳、過程

此次參加的 DaWaK 國際會議，係屬 DEXA 2012 年會的 8 個主題會議之一，由 DEXA 協會每年於各國定期舉辦，今年於 9 月 3 日至 6 日在奧地利維也納科技大學（Vienna University of Technology；簡稱 TU）舉行，是重要的國際型電腦資訊科技研討會之一。以下為會議的網址與舉辦地點「維也納科技大學」。

http://www.dexa.org/

DEXA 2012 年會之 8 個主題會議名稱詳列如下：

- DEXA '12 : 23rd International Conference on Database and Expert Systems Applications
- DaWaK '12 : 14th International Conference on Data Warehousing and Knowledge Discovery
- EC-Web '12 : 13th International Conference on Electronic Commerce and Web Technologies
- TrustBus '12 : 9th International Conference on Trust, Privacy, and Security in Digital Business
- Globe '12 : 5th International Conference on Data Management in Cloud, Grid and P2P Systems
- ITBAM '12 : 3rd International Conference on Information Technology in Bio- and Medical Informatics
- EGOVIS and EDEM '12 : Joint International Conference on Electronic Government, the Information Systems Perspective, and Electronic Democracy
- ICT-GLOW '12 : 2nd International Conference on ICT as Key Technology for the Fight against Global Warming

DEXA 2012 的 8 個主題會議，包含的領域有：資料庫與專家系統應用(DEXA ’12)、資料倉儲與知識探勘(DaWaK ’12)、電子商務與 Web 技術(EC-Web ’12)、數位商務的信任-隱私-安全(TrustBus ’12)、雲端-網格-點對點的資料管理(Globe ’12)、生物與醫學的資訊技術應用(ITBAM ’12)、以資訊系統角度結合電子化政府和電子化民主(EGOVIS and EDEM '12)、以資訊和通訊技術作為對抗全球暖化的關鍵技術(ICT-GLOW '12)等。資料倉儲與知識探勘(DaWaK ’12)會議在這次會期選取了 36 篇論文發表，錄取率約為 32%，所發表的文章最後彙整成冊出版(ISBN：978-3-642-32583-0)。

本屆「資料倉儲與知識探勘」(DaWaK '12)會議，所發表的場次及主題如下：

時　間	主　　題	主持人
第一天 **2012.09.03** [10:30-12:00]	第一場次：資料倉儲設計方法論	Alberto Abello
	◆ **BPMN-Based Conceptual Modeling of ETL Processes** Zineb El Akkaoui, José-Norberto Mazón, Alejandro Vaisman, Esteban Zimányi ◆ **Enhancing Coverage and Expressive Power of Spatial Data Warehousing Modeling: The SDWM Approach** Alfredo Cuzzocrea, Robson N. Fidalgo ◆ **Sprint Planning Optimization in Agile Data Warehouse Design** Matteo Golfarelli, Stefano Rizzi, Elisa Turricchia	
2012.09.03 [13:30-15:00]	第二場次：ETL & DW 方法論及工具	Robson Fidalgo
	◆ **A Case Study on Model-Driven Data Warehouse Development** Thomas Benker, Carsten Jürck ◆ **A Lightweight Stream-based Join with Limited Resource Consumption** M. Asif Naeem, Gillian Dobbie, Gerald Weber ◆ **Integrating ETL Processes from Information Requirements** Petar Jovanovic, Oscar Romero, Alkis Simitsis, Alberto Abelló	
2012.09.03 [15:30-17:00]	第三場次：多維度資料處理和管理	Alfredo Cuzzocrea
	◆ **Automatic Transformation of Multi-Dimensional**	

時　間	主　　題	主持人
	Web Table into Data Cubes Norah Alrayes, Wo-Shun Luk **Differentiated Multiple Aggregations in Multidimensional Databases** Ali Hassan, Frank Ravat, Olivier Teste, Ronan Tournier, Gilles Zurfluh **Genetic Algorithms-based Symbolic Aggregate Approximation** Muhammad Marwan Muhammad Fuad	
第二天 **2012.09.04** [09:00-10:30]	第四場次：資料倉儲與 OLAP 擴展	Pedro Furtado
	A Neural-Based Approach for Extending OLAP to Prediction Wiem Abdelbaki, Riadh Ben Messaoud, Sadok Ben Yahia **Mobile Activity Recognition using Ubiquitous Data Stream Mining** João Bártolo Gomes, Shonali Krishnaswamy, Mohamed Gaber, Pedro Sousa, Ernestina Menasalvas **Warehousing Manufacturing Data. A Holistic Process Warehouse for Advanced Manufacturing Analytics** Christoph Gröger, Johannes Schlaudraff, Florian Niedermann, Bernhard Mitschang	
2012.09.04 [13:30-15:00]	第五場次：資料倉儲的效能和最佳化	Alfredo Cuzzocrea
	Queen-Bee: Query Interaction-Aware for Buffer Allocation and Scheduling Problem Amira Kerkad, Ladjel Bellatreche, Dominique Geniet **Efficient Distributed Parallel Top-Down Computation of ROLAP Data Cube Using MapReduce** Suan Lee, Jinho Kim, Yang-Sae Moon, Wookey Lee **Landmark-Join: Hash-Join based String Similarity**	

時　間	主　　題	主持人
	Joins with Edit Distance Constraints Kazuyo Narita, Shinji Nakadai, Takuya Araki	
2012.09.04 [15:30-17:00]	第六場次：資料挖掘和知識探勘的技術	Qiming Chen
	⊕ **A Fast Algorithm for Frequent Itemset Mining Using Patricia* Structures** Jun-Feng Qu, Mengchi Liu ⊕ **Incremental Itemset Mining Based on Matrix Apriori Algorithm** Damla Oguz, Belgin Ergenc ⊕ **Multi-objective Optimization for Incremental Decision Tree Learning** Hang Yang, Simon Fong, Yain-Whar Si	
第三天 **2012.09.05** [09:00-10:30]	第七場次：資料挖掘和知識探勘的應用(1)	Wo-Shun Luk
	⊕ **Mining Contextual Preference Rules for Building User Profiles** Sandra De Amo, Mouhamadou Saliou Diallo, Cheikh Talibouya Diop, Arnaud Giacometti, Haoyuan D. Li, Arnaud Soulet ⊕ **Polarities, Axiallities, and Marketability of Item Sets** Dan Simovici, Paul Fomenky, Werner Kunz ⊕ **RssE-Miner: A new Approach For Efficient Events Mining from Social Media RSS Feeds** Nabila Dhahri, Chiraz Trabelsi, Sadok Ben Yahia	
2012.09.05 [13:30-15:00]	第八場次：特徵挖掘	Mohamed Gaber
	⊕ **K Nearest Neighbor using Ensemble Clustering** Loai AbedAllah, Ilan Shimshoni ⊕ **Collocation Pattern Mining in Limited Memory Environment Using Materialized iCPI-tree**	

