關聯推理神經網路

葉怡成
中華大學資訊管理學系

杜進明
中華大學資訊管理學系

丁導民
開南大學企業與創業管理學系

王逸芸
中華大學資訊管理學系

劉謹豪
中華大學資訊管理學系

摘要

本研究所提出的關聯推理神經網路(Association Reasoning Neural Networks, ARNN)是修改自倒傳遞神經網路演算法，可以產生關聯規則，為傳統的關聯分析開啟完全不同的研究途徑。經由一個數值例題與一個實際例題的結果歸納出以下結論：(1)ARNN的推論輸出值與資料中的信賴度大約相等。(2)當ARNN的隱藏單元數減少時，檢出率、準確率會跟著變小。(3)ARNN可以利用減少隱藏單元數來避免產生信賴度過低的關聯規則。(4)ARNN產生的規則與由傳統關聯分析法所找到的規則重疊性很高，且能找出一些被傳統關聯分析法忽略的關聯規則。

關鍵字：資料探勘、類神經網路、關聯規則、關聯分析

Association Reasoning Neural Networks

I-Cheng Yeh
Department of Information Management, Chung Hua University

Chin-Ming Tu
Department of Information Management, Chung Hua University

Tao-Ming Ting
Department of Business and Entrepreneurial, Kainan University

I-Yun Wang
Department of Information Management, Chung Hua University

Chin-Hao Liu
Department of Information Management, Chung Hua University

Abstract

In this study, we proposed the Association Reasoning Neural Network (ARNN) which is derived from ANN, can produce association rules, and open a new approach for association analysis. Based on a numerical and a practical example, some conclusions can be gotten: (1) The reasoning output value of ARNN is proportional to the confidence of the association rules hidden in the data set. (2) As the number of neurons in the hidden layer of ARNN is decreasing, the detection rate and the accuracy rate are decreasing. (3) The ARNN can avoid producing the association rules with low support value by reducing the number of neurons in the hidden layer of the network. (4) The association rules generated by ARNN are similar to those generated by traditional association analysis.

Key words : Artificial neural network, association analysis, association rule, data mining.

壹、前言

關聯分析(association analysis)是在一堆看似無關聯的資料中去找到一些關聯性，經常被用來做為分析顧客購物行為的工具，促進商品的銷售。關聯探勘可定義為：「給予一組記錄，每筆記錄登記了一些項目。找出一個能夠以某些項目出現與否來預測其它項目出現與否的關聯規則。」

近年來，關聯分析的技術日益成熟，其中較廣為使用的演算法有Apriori、DHP、FP-Growth等 (Hipp 2000)。雖然這些關聯分析方法具有可產生明確的關聯規則、具機率理論基礎的優點，但其缺點是(Hipp 2000；黃南傑2004)：

(1) 當項目很多時會造成運算時間的爆炸性成長。

(2) 項目(輸入變數)只能限制在0或1。

(3) 無個案的預測能力。例如底下的三條由傳統關聯分析法所得的規則：

關聯規則1：If A Then E (Confidence=60%)

關聯規則2：If B and C Then E (Confidence=80%)

關聯規則3：If A and D Then E (Confidence=50%)

當A、B、C、D同時出現時，傳統關聯分析法無法預測E的機率。

已有許多研究將其目標設定在改善傳統關聯分析方法的效率，與修改演算法滿足不同的應用上(Cai 1998；Chan 1997；Hipp 2000；Liu, et. al. 1999；Sarawagi et. al. 1998；Zheng et. al. 2001；黃南傑2004；王錫中2002)，但還很少有研究把注意力放在第(3)個缺點上。

近年來類神經網路(Artificial Neural Network, ANN)已被視為非常有效的非線性模型建構工具(林瑞山2004；陳建銘2001；葉怡成2006)。傳統上倒傳遞網路是以監督式學習(supervised learning)為主，但使用倒傳遞網路於自聯想學習(auto-association learning)的文獻很早就有人提出，特別是應用在資料壓縮的領域(Sonehara et. al. 1989；Abdel-Wahhab & Fahmy 1997；Arozullah & Namphol 1990；Setiono & Lu 1994a；Huang et. al. 1991; Abidi et. al. 1994；Setiono & Lu 1994b；Wang & Oja 1993；劉向陽、王如雲2006)。資料壓縮是指將高維次資料以較低維次數據來表現，一個具有 N維次的資料可用N-M-N架構的倒傳遞網路將它壓縮到M維次，其中M小於 N，即網路的輸入、輸出相同，均為原N維次資料，網路的隱藏層輸出即壓縮後的數據。當M越小時，壓縮的比例越大，但失真的情形也可能更嚴重。這種學習模式的特色是其輸入與輸出(目標輸出)是相同的，因此被稱為自聯想學習，與Hopfield network (Paik & Katsaggelos 1992)的associate memory有異曲同工之妙。

因為類神經網路本身具有推論及聯想的能力，具有改善前述傳統關聯分析方法無個案的預測能力缺點的潛力。然而利用神經網路來產生關聯規則的研究卻很少。故本研究修改倒傳遞神經網路(Back-Propagation Network, BPN) 演算法，提出的關聯推理神經網路 (Association Reasoning Neural Networks, ARNN)，來建立具有個案的預測能力的關聯模

型，以克服傳統的關聯分析方法的缺點。

ARNN是一個多層神經網路，其架構如圖1所示，輸出層的節點數的數目和輸入層的節點數相同，每個節點均代表一個項目，此外還包含一層以上的隱藏層，其節點數目是可調整的變數。ARNN將每一筆記錄視為一個範例，在記錄中出現的項目，其對應的節點值為1，否則為0。在訓練過程中，ARNN會逐一載入每一筆記錄，並將記錄中出現的項目以「隨機」的方式決定是否放入神經網路的輸入層，但以全部的出現項目都會放入神經網路的輸入層，當作神經網路輸出層的目標值。

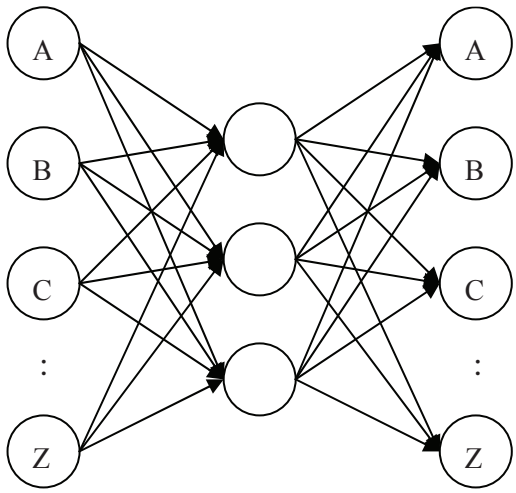


圖1：ARNN的架構

一旦ARNN訓練完成後，隱含在記錄中的項目之間的關聯已被儲存在神經網路的連結加權值當中。當需要產生關聯規則時，只要將有興趣探討的關聯規則左端項目X對應的ARNN輸入層的節點值設為1，其它項目對應的節點值設為0，輸入ARNN，即可計算出輸出層的節點值，其值超過預設的「法則門檻」的節點其對應的項目即關聯規則左端項目Y，產生X=>Y的關聯規則。因此ARNN產生規則前，並不需產生最大項目集，也不需設定最小支持度和最小信賴度，其關聯規則的產生受「隱藏層節點數目」與「法則門檻」控制。ARNN的優缺點正好與傳統的關聯分析方法相反，ARNN的優點有：在項目很多時，不會造成運算時間的爆炸性成長；項目(輸入變數)不限制在0或1，可以是0~1之間的任意值；具有個案的預測能力，即在特定項目出現下，可預測其他項目出現的可能性。而ARNN的缺點有：無法產生明確的、完整的關聯規則；不具有機率理論基礎。

本文第二節介紹ARNN的理論與方法，第三節以一個數值例題探討ARNN的性能，第四節以一個應用實例來比較ARNN與傳統關聯分析方法，第五節說明本研究的主要貢獻和未來的研究方向。

貳、關聯推理神經網

一、學習模式

本研究將關聯分析視為「給予一組交易記錄，每個記錄包含若干個項目，試著建立一系統，其輸入是記錄的部份項目，透過系統得到的推論輸出是該記錄的完整項目。」故假設ARNN的誤差函數如下：

$$E=\frac{1}{2}\sum_{j}(T_j-Y_j)^2 \tag{1}$$

其中

T_j＝神經網路輸出層的目標值＝X_j (2)

Y_j＝神經網路輸出層的推論值，其公式如下：

$$Y_j=\frac{1}{1+\exp(-net_j)} \tag{3}$$

$$net_j=\sum_{k}W_{kj}H_k-\theta_j \tag{4}$$

其中net_j＝隱藏值之加權乘積和；W_{kj}＝第 k 個隱藏單元與第j個輸出單元間的連結加權值；θ_j＝第 j 個輸出單元的門限值。H_k＝第 k 個隱藏單元的輸入值。

$$H_k=\frac{1}{1+\exp(-net_k)} \tag{5}$$

$$net_k=\sum_{i}W_{ik}R_iX_i-\theta_k \tag{6}$$

其中net_k＝輸入值之加權乘積和；W_{ik}＝第 i 個輸入單元與第 k 個隱藏單元間的連結加權值；θ_k＝第k個隱藏單元的門限值。$X_i\in\{0,1\}$，由已知的交易記錄決定，在記錄中出現的項目，其對應的X_i值為1，否則為0。$R_i\in\{0,1\}$，以隨機亂數設定。

在(6)式中，由於$R_i\in\{0,1\}$，是一個隨機亂數，故$\{R_1X_1,R_2X_2,...,R_nX_n\}$向量所表現的是「交易的部份項目」，而在(2) 式中，$T_j=X_j$，故$\{T_1,T_2,...,T_n\}=\{X_1,X_2,...,X_n\}$，此向量所表現的是「交易的全部項目」。

在(6)式中，$R_i\in\{0,1\}$，並隨機設定的理由是：如果真的有關聯規則$X=>Y$存在，則它滿足$Support(X\cup Y)$大於支持度下限，且信賴度大於信賴度下限的要求。故會有相當多的樣本(交易)會同時含有項目X與項目Y。藉由隨機設定$R_i\in\{0,1\}$，可以使ARNN的輸入端有相對高的機會只含這些交易的部份項目X，而輸出端則含其全部項目$X\cup Y$。在ARNN

中，輸出端的推論輸出值Y是網路連結權值與輸入端輸入值X的函數，故如果能調整網路連結權值，來最小化(1)式的誤差函數，即可使神經網路輸出層的推論值Y逼近其目標值T，滿足前述的「建立一系統，其輸入是記錄的部份項目X，透過系統得到的推論輸出是該記錄的完整項目$X \cup Y$」之要求，此即關聯推理神經網路(ARNN)之基本原理。

第(1)~(6)式的ARNN架構與傳統的倒傳遞網路相近，二者的差異在於第(6)式，但一樣可用最陡坡降法求解。其學習過程是以一次一個訓練範例的方式進行，直到學習完所有的訓練範例，稱為一個訓練循環 (learning cycle)。一個網路可以將訓練範例重覆學習數百甚至數萬個訓練循環，直至達到收斂。

關聯推理神經網路演算法整理如下：

1 設定網路參數。

2 以均佈隨機亂數設定加權值矩陣與門限值向量初始值。

3 讀入一個訓練範例做為目標輸出向量T；並以下法產生輸入向量X

```
For i=1~n
    {
    產生一個(0,1)間的均佈隨機亂數R
    If (R>0.5) Then X[i]=T[i]
    Else X[i]=0
    }
```

4 計算推論輸出向量 Y

用(3)~(6)式計算推論輸出向量 Y

5 計算差距量 δ

(a) 計算輸出層差距量 δ

$$\delta_j = (T_j - Y_j) \cdot Y_j \cdot (1 - Y_j) \tag{7}$$

(b) 計算隱藏層差距量 δ

$$\delta_k = \left(\sum_j \delta_j W_j \right) \cdot H_k \cdot (1 - H_k) \tag{8}$$

6 計算加權值矩陣修正量，及門限值向量修正量

(a) 計算輸出層加權值矩陣修正量，及門限值向量修正量

$$\Delta W_{kj}(n) = \eta \delta_j H_k + \alpha \Delta W_{kj}(n-1) \tag{9}$$

$$\Delta \theta_j(n) = -\eta \delta_j + \alpha \Delta \theta_j(n-1) \tag{10}$$

(b) 計算隱藏層加權值矩陣修正量，及門限值向量修正量

$$\Delta W_{ik}(n) = \eta \delta_k X_i + \alpha \Delta W_{ik}(n-1) \tag{11}$$

$$\Delta \theta_k(n) = -\eta \delta_k + \alpha \Delta \theta_k(n-1) \tag{12}$$

7 更新加權值矩陣，及門限值向量

(a) 更新輸出層加權值矩陣，及門限值向量

$$W_{kj} = W_{kj} + \Delta W_{kj} \tag{13}$$

$$\theta_j = \theta_j + \Delta \theta_j \tag{14}$$

(b) 更新隱藏層加權值矩陣，及門限值向量

$$W_{ik}=W_{ik}+\Delta W_{ik} \quad (15)$$

$$\theta_k=\theta_k+\Delta\theta_k \quad (16)$$

8 重覆步驟 3 至步驟 7，直到收斂。

二、規則擷取模式

經由倒傳遞神經網路的學習之後，可從網路中擷取關聯規則。關聯規則的擷取可分為一階和二階關聯規則擷取，即以特定的關聯規則左端項目組合做「關聯規則種子」，利用上述聯想模式，推論右端項目出現的可能性，並根據所設定的「法則門檻」來產生規則。

所謂一階關聯規則即是在ARNN網路的輸入單元同時只允許一個單元(X)的輸入值為1，其餘為0。再透過ARNN的推論得到各輸出單元的推論輸出值，當一輸出單元(Y)的此值大於一預設的「法則門檻」時，產生對應的關聯規則(X=>Y)。

所謂二階關聯規則即是在ARNN網路的輸入單元同時只允許二個單元的輸入值為1，其餘為0，以透過ARNN產生對應的關聯規則。

參、數值例題

一、簡介

本節將舉一個例子讓讀者更容易了解關聯神經網的運作方式。假設一五金行，總共有20種商品：A=鐵鎚，B=油漆，C=老虎鉗子，D=鋸子，E=鐵釘，F=刷子，G=螺絲起子，H=鑽刀，I~T為其它五金工具。共100筆的交易資料。在此資料集中：