時　間	主　　題	主持人
	Pawel Boinski, Maciej Zakrzewicz **Mining Popular Patterns from Transactional Databases** Carson Leung, Syed Tanbeer	
2012.09.05 [15:30-17:30]	第九場次：資料串流挖掘	Osmar Zaiane
	Rare Pattern Mining on Data Streams David Huang, Yun Sing Koh, Gillian Dobbie **A Single Pass Trellis-based Algorithm for Clustering Evolving Data Streams** Simon Malinowski, Ricardo Morla **New Management Operations on Classifiers Pool to Track Recurring Concepts** Mohammad Javad Hosseini, Zahra Ahmadi, Hamid Beigy **Extrapolation Prefix Tree for Data Stream Mining using a Landmark Model** Yun Sing Koh, Russel Pears, Gillian Dobbie	
第四天 **2012.09.06** [09:00-10:30]	第十場次：資料挖掘和知識探勘的應用(2)	Russels Pears
	Dynamic Topography Information Landscapes – An Incremental Approach to Visual Knowledge Discovery Kamran Ali Ahmad Syed, Mark Kröll, Vedran Sabol, Arno Scharl, Stefan Gindl, Micheal Granitzer, Albert Weichselbraun **Classifying Websites into Non-topical Categories** Chaman Thapa, Osmar Zaiane, Davood Rafiei, Arya M. Sharma **Improving Cross-Document Knowledge Discovery Using Explicit Semantic Analysis** Peng Yan, Wei Jin	

時　間	主　　題	主持人
2012.09.06 [14:00-15:00]	第十一場次：資料倉儲的保密性與安全性	Ladjel Bellatreche
	⊕ **Implementing a Data Lineage Tracker** Colin Puri, Doo Soon Kim, Kunal Verma, Peter Yeh ⊕ **Evaluating the Feasibility Issues of Data Confidentiality Solutions from a Data Warehousing Perspective** Ricardo Jorge Santos, Jorge Bernardino, Marco Vieira	
2012.09.06 [15:30-17:00]	第十二場次：分散式處理概論與演算法	Carson Leung
	⊕ **A New Paradigm for Collaborating Distributed Query Engines** Qiming Chen, Meichun Hsu ⊕ **Using OCL for Automatically Producing Multidimensional Models and ETL Processes** Faten Atigui, Franck Ravat, Olivier Teste, Gilles Zurfluh ⊕ **MIRABEL DW: Managing Complex Energy Data in a Smart Grid** Laurynas Siksnys, Christian Thomsen, Torben Bach Pedersen	

參、會議內容摘錄

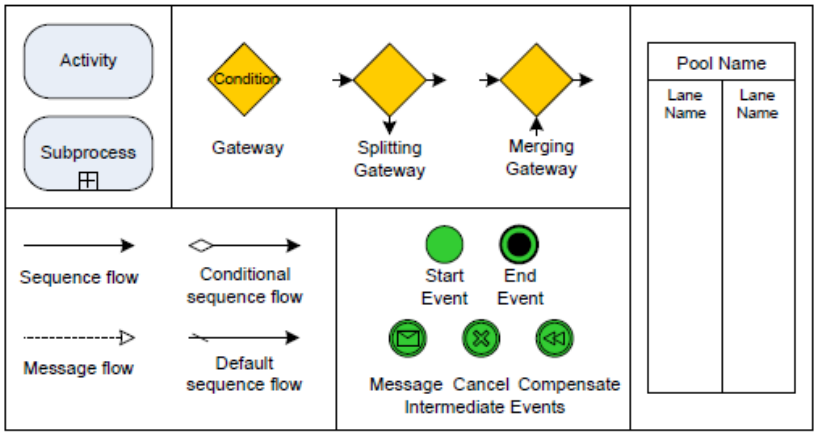
一. 以 BPMN 為基礎將 ETL 過程概念模組化(Conceptual Modeling)

商業智慧(Business Intelligence；簡稱 BI)是一系列商業活動行為的資料收集、定義、分析與資訊化技術，透過持續性的監測、分析、管理，作為組織決策支援的工具。BI 應用包含廣泛的分析能力，包括線上分析處理系統(On-Line Analytical Processing；簡稱 OLAP)和資料挖掘(Data Mining)工具等。在大多數情況下，BI 是應用整合在資料倉儲(Data Warehouse)裡的資料，所謂整合性資料指的是資料倉儲的流程，包含從原始資料抽取(Extraction)、轉換(Transformation)和裝載(Load)到資料倉儲，成為已淨化和轉換的資料，這樣的流程稱為 ETL(Extraction, Transformation, Load)過程。而 ETL 過程廣泛被認為發展複雜，容易出錯且非常耗時。

現今的商業動態需要靈活異動的 BI 資料，相當於對 ETL 過程發展產生新的挑戰，資料倉儲的資料必須具備即時性(Real-time)和時間正確性(Right-time)。為了應付這些一直改變的需求，需要敏捷和靈活的 ETL 工具，可以很快產生和修改可執行碼。同時，為了應對這些挑戰，提出另一種思考的觀點，建議 ETL 過程加入業務流程建模標記法(Business Process Model and Notation；簡稱 BPMN)，了解業務流程(Business Processes)，以便輕鬆辨識那些資料是需要的。

業務流程建模標記法(BPMN)是物件管理組織(Object Management Group；簡稱 OMG)維護的關於業務流程建模的行業性標準。它建立的業務流程圖(Business Process Diagram)非常類似程式語言的流程圖法(flowcharting)，希望透過一套符合業務人員直觀又能展現複雜流程語意的標記法，可以同時為業務人員從事的業務流程，管理監督業務流程的

經理人，以及資訊技術人員，充當不同專業領域的共同語言，便於溝通及提供支援，同時，BPMN 規範還提供從標記法圖形到執行語言的基礎構造映對，稱之為業務流程執行語言(Business Process Execution Language；簡稱 BPEL)。相同的，組織裡的其它流程也可以比照類似程序順利完成。下圖是 BPMN 主要使用的物件和標記。

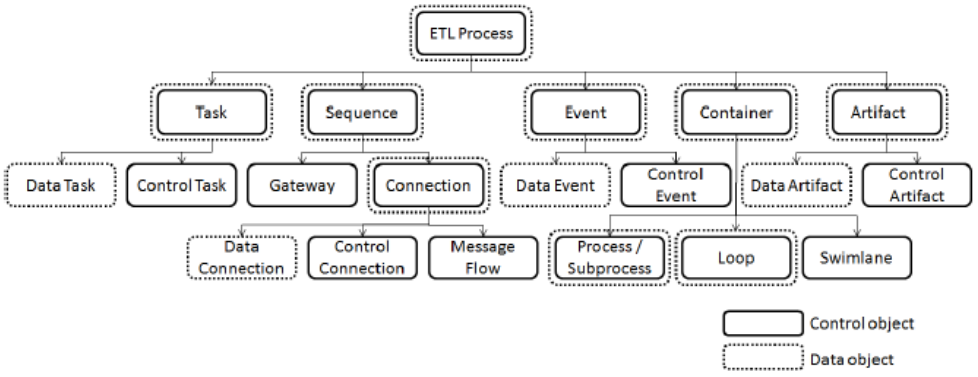


BPMN main objects and notation.

下圖是 ETL 過程模組化時以 BPMN 為基礎建成 ETL 分類樹，ETL 元件可能有下列五種要素：任務(Task)、順序(Sequence)、事件(Event)、容器(Container)和加工品(Artifact)，分類樹的元素可能是資料物件(Data object)、控制物件(Control object)或二者皆是。然後可以建立出二種不同的分類樹：

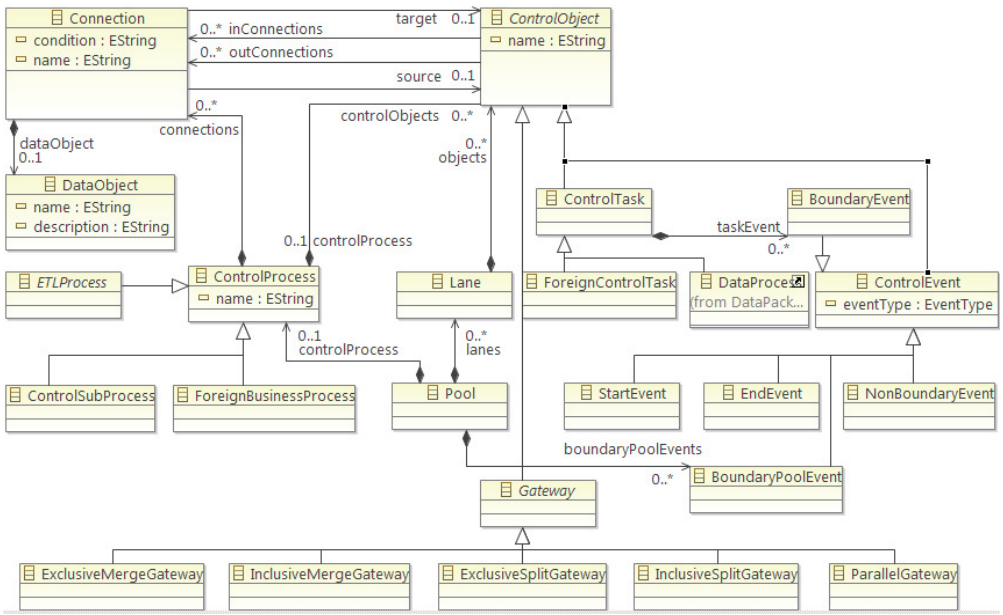
(一). ETL 控制分類樹(ETL control classification tree)。

(二). ETL 資料分類樹(ETL data classification tree)。



ETL classification tree

接續以 BPMN 2.0 的基礎模型(metamodel)為基礎，建立 ETL 過程的基礎模型，依據前述分類可產生控制基礎模型(Control metamodel)和資料基礎模型(Data metamodel)，則 ETL 過程的概念模組(Conceptual models)就可以產生，接續可直接對應成邏輯模組(logical model)，使用有模組轉換成文字功能(Model-to-Text transformations)的工具，就可以直接產生程式。OMG 標準要求重複使用性(reusability)、延展性(extensiblity)及協同工作(collaborative work)。當資料倉儲在生命週期期間發生需求改變時，不論是資料來源或是資料倉儲內部，也可應用這些模組，配合需求的改變順利調整完成 ETL 過程。



控制基礎模型(Control metamodel)

二. 隨處可得的資料流挖掘(Data Stream Mining)達成移動式行為識別(Mobile Activity Recognition)

隨著網際網路、通訊科技和感知器設備(Sensor Device)快速進步，例如：手機、平板電腦等，其計算能力提升與記憶體增加，形成無所不在的運算(Ubiquitous Computing)行為出現；行動使用者透過手機、平板電腦可以在任何時間、任何地點收發 e-mail，召開即時線上行動會議，彼此交換檔案資料等，因此資料成長速度每天大幅度的大量增加，以串流(Stream)的形式傳遞，稱之為資料串流(Data Stream)，它不同於傳統靜態資料庫的資料型態，有較多的資源限制，但具有時效性，資料若不立即進行分析，過了一段時間後就會失去其價值，例如透過手機即時接收股票市場資訊，進行股票各類分析等。資料串流具有下列六項特質(Ho, Li, Kuo and Lee, 2006)：

(一). 輸入的資料是無窮。

(二). 資料持續快速到達。

(三). 主記憶體有限

(四). 資料無法永久儲存，而且只能作一次處理。

(五). 當使用者有需求時，其分析結果能立即產生。

(六). 分析結果的偏差，必須在使用者能容許的範圍。

彙整日常生活使用的小型無線感知器(small wireless sensor)，可以創造一個豐富的非侵入性的感官數據環境。由於感知器硬體和行動式設備因小型化而成本降低，導致出現移動式行為識別(Mobile Activity Recognition)的研究，在一些現有的研究，人們在日常活動中使用可穿戴式感知器(wearable sensors)，也有一些人將感知器嵌入到日常生活居住環境的工具或器皿中，使能獲得更精細的活動數據分析。

移動式行為識別可以規範為一個分類的問題，其中監督機器學習是經由感知數據解譯成一種活動(activities)，學習過程通常會經過以下幾個階段：

(一). 數據資料收集：從一個或多個行動裝置使用者標記的活動，收集一段特定期間的感知資料。

(二). 數據資料傳輸：收集好的資料會傳輸和集中存放在一儲存體中。

(三). 學習模型的建立：利用收集的資料訓練並測試行為識別分類模型。

(四). 模型的部署：已學習模型部署在行動式裝備，利用感知資料來識別和分類活動。

移動式行為識別的方法是利用無所不在的感知器，對於一般的行為識別可達到高的辨識率，而感知數據行為識別提供監督機器學習有下列幾種演算法：

(一). 決策樹法(Decision Trees)。

(二). 類神經網路(Artificial Neural Networks)的 Hidden Markov 模組。

(三). Naive Bayes 的 K-Nearest Neighbour。

(四). Support Vector Machines。

近期有相關研究，使用手機內建的加速計(accelerometer)進行移動式行為識別，收集的數據來自 29 名對象的日常活動，如：散步、慢跑、爬樓梯、坐著、站立等行為，只要在活動進行時將手機放在固定位置，如：前腿褲的口袋裡，現今 Android 平台的智慧型手機，對於行為識別有超過 90%的正確辨識率，但現有方法會將資料經由 internet 傳遞到另一處的 Server，得到的模型是靜態的，是建立在行動裝置的離線環境，忽略了個人化的通用模型和隱私。為了解決這些問題提出移動式行為識別系統(Mobile Activity Recognition System；簡稱 MARS)，將分類模型