Support(A∪E)=32	*Conf(E｜A)*=80%	*Conf(A｜E)*=100%
Support(B∪F)=16	*Conf(F｜B)*=40%	*Conf(B｜F)*=100%
Support(C∪G)=16	*Conf(G｜C)*=80%	*Conf(C｜G)*=100%
Support(D∪H)=8	*Conf(H｜G)*=40%	*Conf(D｜H)*=100%

其餘I~T項目仍有一些是1，屬於雜訊。理論上應該產生八條關聯規則：

Rule 1: A=>E (Conf=80%, Support=32)
Rule 2: B=>F (Conf=40%, Support=16)
Rule 3: C=>G (Conf=80%, Support=16)
Rule 4: D=>H (Conf=40%, Support=8)
Rule 5: E=>A (Conf=100%, Support=32)
Rule 6: F=>B (Conf=100%, Support=16)
Rule 7: G=>C (Conf=100%, Support=16)
Rule 8: H=>D (Conf=100%, Support=8)

二、方法與結果

接下來我們將以關聯推理神經網路演算法來實作這份資料。其步驟如下：

1. 首先複製成二份交易記錄分別作為ARNN的訓練和測試範例，雖然訓練和測試範例相同，但因為ARNN的輸入向量是$\{R_1X_1,R_2X_2,\cdots,R_nX_n\}$，不是$\{X_1,X_2,\cdots,X_n\}$，故具有隨機性，因此雖然訓練和測試範例相同，仍會有獨立的測試結果。
2. 設定ARNN的相關參數如表1所示，並以前節的算法訓練ARNN。訓練完成後將產生的連結加權值儲存。為證明ARNN的推論輸出值會逼近信賴度，在此繪一散佈圖，橫軸為實際在資料中的X=>Y中的信賴度，縱軸為ARNN的推論輸出值，如圖2。由圖可知ARNN的推論輸出值會逼近信賴度。
3. 以2.2節方法產生一階關聯規則，法則輸出門檻值設為0.6。產生的關聯規則如圖3。為了讓讀者了解，圖3除了列出推論輸出值大於0.6的項目，也列出推論輸出值在0.35~0.6之間的項目。

將以上關聯推理神經網路產生的關聯規則與理論上應該產生的關聯規則比較得知，原設計的8條關聯規則中，總共有6條規則被ARNN找出來，只有Rule 2(B=>F)和Rule 4(D=>H)二條規則沒有被產生出來，其推論輸出值分別為0.387和0.484。這是因為原資料中B=>F與D=>H的Conf=40%，以致於用此資料訓練出來的網路，將B或D輸入時，由圖3知，其F或H的推論輸出值大約會在0.4左右，低於法則門檻值0.6，因而未產生規則。

表1：數值例題之關聯推理神經網路的參數設定

參數	設定值
輸入層節點數 (Ninp)	20
隱藏層節點數 (Nhid1)	8
輸出層節點數 (Nout)	20
學習循環 (Ncycle)	3000
學習速率 (eta)	1.0

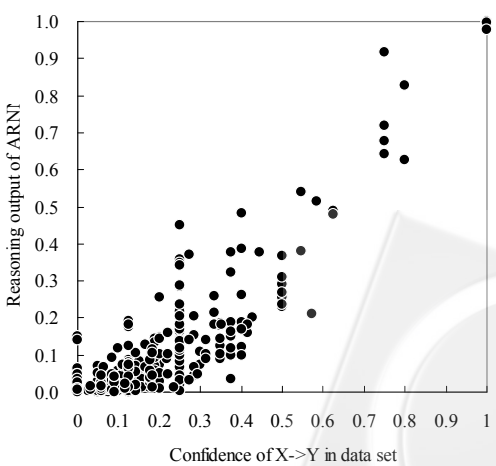


圖2：數值例題的ARNN的推論輸出值與資料中的信賴度之關係(隱藏單元數=8)

```
Rule 1:   A,   => E (0.625)
Rule 2:   B,   => F (0.387)
Rule 3:   C,   => G (0.827)
Rule 4:   D,   => H (0.484)
Rule 5:   E,   => A (0.999)
Rule 6:   F,   => B (0.999)
Rule 7:   G,   => C (0.997)
Rule 8:   H,   => D (0.978), R (0.378)
Rule 9:   I,   => B (0.718), A (0.367), P (0.357), R (0.451)
Rule 10: J,   => B (0.489)
Rule 11: K,   =>
Rule 12: L,   => B (0.513)
Rule 13: M,   =>
Rule 14: N,   => C (0.378)
Rule 15: O,   => A(0.642), D(0.917), E(0.678),
Rule 16: P,   =>
Rule 17: Q,   => B (0.478)
Rule 18: R,   =>
Rule 19: S,   =>
Rule 20: T,   => H (0.489), E (0.381), N (0.372)
```

圖3：數值例題的ARNN產生的一階關聯規則(隱藏單元數=8)
(加方框者為推論輸出值在0.35~0.6之間的項目)

另外，由圖3發現，除了前所指的8條規則外，ARNN也多找出2條規則出來，分別是

Rule 9: I=>B

Rule 15: O=>A, D, E

經統計後發現

I=>B的Support(I,B)=3，Conf(B|I)=75%

O=>A的Support(O,A)=3，Conf(A|O)=75%

O=>D的Support(O,D)=3，Conf(D|O)=75%

O=>E的Support(O,E)=3，Conf(E|O)=75%

探究其原因可能是在巧合的情況下，使得其Support與Confidence夠大，以致於用此資料訓練出來的網路，將I(或O)輸入時，其B(或A,D,E)的推論輸出值高於法則門檻值0.6，而產生關聯規則。

Rule 9與Rule 15在資料中的支持度很低，如果使用傳統的關聯分析，可以透過設定支持度的最低門檻來排除之。但ARNN缺少此一排除機制。為此，將ARNN的隱藏單元數採用很小的值重作，希望能透過縮小隱藏單元數，達到排除在資料中的支持度低的關聯規則。其推論輸出值與信賴度的關係如圖4。由圖可知，支持度較小的I=>B等四條規則其推論輸出值降低，而支持度較大的關聯規則(A=>E，C=>G，E=>A，F=>B，G=>C，H=>D)其推論輸出值不變。

此網路產生的關聯規則如圖5。因為支持度較小的I=>B等四條規則其推論輸出值降低，低於法則門檻0.6，故被移除。因此在隱藏單元數為2下只產生6條關聯規則。可見減少隱藏單元數可以排除在資料中的支持度低的關聯規則。

總結上述分析可知，傳統的關聯分析很重視支持度，但支持度對它而言是一個門檻，只有通過與不通過。ARNN沒有明確的支持度門檻機制，這一點與傳統的關聯分析很不相同。取而代之，ARNN可以利用提高法則門檻值來減少關聯規則的數目；利用減少隱藏單元數來抑制信賴度較低的關聯規則的產生。