建立在 Android 平台的個人行動裝置板上，經由無處不在的資料流挖掘(Ubiquitous Data Stream Mining)的增量方法(Incremental Naive Bayes)來開採，它的優點是：

(一). 個人化模型，指可為單獨一個使用者而建立。

(二). 加強隱私性，因為資料未傳送到另一個存放地點。

(三). 模組可以依據使用者的活動概況調整和學習。

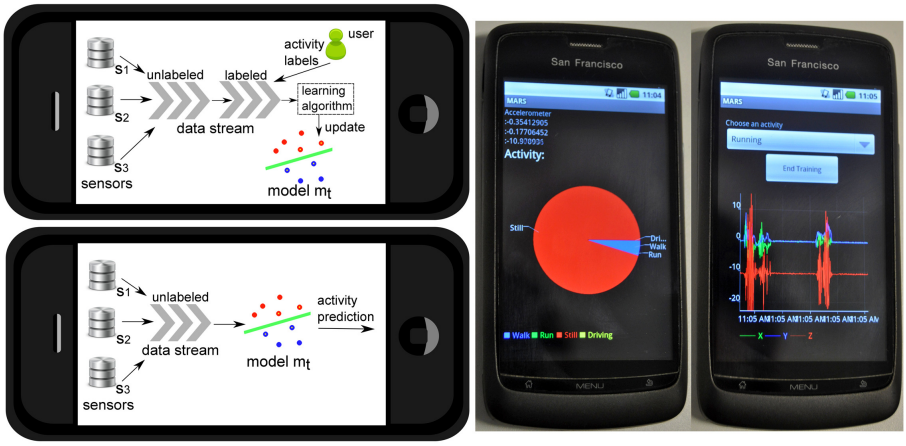
移動式行為識別系統(MARS)的學習過程可分為二個階段：

(一). 訓練階段(Training)：

使用者進行的活動不論是事先計劃好的或是任使用者自行選擇，感知器藉由 user-friendly 的界面執行註記活動和資料收集，例如：選項清單中事先列出之前進行過的活動。然後資料流直接在行動裝置上經由線上漸近式(incremental)學習演算法處理，進而預估並調整模組的正確性。下圖是 MARS 的訓練階段，經由學習過程和收集的資料，可以隨時對模組進行修正。

(二). 行為識別(Activity Recognition)：

未註記的新資料記錄會進行分類，並可提供模組做下一次的行為預測，下圖亦有這階段的展示。



MARS: framework and implementation

三. 動態地形資訊地景(Information Landscapes)－增量的(Incremental)視覺化的知識發現(Visual Knowledge Discovery)

增量計算的資訊地景對於在大型文件庫裡的縱向顯現改變是一種有效的方法，類似自然世界的構造過程。動態渲染反映了文件庫長期趨勢和短期波動。為了某主題的上升和衰退視覺化，對應的演算法提升，並降低相關同中心點的等高線。愈來愈多的文件採用最先進的知識發現應用來處理，本文介紹一種增量的，可擴展的方式來產生地景，處理管道包含一些連續的任務，從爬行(Crawling)、過濾(Filtering)和對網頁內容作一些前置處理，如：規劃、標記和渲染這些聚集的資訊。增量處理步驟被定位在推算預估階段，是由文件的同質分類(Clustering)構成，群組(Cluster)可促使力導向定位(Force-Directed Placement；簡稱 FDP)和快速文件定位，本文採用增量版和非增量版的輸出結果來作比較評估。實驗的文件取自環境網站 Media Watch on Climate Change (www.ecoresearch.net/climate)的部落格樣版，實驗結果證明增量計算方法可以正確的產生動態資訊地景。

近來我們面對的不只是不斷增長的，也是常常瞬息萬變的”大量資料”儲存庫，資訊地景是一種功能強大的視覺化技術，適用於大型文件儲存體中傳遞局部相關性資料。然而，資訊地景的觀念只能處理靜態條件視覺化。在之前的研究中，我們引入了動態地形資訊地景，可同時滿足：局部關聯及數據變化的視覺化。因此，動態地景已經證明了它在企業觀點中的重要價值，涉及在大型動態文字儲存庫裡，有關於視覺的知識發現(Knowledge Discovery)，已被應用在追蹤局部的關係、媒體趨勢和獲得專利的資料庫。

動態地形資訊地景是以地圖的地形特徵來作為視覺化展現，其局部關聯性在視覺化的空間裡，經由相近的空間傳遞，如：丘陵群(同質性的

群組 Clusters)都有局部類似的文件，山從低層文件就有重要項目的標記，以方便使用者定位，當一個文件庫隨著時間的變化，如：新的文件加上去或是舊文件移除，整理局部跟著結構也會改變。動態資訊地景傳達這些構造過程的改變，修改相對應的地景地形，在這個過程中產生的資訊地景，高維度數據資料投射對應到低維度空間，最常採用的就是降維(Dimensionality Reduction)演算法。

在很多科學研究領域裡，常需利用向量的形式來描述所欲探討之資料物件。以視訊影像為例，一張解析度為 64 x 64 像素的灰階影像可以表示成一個 4096 維度的特徵向量(feature vector)，每一維度儲存了相對應像素的灰階值。將物件以特徵向量之數學化形式呈現後，一個物件即可被視為向量空間中的一點，而一群物件將會在空間中形成某種分布，這有益於實際應用上的分析，如物件的特徵萃取(feature extraction)、分群(clustering)、分類(classification)或辨識(recognition)。

將資料物件表示的越詳盡，相對提高了向量空間的維度，高維度空間除了大幅增加計算的複雜度，且需要以指數成長的樣本數才能正確地分析資料(curse of dimensionality)，但實際應用上可用的資料往往是有限的。近來由於多媒體應用蓬勃發展，常需要分析如視訊、聲音、超文件或高解析度醫學影像等資料，這些物件其特徵向量在本質上往往就是高維度的，可於初始的物件表示上，使用較高維度的向量來包含特徵資訊，再透過降維(dimensionality reduction)演算法，來萃取或保持其在高維度空間裡所隱含的重要性質。

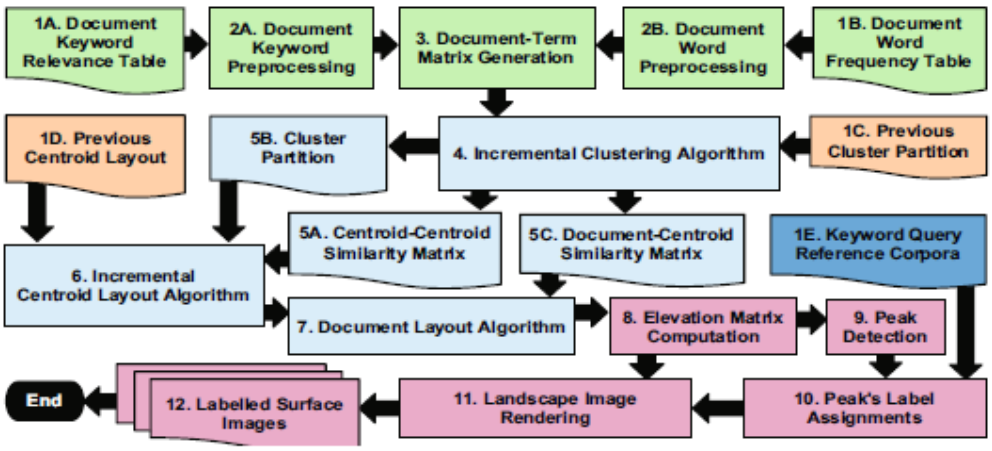
而現有的降維(Dimensionality Reduction)方法缺少以下幾種觀點：

(一). 支援增量計算模式。

(二). 數據集的大小和高維度。

(三). 令人容易瞭解且美觀的輸出，是必要的視覺化應用。

資訊地景在大型文件儲存庫裡常用來視覺化局部關連性。靜態地景可視覺化，但不能傳遞變化。如：ThemeRiver 是用來做視覺化展現，設計來傳達局部性群組的改變，但它不能表達文件和局部性群組的關連性，局部性改變的視覺化經由資訊地景與動態地形景觀，其中 Sabol 證明對大型資料集是有效的方法，它依賴地景幾何形狀動畫轉型的 3D 加速器。視覺化技術用在處理現今不斷增長的數據生產和數據消費，增量演算法提供處理”大量資料”所需的功能，當新的數據項目產生時，增量演算法不會對內部模組重新計算，因此能夠處理和整合不斷變化和不斷成長的數據資料，在產生動態資訊地景的背景之下，我們採用增量式降維方法(Incremental Dimensionality Reduction)和增量式分組(Incremental Clustering)技術來實作。



Workflow diagram for the incremental landscape computation framework

上圖是整體的工作流，分成三個主要部分：

(一). 首先是綠色顯示部分，是準備增加的文件項目矩陣，由適當關鍵字和詞頻率表的組合。

(二). 再來是青色部分，建立群組和置放文件。使用 k-means 的同質分組演算法來切割文件，變成局部相關的群組。然後力導向定位

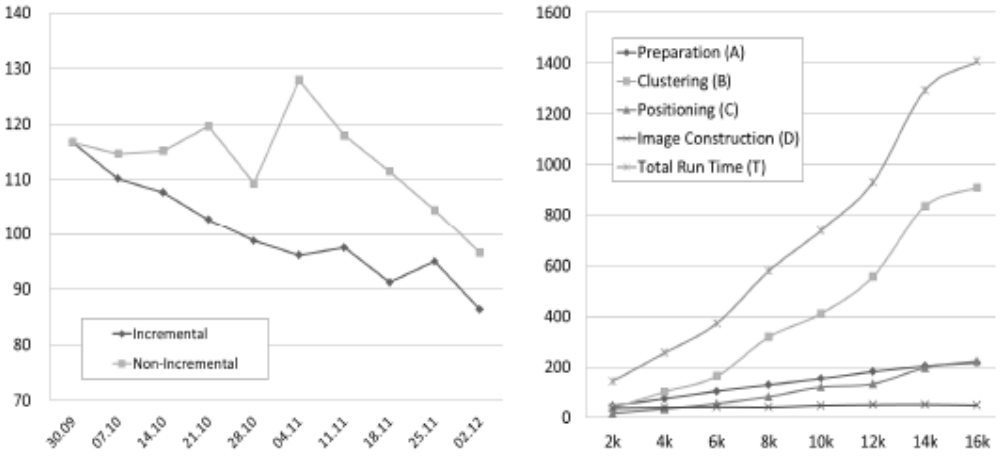
(Force-Directed Placement；簡稱 FDP)群組至 2D 的視覺化空間，並以 2D 群組位置為基礎快速定位文件位置。

(三). 最後是粉紅色部分，使用文件的輸出位置模擬一個局部地景，實際上是在 2D 網格矩陣中，一個高度海拔地景。

利用著色來建構地景和表面圖形，峰值檢測算法找到主要的峰群(丘陵)，然後收集相關文件，計算文字描述以標記峰值，需注意的是之前的計算結果，為增量處理的初始狀態(橘色部分)。

增量式和非增量式二種類型的壓力值計算和總結在下圖。下圖左：初始樣本是 2000 筆文件，從 2011 年 9 月 30 日開始，每一星期這個文件的變化，即新的文件到達，為維持一定數量的文件，會刪除最舊的。在非增量情形下，例如 11 月 4 日曲線顯示出更多的波動，這是由於 K-means 和 FDP 對初始情況的敏感性；而增量式的壓力測試值結果較低，推測是因為每週的增量改變僅是局部極小的影響 FDP 過程，表示增量式計算方法可改善其效能，並能準確的產生動態資訊地景。

右圖則是對八種不同大小的文件集，從 2000 筆到 16000 筆地景計算總結，包括(A)前置處理的文件項目矩陣，(B)同質分組，(C)文件定位，(D)群組採用 FDP 定位及固定群組數、峰值檢測、標籤定位，(T)針對不同大小的資料集，產生動態資訊地景圖形的總執行時間。



Left: The stress values (y-axis) for incrementally computed documents layout and for non-incrementally computed documents layout over a period of 10 weeks; Right: Run times in seconds (y-axis) for landscape computation framework with different document sets (x-axis)

四. 非局部性分類(Non-topical Categories)的網站分類

在過去幾年全球資訊網(World Wide Web；簡稱 WWW)有極大的成長，只要有網路存在，終端使用者很容易接觸到一般大眾。隨著愈來愈多的個人、組織和政府發佈資訊在網路上，就愈來愈難在網路上找到令人滿意的資訊。例如：有人想從健康診所的網站上知道，它是否是政府資助的，因此醫療費用是由公共醫療保險支付，在這樣的情況下，相關的網站上若有令人滿意的標籤時，則對搜索帶有查詢的鏈結標籤(linking labels)有很大的幫助，同時對使用者過濾網站也更容易。對自動分類網站自動建立 web 目錄時也很有幫助，相對於若是手動標記網站可是要付出相當的代價。

網站分類在可以處理文字分類(Text Classification)的假設之下，網站是網頁和文件的集合。有一種情況是原文分類到非局部性的類別時，這些類別可能無法在文字上描述的很好，而必需要維護一組特徵集，這種情況類似對情緒或心情的文件分類，識別文字風格等，例如：特徵好比