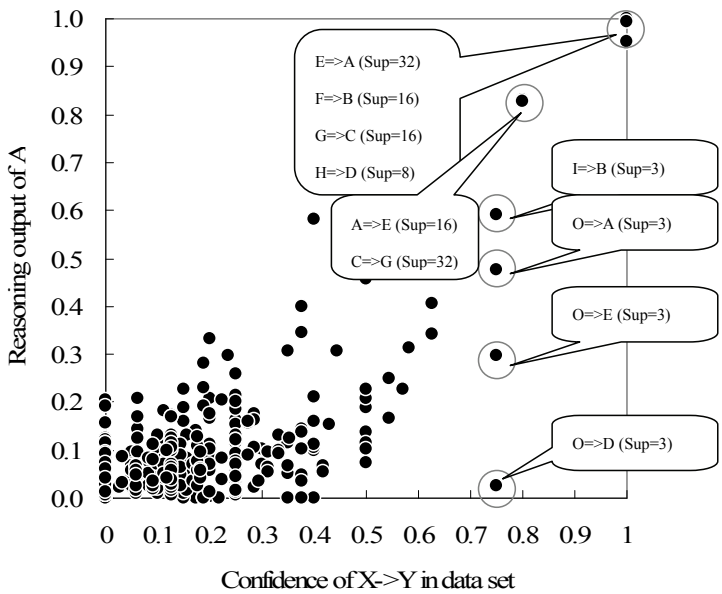


圖4：數值例題的ARNN的推論輸出值與資料中的信賴度之關係(隱藏單元數=2)

```
Rule 1:   A,   => E (0.832)
Rule 2:   B,   =>
Rule 3:   C,   => G (0.829)
Rule 4:   D,   => H (0.581)
Rule 5:   E,   => A (0.949)
Rule 6:   F,   => B (0.999)
Rule 7:   G,   => C (0.951)
Rule 8:   H,   => B (0.457), D (0.993), R (0.399)
Rule 9:   I,   => B (0.591)
Rule 10: J,   => B (0.405)
Rule 11: K,   =>
Rule 12: L,   =>
Rule 13: M,   =>
Rule 14: N,   =>
Rule 15: O,   => A (0.475)
Rule 16: P,   =>
Rule 17: Q,   => B (0.341)
Rule 18: R,   =>
Rule 19: S,   =>
Rule 20: T,   =>
```

圖5：數值例題的ARNN產生的一階關聯規則(隱藏單元數=2)
(加方框者為推論輸出值在0.35~0.6之間的項目)

三、聯想能力的評估

為評估ARNN之聯想能力，可用檢出率(detect ratio)、準確率(precise ratio)和失誤率(error ratio)，其計算方法如下：

1. A值(應有且發現項目數)：是指輸入的每一筆交易中其值為1的項目之總數扣除規則左端為1值的項目之數目，且出現在輸出的資料中其值大於期望輸出門檻值的項目之數目，亦即在理想上應該找到的右端之項目數目中，而ARNN實際所找出的右端之項目數目。
2. B值(應有項目數)：是指輸入的每一筆交易中值為1的項目之總數扣除規則左端為1值的項目之數目，亦即理想上應該找到的右端之項目數目。
3. C值(發現項目數)：是指輸出的資料中，其值大於期望輸出門檻值的項目之數目扣除規則左端為1值的項目之數目，亦即在ARNN推論後所找到的右端之項目數目。
4. 檢出率：計算公式為A/B，其目的是希望經由ARNN所找出的右端之項目數目和理想上應該輸出的右端之項目數目相比。
5. 準確率：計算公式為A/C，其目的是希望經由ARNN所找出的右端之項目數目和推論上應該輸出的右端之項目數目相比。
6. 失誤率：計算方法為D/E，其中D值為期望輸為1而且輸出不為1的項目數，加上期望輸出為0且輸出不為0的項目數，而E值為所有輸出單元數目。

上述數值例題在不同的隱藏單元數目與分類門檻下，其聯想能力的評估結果如表2。由表可知

1. 分類門檻：當分類門檻越高時，檢出率越低，準確率越高。故檢出率越高時，準確率越低；兩者之間為反比的關係。
2. 隱藏單元：當隱藏單元數越多，失誤率越低。這可能是因為神經網路的隱藏單元數多時，可以記憶大量的關聯規則，而隱藏單元數少時，則不足以記憶大量的關聯規則，因此而有較大的誤差。

表2：數值例題的檢出率和準確率的評估

隱藏 單元數	分類門檻	應有且發現 項目數(A)	應有 項目數(B)	發現 項目數(C)	檢出率% (A/B)	準確率% (A/C)	失誤率 (%)
2	0.7	32	153	35	20.92	91.43	9.66
	0.5	35	153	44	22.88	79.55	9.34
	0.3	42	153	69	27.45	60.87	10.14
4	0.7	35	149	41	23.49	85.37	8.32
	0.5	44	149	60	29.53	73.33	7.75
	0.3	59	149	112	39.60	52.68	8.80
6	0.7	26	154	30	16.88	86.67	6.99
	0.5	32	154	41	20.78	78.05	6.57
	0.3	47	154	87	30.52	54.02	7.39
8	0.7	27	147	31	18.37	87.10	6.45
	0.5	37	147	48	25.17	77.08	6.15
	0.3	51	147	90	34.69	56.67	6.84

肆、資訊類書籍關鍵字的關聯分析

一、簡介

隨著網際網路的發達，不出門即能輕鬆購物不再是幻想。網路書店也從美國的亞馬遜熱潮延燒到台灣。許多書商紛紛成立網路書局服務書友，以平實的價格和詳細的分類獲得顧客的支持。網路書局中為吸引顧客常以電子郵件通知讀者與會員新書資訊。但是一封提供數以千計新書資訊的電子郵件並無法引起顧客興趣，反而只會被當成垃圾郵件棄之不顧。網路書店要能善用顧客的歷史資訊，針對顧客投其所好，寄發個人化推薦的新書資訊，才是吸引顧客，提高回應率與購買率的良策 (葉怡成、王逸芸2006)。

本個案的主要目的是挖掘專業書籍的關鍵字之間的關聯規則，再利用這些規則產生個人化推薦。例如某顧客曾買過某書，該書含有關鍵字{Data mining, Business, Data processing}，自然我們應該為顧客推薦含有關鍵字Data mining、Business、Data processing的書。但假設存在下列關聯規則：

規則1：當一本書存在關鍵字｛Data mining, Business, Data processing}時，有80% 的機率會出現關鍵字｛SQL}

規則2：當一本書存在關鍵字｛SQL}時，有70% 的機率會出現關鍵字{Database}

規則3：當一本書存在關鍵字{Data mining,SQL}時，有60% 的機率會出現關鍵字{Machine learning}

顯然也可以推薦含有關鍵字SQL、Database、Machine learning的書。如此即可從顧客少量曾買過的書去擴大推薦他們可能有興趣的書。

為了探討關聯分析產生合理、有用的關聯規則之可能性，本研究以資訊管理系的資訊類課程的原文教科書為研究範圍。在本研究中，每一本書都被視為一筆「記錄」，每本書中的每個關鍵字視為一個「項目」。其研究進行步驟如下：

(1) 建立關鍵字資料庫：首先以資管系的資訊類課程名稱為關鍵字，例如「datamining」、「Information Management」、「Database management」，利用這些關鍵字到亞馬遜網站和大學圖書館搜尋書本，再將書裡的關鍵字逐一登錄。由於大多數的教科書至多只有兩、三項關鍵字，本研究依照書本題目以人工方式增加關鍵字，以獲得更有用的關聯規則。

(2) 關鍵字規格化：將收集到的關鍵字做正規化。由於關聯分析軟體不會區別大小寫和空格，例如「Data Mining」與「Data mining」會被視為兩種不同的關鍵字，需要統一規格。其他如關鍵字後面加了複數（如computers）也需要逐一統一規格，才能提升關聯分析的品質。經過整理有902筆資料(書籍)，479個項目(關鍵字)，所有項目(關鍵字)共計出現2140次，每個項目(關鍵字)平均出現4.5次(=2140/479)，每筆資料(書籍)平均含2.4個關鍵字(=2140/902)。