是詞性(part-of-speech)標籤，標點符號，及從本文的詞來命名等已被證明是有用的，而除了本文和詞性(part-of-speech)樣式，網站的 link 架構、URL 樣式和 HTML 標籤，都可以提供有用的資訊，幫助正確分類網站。

在這篇研究中，我們將網站分類歸類於四種非局部性的種類：公立、私立、非營利性、和以商業營利為主的，我們所關注的是在加拿大有關減肥或控制肥胖的網站。有許多網站對肥胖控制提供許多的服務，因此，對整個網站進行分類，將會對這些組織揭示重要的事實，進而告知使用者，例如：這些機構提供的服務的成本和可靠性。相關領域專家也證實這些事實將有益於肥胖患者，從網路上有效的過濾取得所需的資源。

為了對整個網站進行分類，我們認為網站是許多網頁的集合，視為一個單一文件，而不是只憑一個網頁或部分網頁內容就對網站進行分類。一個網站可以包含數百頁網頁，每一網頁以詞袋模型(Bag-of-Words)為基礎加入一些特徵，它可能需要一段時間來萃取詞性(part-of-speech)和命名實體。而更多的屬性也意味著分類階段需要花更多的時間。故採用降維法(dimensionality reduction)分析下列特徵：資訊增加和網頁點擊深度、使用者找到網頁所須點擊次數、網站的點擊深度增加，網頁的增加數量等。在多標籤分類(Multi-label Classification)設定下採用支援向量機(Support Vector Machine；簡稱 SVM)分類器。

資訊擷取(Information Retrieval，簡稱 IR)常使用在數位圖書館或網際網路搜尋引擎上，為確保其系統的效率，常使用 Recall 與 Precision 二個評量方法作檢視；評量搜尋的召回率(Recall)是找到有意義於所有有意義的比例；精確性(Precision)是找到中有意義的比例。其二者之間，當系統儘可能找到所有相關資訊給搜尋者時，即 Recall 值越大，其中包含不相關的資訊會越多，其精確率會下降。而 F-measure 是精確率(Precision)

與召回率(Recall)二個數值的協調平均值，其值介於 0 與 1 之間，可評量系統抓取資料的精確率與召回率，公式如下所示，在物理上的意義，數值 2 是因為分母 Precision 與 Recall 相加其最大值為 2，所以 F-measure 的值表示約當 Precision 與 Recall 任一個數值越小時，其二者相乘之結果會越小。

$$F - Measure = \frac{2 * \Pr ecision * \mathrm{Re} call}{\Pr ecision + \mathrm{Re} call}$$

Recall 與 Precision 協調值-- F-Measure

本篇研究結果顯示以關鍵字為基礎，謹慎選擇和處理關鍵字，可以達到 F-measure 的 70%，若再加上網站結構，詞性(part-of-speech)和實體命名等特性，可以更進一步達到 F-measure 的 74%，尤其是當網頁文字特徵不夠正確或是不足以提供準確資訊時，對網頁分類提供很大的改善。

肆、心得與建議

主計資訊處統計資訊科長期以來提供國勢普查處有關人口、農漁、工商三大普查相關資訊服務，由於普查的資料量非常多，一直以來都是使用大主機在處理資料及執行相關作業，包括：調查前的名冊整理，調查資料回收後的檢核、推計、彙編等資訊作業，另外還有大量的公務資料整理作業等。普查的大量資料，往往花費很長的時間在做資料檢核，好不容易檢核完資料，最後只能在有限時間內趕緊完成厚厚的分析報表，調查資料整理成文字檔案型態儲存，接續又要趕緊投入下次的普查前置作業。近年來資訊科技進步日新月異，不論是硬體或軟體皆不可同日而語，硬體體積愈來愈小，處理能力愈來愈強，成本也愈來愈親民；雖然大主機作業有前輩們的經驗傳承，和良好的架構設計，但軟硬體設備技術人力培訓不易，且維運成本一直很高的考量之下，轉換到新的硬體平台和使用新的資訊技術，已是接續未來的發展趨勢。

想縮短資料檢核時間，最根本的解決方式就是出外訪查時，當下即取得正確的調查資料，而為了取得良好的調查品質資料，目前仍和下列幾項因素有密切的關係：

一. 輕便的硬體設備供調查人員執行訪查工作：

讓調查人員願意帶設備出去訪查，使在訪查時可經由檢核系統即時取得乾淨正確的調查資料，同時又可減輕紙本問卷及參考文件重量。例如：智慧型手機、平版電腦。

二. 硬體設備電池電力足夠支撐在外調查時間：

由於不可能向訪查對象借用充電的情況下，在外訪查期間硬體設備需有足夠電力，才能取回正確調查資料。例如：智慧型手機、平版電腦搭配還算輕便的行動電源。

三. **無線網路的成本及鄉鎮普及率：**

為了調查檢核系統軟體的使用和調查資料的傳遞。這一項在目前的環境是每個月的成本支出，隨著調查人員的人數，長期下來會變成一項很可觀的支出，因此這一項目前還未到可接受的程度。

相信隨著資訊科技軟硬體的進步，上述的限制，眼看已是不久即可實現的願望。將繁雜的資料處理作業整合系統化，簡化和縮短資料檢核時間，利用資訊技術將大量的調查資料和歷史資料，建立成資料倉儲(Data Warehouse)的環境，經由知識探勘(Knowledge Discovery)及線上分析處理系統(On-Line Analytical Processing；簡稱 OLAP)的應用，由新的平台提供更好的資訊服務，如：