(3) 產生關聯規則：採用ARNN分析產生關聯規則，根據網路參數的設定得到關聯規則。

二、傳統關聯分析方法的結果

本文採用SQL server2005的關聯分析功能產生關聯規則。根據不同參數得到的關聯規則數，其中信賴度嘗試了30%, 40%, 50%, 60%, 70%，支持度嘗試了2, 3, 4, 5, 6，共25種組合。一般而言，當信賴度、支持度太低時，會出現太多的關聯規則，其中有許多不合理的規則存在；當信賴度、支持度太高時，又會出現太少的關聯規則，其中有許多合理的規則被濾掉。這與分類問題中，分類門檻的取捨會造成第一類錯誤與第二類錯誤的矛盾是一樣的，並不會有一個客觀的最佳門檻存在，而取決於第一類錯誤與第二類錯誤的相對嚴重性。對關聯規則而言，也不會有一個客觀的最佳信賴度、支持度存在，而取決於對關聯規則的品質(可靠性)與數量(完整性)的相對關心程度。如果要求關聯規則要有高品質(可靠性)，則可提高信賴度、支持度；反之，要求大數量(完整性)，則可降低之。經檢視各參數產生的關聯規則後發現，以信賴度=50%，支持度=3的情況下所獲得的90條關聯規則的品質(可靠性)與數量(完整性)較為均衡。因此，本研究採用此參數，並利用這些關聯規則可繪得「相依性網路圖」(圖6)。圖中箭頭是由關聯規則前項項目指向關聯規則後項項目。

由圖6可知，各種關鍵字的相互關聯可以分成19個「社群」(communities)：

1. 資料探勘社群：包含Data mining、SQL、Data processing、Business。
2. 人工智慧社群：包含Artificial intelligence、PROLOG、LISP。
3. 模糊系統與控制社群：包含Fuzzy systems、Automatic control、Control theory。
4. 統計社群：包含Statistics、Commercial statistics。
5. 電腦輔助教學社群：包含Computer-assisted instruction 、Internet in education、Education、Instructional systems。
6. 電腦繪圖社群：包含Computer animation、Three-dimensional display systems、Computer sound processing、Computer graphics。
7. 人機互動社群：包含Human-computer interaction、Virtual reality。
8. 生產管理與MIS社群：包含Management information systems、Information resources management、ERP、Production management。
9. 資料庫社群：包含Relational databases、Oracle。
10. 資料結構社群：包含Data structures、Pascal、Modula-2。
11. 電腦程式社群：包含Computer programming、C++、C、Electronic computers。
12. 軟體工程社群：包含Computer software、UML、Object-oriented methods、Computer science、Microprocessors、Computer systems。
13. 資訓安全社群：包含Computer security、Data encryption、Data protection。
14. 電腦網路社群：包含Computer networks、Data transmission systems。
15. 網際網路社群：包含World wide web、Client/server computing。
16. 網路應用社群：包含Web sites、Web publishing。
17. 微波電路社群：包含Microwave circuits、Radio circuits。
18. 無線通訊社群：包含Wireless communication systems、Microwave communication

systems、Mobile communication systems。

19. 作業系統社群：包含Operating systems、Linux。

以上結果大致上符合資訊方面的專業知識。由此可證，利用資料探勘中的關聯分析確實可以自動找出合理的關聯規則。網路書店以具有密切關聯性的社群來經營將有助於更有彈性地運用產品、定價、通路及推廣等行銷策略，以提升顧客滿意度(Wenger 2002)。

三、ARNN方法的結果

1. 學習循環對ARNN的影響

在此以表3的參數建構ARNN。圖7是檢出率、準確率對學習循環的折線圖。由此可以看出，在學習過程中，其檢出率、準確率確實同時得到很大的提升，並達到一個穩定的狀態。

表3：資訊類書籍關鍵字實例之ARNN的參數

參數	設定值
輸入層節點數 (Ninp)	479
隱藏層節點數 (Nhid1)	44
輸出層節點數 (Nout)	479
學習循環 (Ncycle)	10000
學習速率 (eta)	1.0

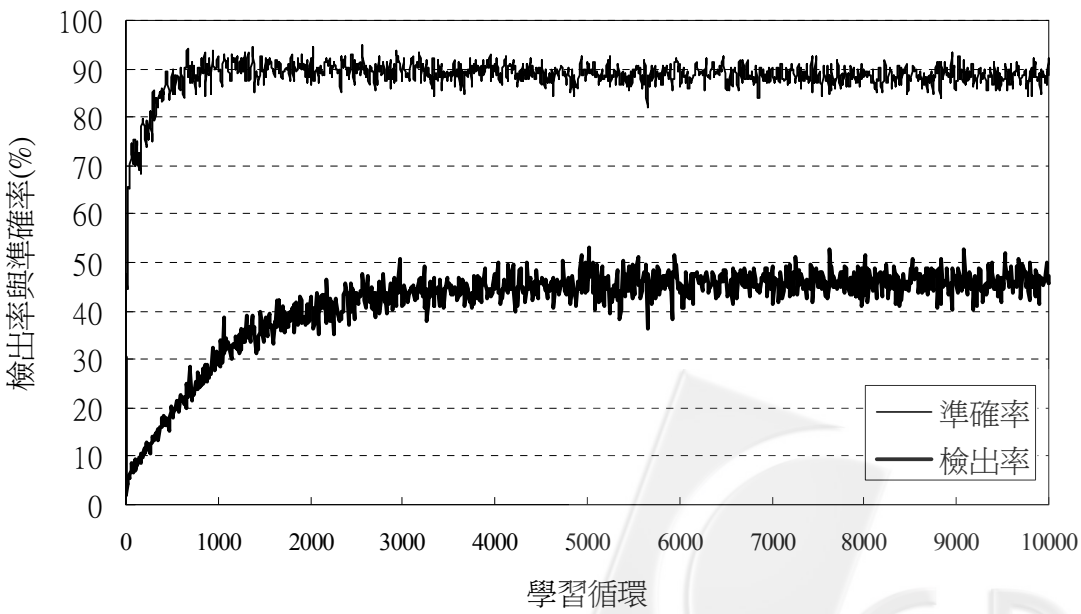


圖7：資訊類書籍關鍵字實例之ARNN的學習循環對檢出率和準確率的影響

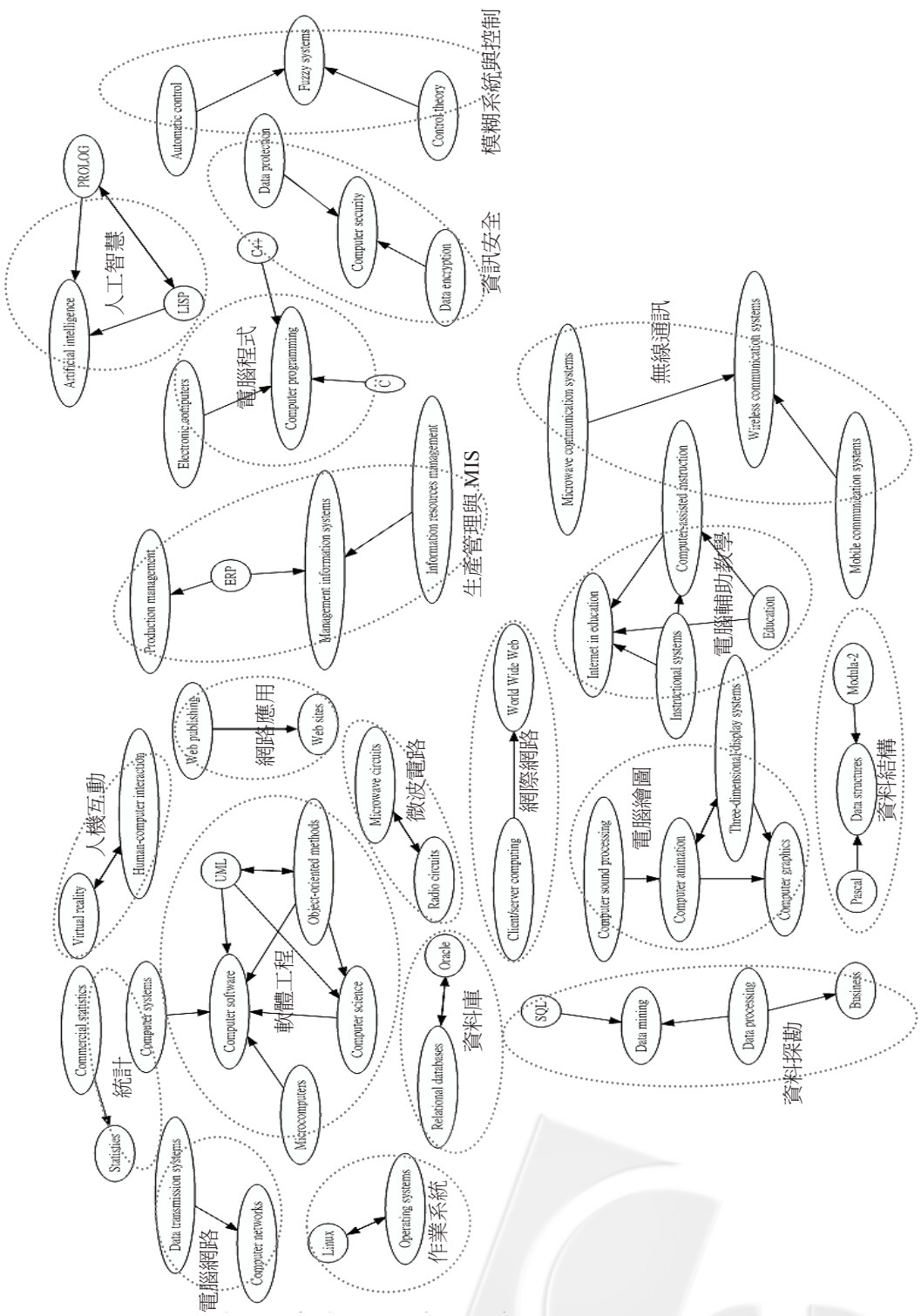


圖6：資訊類書籍關鍵字實例之傳統關聯分析相依性網路圖

2. 隱藏單元對ARNN的影響

從上一節的ARNN分析中，當隱藏單元數目為44個的時候，總共可以產生306條規則；而在此節，我們希望能排除掉支持度太小的規則，並希望藉由減少隱藏單元數目來達到目的。在此分別以2, 5, 10個隱藏單元重作一次，結果如表4。由表可知，當隱藏單元數減少時，檢出率會跟著變小，而準確率也有些微變小的傾向。其原因可能是當隱藏單元數目變小時，ARNN記憶規則的能力也變差，造成檢出率、準確率均變小。

雖然隱藏單元數目從44個降為5個時，檢出率、準確率、失誤率均較差，但對關聯分析而言，檢出率、準確率達100%、失誤率達0%並非代表模型找到完美的關聯規則，例如在超商的交易資料中，可能會出現一個同時買信封、信紙、指甲刀的交易，其中信封、信紙可能真有關聯，而{信封、信紙}與指甲刀應該只是巧合，這種巧合可以透過較高的支持度門檻加以排除。如果ARNN的檢出率、準確率要達100%、失誤率要達0%，則必然也要在ARNN的輸入端輸入{信封、信紙}時，輸出端必須輸出「指甲刀」。然而，顯然這不是關聯分析的目的。因此，表4中的檢出率、準確率、失誤率只是用來證明「當隱藏單元數目變小時，ARNN記憶規則的能力也變差」，但「記憶規則能力」並非越大越好，例如在使用倒傳遞網路建分類模型時，大量的隱藏單元數目與學習循環次數常導致「過度學習」現象(over-learning)，造成分類模型對訓練集裡的樣本很準確，對測試集裡的樣本不準確的結果，究其因，就是分類模型「記憶」了所有的訓練集樣本，而非「歸納」出隱藏在訓練集樣本中的分類規則。

同理，在使用ARNN建關聯模型時，大量的隱藏單元數目與學習循環次數也會導致「過度學習」現象。為了避免此現象，必須檢視其產生的關聯規則，評估其品質(可靠性)與數量(完整性)。結果發現，隱藏單元數目為44所產生的306條規則中有許多是不合理的，這些規則可能是「過度學習」的產物，將偶然出現在同一筆交易的項目關聯在一起；反之，藏單元數目為2只產生55條規則，只找出交易記錄中的部份關聯，可以說是「不足學習」(under-learning)的產物；而藏單元數目為5時產生142條規則，其關聯規則的品質(可靠性)與數量(完整性)較為均衡，可以說是「適度學習」(on-learning)的產物。

表4：資訊類書籍關鍵字實例之ARNN的隱藏單元數目對檢出率、準確率的影響

隱藏單元數目	分類門檻	應有且發現項目數(A)	應有項目數(B)	發現項目數(C)	檢出率%(A/B)	準確率%(A/C)	失誤率%
44	0.5	397	841	448	47.21	88.62	0.26
10	0.5	285	858	305	33.22	93.44	0.38
5	0.5	142	859	181	16.53	78.45	0.57
2	0.5	55	844	79	6.52	69.62	0.66

3. 法則門檻對ARNN的影響

如表5與圖8為不同的隱藏單元數目和不同的法則門檻值之下所產生的結果。從中可以發現關聯規則的數目會隨著隱藏單元數目的變小、和法則門檻值的變高而變小。簡言之，關聯規則的數目和隱藏單元數目成正比，而和法則門檻成反比。

表5：資訊類書籍關鍵字實例之不同隱藏單元數與法則門檻下產生規則數目統計

法則門檻 / 隱藏單元數目	0.2	0.3	0.4	0.5	0.6	0.7	0.8
44	451	455	418	398	364	328	306
10	424	398	370	349	325	308	288
5	304	281	253	231	212	190	156
2	102	93	89	85	80	70	53