一. **針對主辦調查單位：**

因為調查回收的都是乾淨正確的調查資料，可以縮短調查及資料檢核過程，很快就能呈現調查結果，進而加以應用。

二. **針對統計專業人員：**

應將工作重點放在普調查資料的研究分析上，不論是展現實際情況或是預估未來走向，都必需提供及時、有效及精確的統計數據供高層決策支援使用。

三. **針對組織決策高層：**

組織決策高層是帶領組織前進的指標，因應社經環境快速變遷，決策高層希望可以快速取得各類統計數據，甚至可能是多維度或是跨系統的分析結果。

四. **針對一般民眾查詢：**

民眾對於政府施政及調查統計都有知的權利，所以在政府對外揭露資訊的網路平台上，應避免教條學術性的宣導展示，必須有使用者親和性

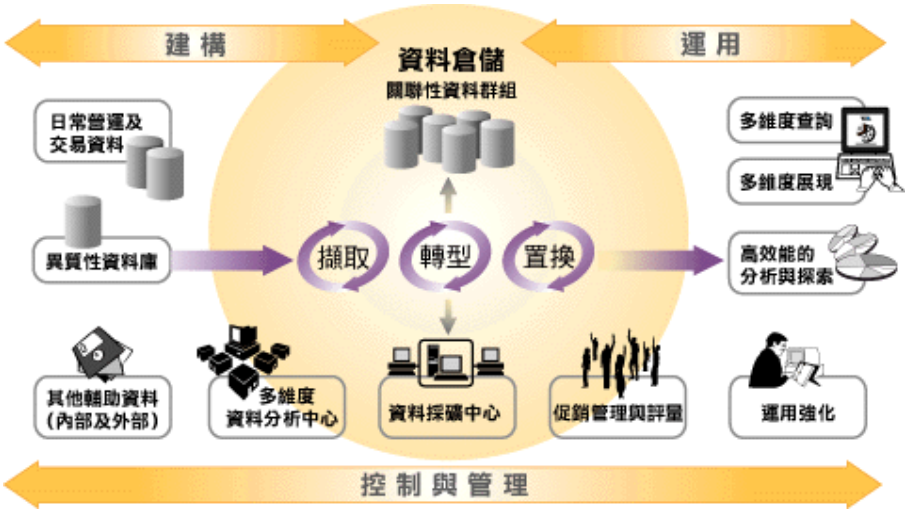
(user-friendly)的界面，除此之外，要能達到不論年紀大小、教育程度高低，所有人都能充分明瞭的目標。

五. **針對調查資料交換：**

花費大量人力及金錢所做的調查，希望可以達到它最高經濟效益價值，故在對受調查對象個別資料嚴格保密的前提之下，公務機關或學術機構在法律的約制下，可以相互提供或索取調查資料，供進一步的分析研究使用。

六. **針對資訊專業人員：**

提供上述人員使用的資訊平台，是資訊人員專業技能的挑戰，以資料倉儲(Data Warehouse)為基礎的環境，最重要的是如何設計規劃資料倉儲的架構，如何轉換和整合資料的過程，即 ETL(Extraction, Transformation, Load)過程，才能在未來有彈性的應用，且能快速支援外界各類不同的需求，提昇組織的效率與生產力。



（圖片來源：輔仁大學統計資訊學系謝邦昌教授的 Data Mining 授課講義）

作業資訊化幫助我們簡化工作流程，因應科技進步和作業系統的改變，本科長期以來自行規劃、分析、設計、撰寫、上線，協助國勢普查處完成了

許多應用系統及系統版本的更新，例如：人口、農漁、工商等三大普查行政作業管理系統，從 DOS 系統版本、Windows 系統版本，到近期的網路版本 CAS；大主機處理三大普查資料的檢核、推計、編表系統，公務交換資料檔的彙整及應用；協助各縣市主計單位管理考核基網人員的基層統計調查網作業管理系統；及應用在人力資源調查，各縣市按月執行的電腦輔助面訪調查系統(Computer Assisted Personal Interviewing，簡稱 CAPI) 等等，都是自行開發撰寫的應用系統，因此普遍獲得好使用、好維護的回應，但都是各自獨立運作的應用系統，系統間偶有少部分資料交流。而藉由此次國際研討會的機會，吸收一些新知識及他人經驗，希望對未來的統計資訊平台建置，提供幾點參考建議：

一. **整合規劃：**

未來統計資訊科應該針對這些個別獨立運作的應用系統，審慎考量各個應用系統間的關連性和共用性，如：類似作業的應用系統是否能整合需求彈性共用；應用系統間的資料如何交流，避免重複的輸入作業；因應查回資料特性，可能在一般檢核作業中，彈性加入額外處理程序等等。在建立整合性資料倉儲環境及統計資訊平台時，都應一併納入考量來作整體規劃。

二. **資訊安全：**

普查資料一直以來是存放在封閉式大主機環境，不論是對存放的資料，或是登入系統的使用者，都有嚴格的權限控管；同時，它不屬於主流性作業系統，熟悉的技術人員不易培養，因此安全性良好。今年十月起正式施行個人資料保護法，公務機關應指定專人辦理安全維護事項，防止個人資料被竊取、竄改、損毀、減失或洩漏，個資外洩最嚴重可能涉及刑責和懲罰性賠款。故一旦將作業環境從大主機環境轉換到其它資訊平台，不論採用何種系統架構，對於內含大量個資的調查資料，一定要有完整的資訊安全計畫，使資料

不論是在儲存時、系統運作時、資料展現或交換時，都能對受調查對象提供生命財產最佳保護。

三. **資料交換：**

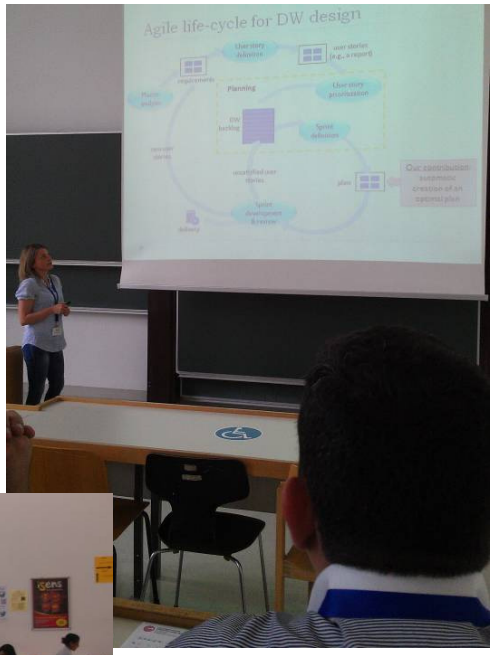
不論是組織內的應用系統之間，或是與外單位的資料交換，在確保資訊安全的前提下，應規劃分析如何快速產出符合需求的資料集，資料來源可能是從一個或是跨多個應用系統的產出，可能是整合彙總後的結果，也可能是某種階段的明細資料，資料集容易再加工應用，如：資料集可轉到 Word、Excel 或製成 PDF，再做後續的應用；或是資料集可輸入到另一個應用系統，供其它應用系統處理作業，可簡化資料輸入作業，及相關資料的完整正確性。因此，資料交換程序和資料格式在建置統計資訊平台時，也應詳加考量規劃。

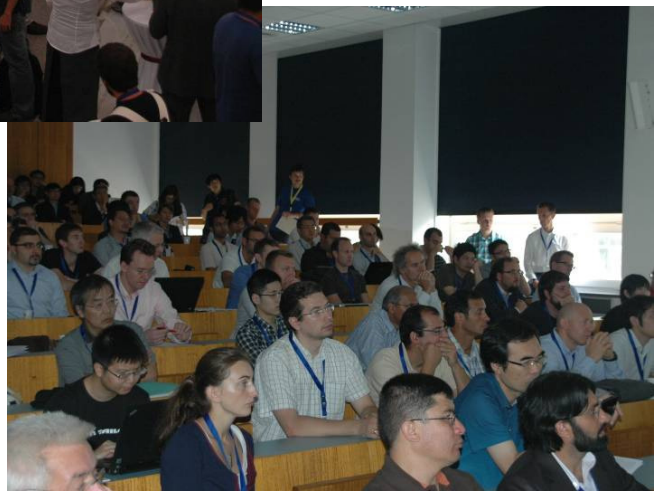
四. **資料展現：**

不論是查詢資料的產出，或是統計資料結果呈現，隨著資訊套裝軟體的進步，資料呈現除數據資料外，也應搭配易讀易懂的圖表顯示，如：2D 或是帶有空間意義的 3D 圖示，搭配不同顏色解釋說明；或是含有地域意義的 GIS 圖示；甚至更進一步與使用者有互動的視覺化動態圖形展示。使資料展現不僅僅是數字的排列，終極努力目標是年齡無論老幼，學歷無論高低，對於資料需求者，都能簡單清楚明瞭。