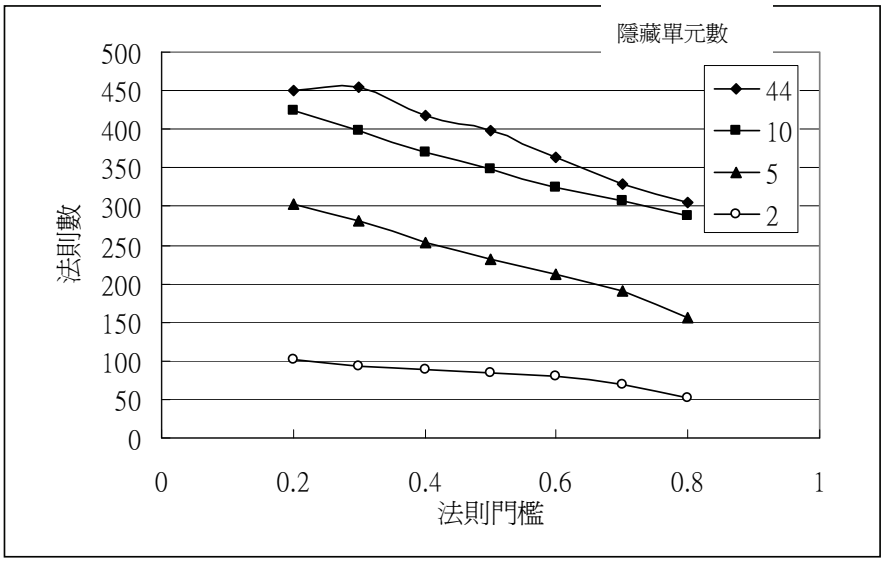


圖8：資訊類書籍關鍵字實例之不同隱藏單元數與法則門檻下產生規則數目統計

4. ARNN的關聯規則之解析

圖9為隱藏單元數為5所建立的ARNN，在法則門檻為0.5下所產生的231條關聯規則所繪的關聯圖，其細部圖見圖10。觀察圖9與圖10，可發現這些關鍵字可以組成21社群：

1.電腦視覺社群
2.基因演算法社群
3.網路程式社群
4.作業研究社群
5.電腦網路社群
6.無線通訊社群
7.多媒體社群
8.軟體工程社群
9.模糊集合社群

10.模糊系統與控制社群
11.電腦程式社群
12.電腦輔助教學社群
13.網際網路社群
14.電腦繪圖社群
15.資料探勘社群
16.電子商務與資安社群
17.人工智慧社群
18.神經網路社群
19.科技管理社群
20.生產管理與MIS社群
21.企業管理與統計社群

將這21個社群與之前傳統關聯分析所找出的19個社群相較下發現，ARNN沒有找出傳統關聯分析中的4個社群，分別為的資料結構社群、資料庫社群、微波電路社群、和作業系統社群，但ARNN有多找到下列8個社群：電腦視覺社群、基因演算法社群、網路程式社群、作業研究社群、多媒體社群、模糊集合社群、神經網路社群、科技管理社群。

仔細分析圖10可以發現一些有趣的關聯，例如在圖10的「17. 人工智慧社群」中有「Artificial Intelligece←Cognitive science」之關聯；而在圖10的「12. 電腦輔助教學社群」中有「Cognitive science→Human-computer-interaction」與「Human-computer-interaction←Distance-eduction」之關聯。可以發現，透過Cognitive science與Human-computer-interaction的串聯，使得感覺上相當遙遠的Artificial Intelligece與Distance-eduction有了合理的聯結，其關聯如圖11所示。

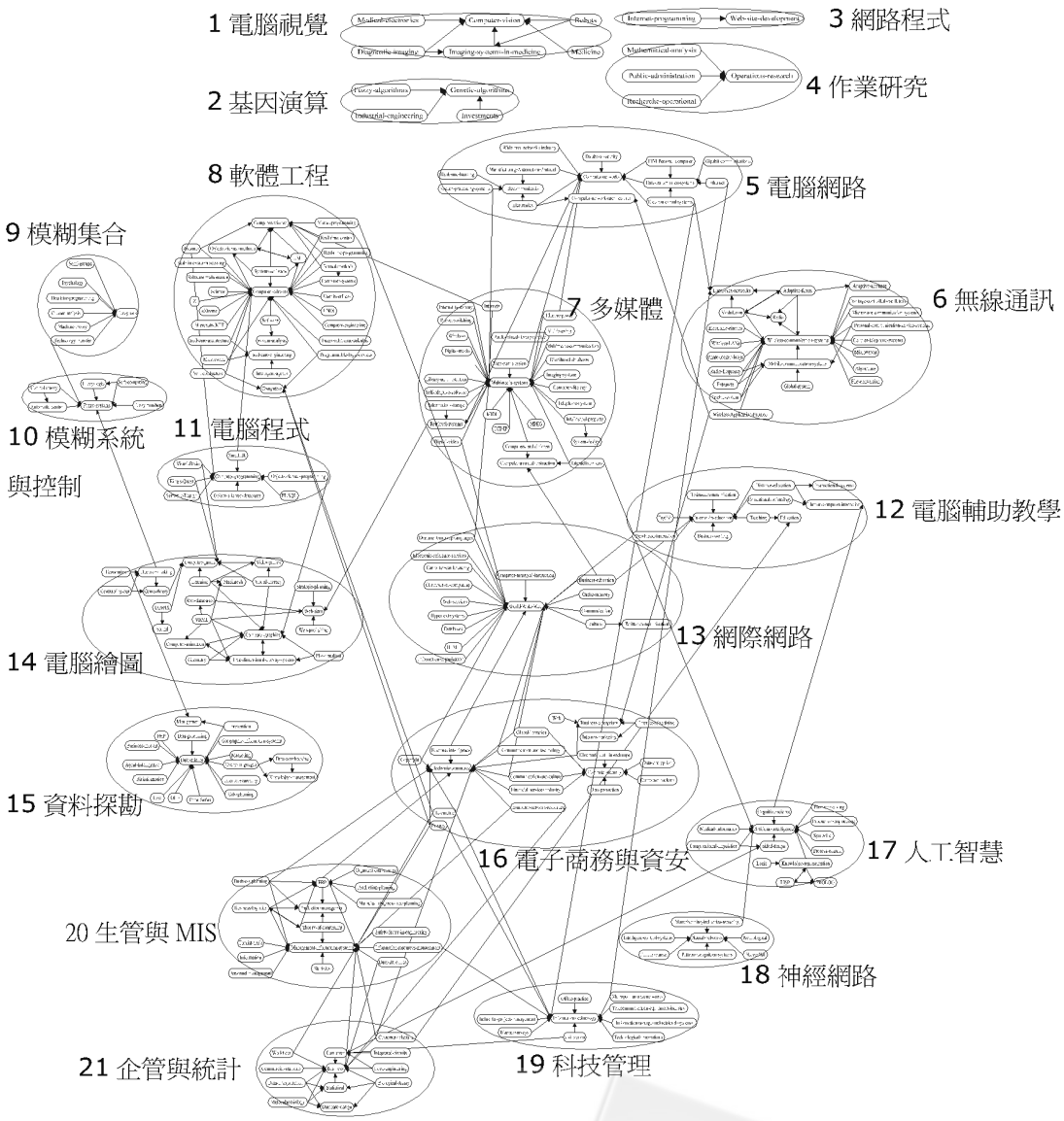
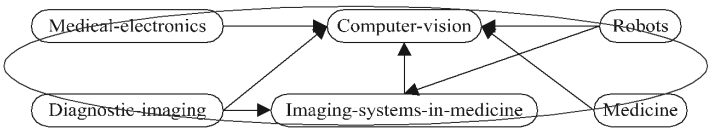
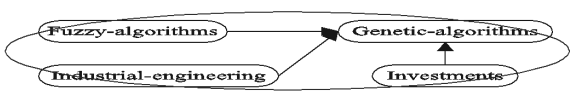