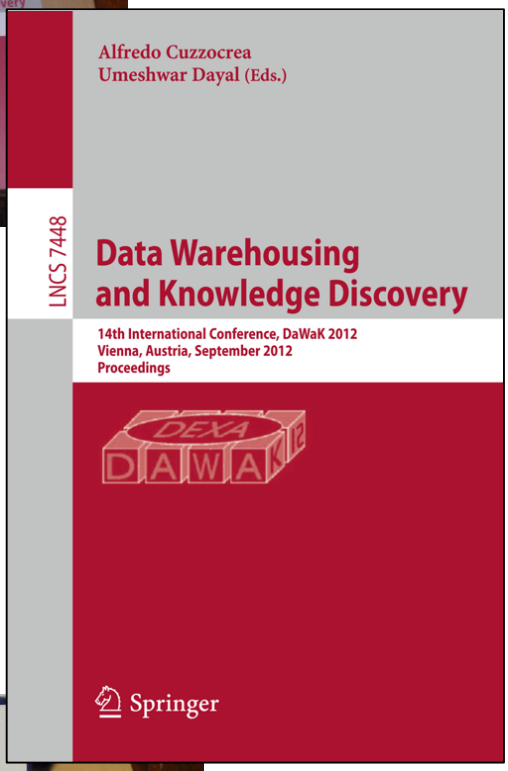
伍、參考資料

DEXA 2012年會會場花絮

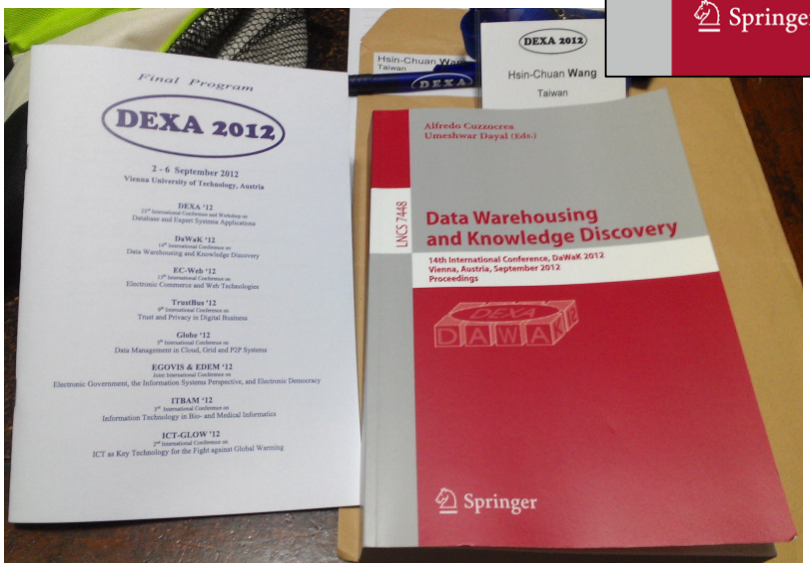




DEXA 2012年會提供DaWaK發表論文集



DaWaK論文集ISBN
978-3-642-32583-0



DEXA 2012年會參加證

CERTIFICATE OF PARTICIPATION

This is to certify that

Hsin-Chuan **Wang**

has participated in the **DEXA Event 2012**

DEXA 2012 - *23rd International Conference on Database and Expert Systems Applications*
DaWaK 2012 - *14th International Conference on Data Warehousing and Knowledge Discovery*
EC-Web 2012 - *13th International Conference on Electronic Commerce and Web Technologies*
TrustBus 2012 - *9th International Conference on Trust and Privacy in Digital Business*
Globe 2012 - *5th International Conference on Data Management in Grid and P2P Systems*
ITBAM 2012 - *3rd International Conference on Information Technology in Bio- and Medical Informatics*
EGOVIS & EDEM 2012 - *Joint International Conference on Electronic Government, the Information Systems Perspective, and Electronic Democracy*
ICT-GLOW 2012 - *2nd International Conference on ICT as Key Technology for the Fight against Global Warming*
DEXA Workshop 2012 - *23rd International DEXA Workshop on Database and Expert Systems Applications*

DEXA 2012

3 - 6 September, 2012
Vienna, Austria

Prof. Dr. Roland R. Wagner
President of the DEXA Association

DEXA Association
Gesellschaft für Datenbank- und Expertensystemanwendungen
Biesenfeldweg 12
A-4040 Linz

DEXA 2012年會參加費用

DEXA
http://www.dexa.org

Society for Database and Expert Systems Applications

ORIGINAL

Austria, Linz, 8/27/2012

DGBAS
Ms. Hsin-Chuan Wang

No.2, Guangzhou Street
Zhongzheng Dist.

10065 Taipei
Taiwan, R.O.C.

Invoice No. A-DEX12-288

For publication services for

Ms. Hsin-Chuan Wang (1506)

in the conference/workshop proceedings of the

DEXA 2012 Event
September 3 - 7, 2012
Vienna University of Technology, Vienna, Austria

we charge the amount of:

PAID

Fee	EUR	590,91
VAT (10%)	EUR	59,09
Total	**EUR**	**650,00**

Conference bank account:
DEXA 2012
Account No. 5.727.276
Raiffeisenlandesbank Oberösterreich
Zweigstelle Biesenfeld, A-4040 Linz
Austria

IBAN AT36 3400 0000 0572 7276
BIC RZOOAT2L

DEXA Society
Gesellschaft für
Datenbank- und Expertensystemanwendungen
Biesenfeldweg 12
A-4040 Linz

Gesellschaft für Datenbank- und Expertensystemanwendungen (DEXA)
Biesenfeldweg 12
A-4040 Linz

T +43 (676) 846 732 11 F +43 (7236) 3343 782

ZVR-Zahl: 641491400

UID-Nr.: ATU62401056
Steuernr.: 159/6700

Raiffeisenlandesbank Oberösterreich, Zweigstelle Biesenfeld
BLZ: 34000, Konto-Nr.: 05.727.276
IBAN: AT36 3400 0000 0572 7276, BIC: RZOOAT2L

1/1