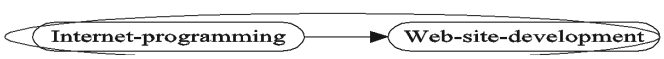
圖9：資訊類書籍關鍵字實例之關聯圖總覽圖



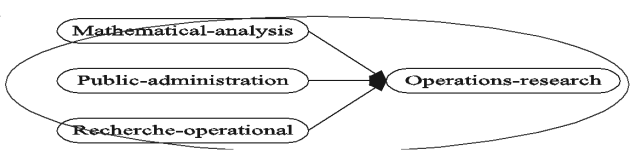
(1. 電腦視覺社群)



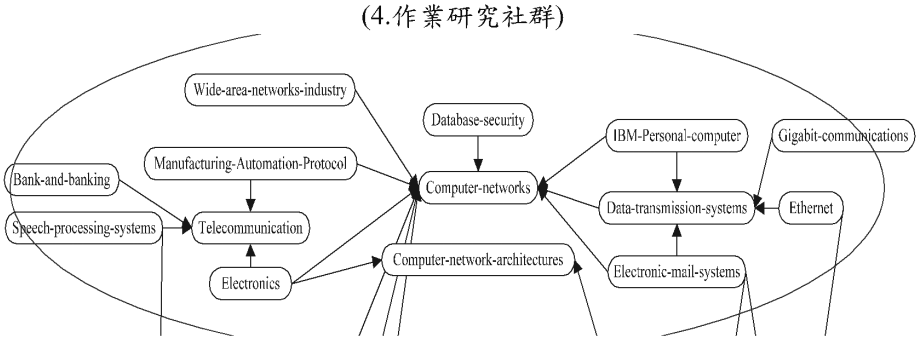
(2. 基因演算法社群)



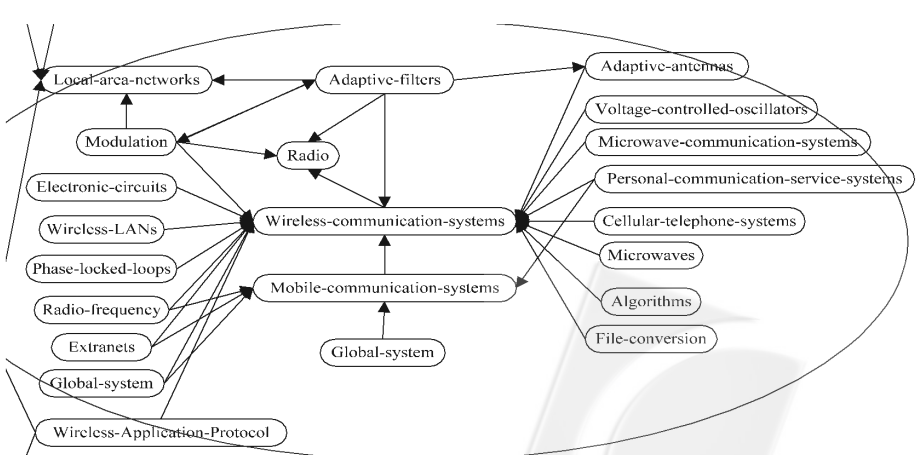
(3. 網路程式社群)



(4.作業研究社群)

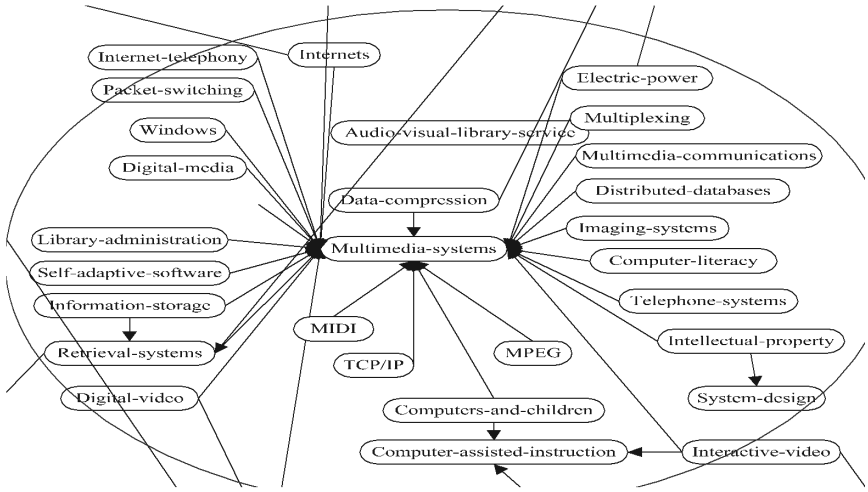


(5. 電腦網路社群)

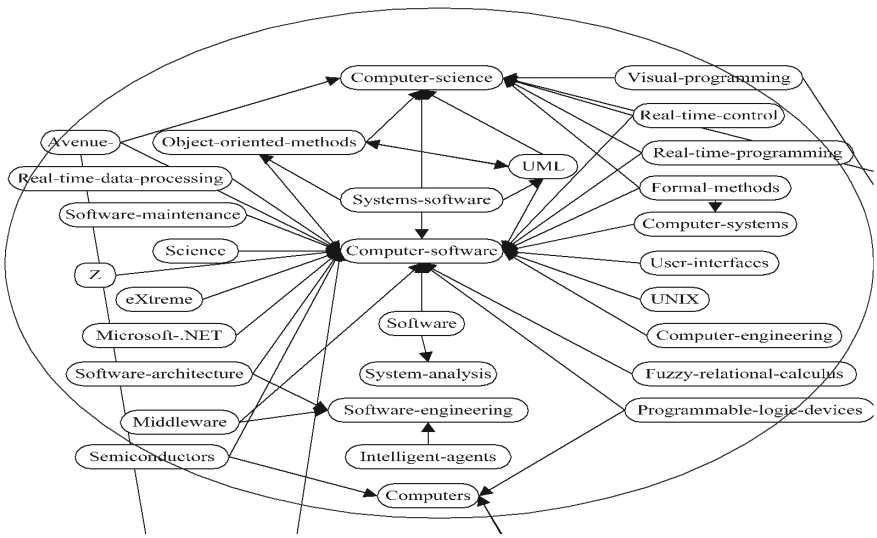


(6. 無線通訊社群)

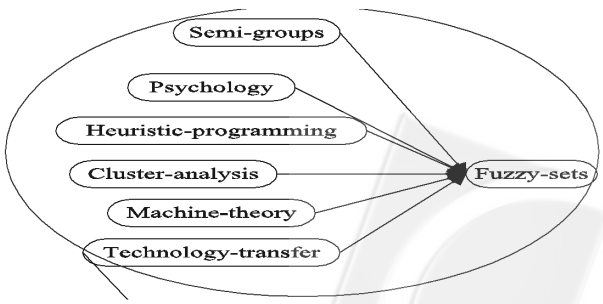
圖10：資訊類書籍關鍵字實例之關聯圖細部圖



(7.多媒體社群)

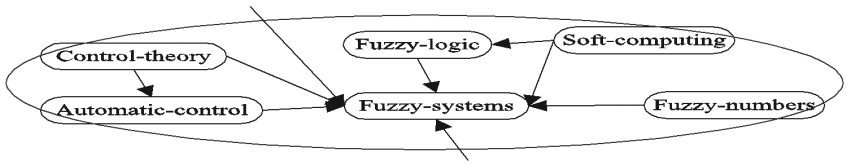


(8.軟體工程社群)

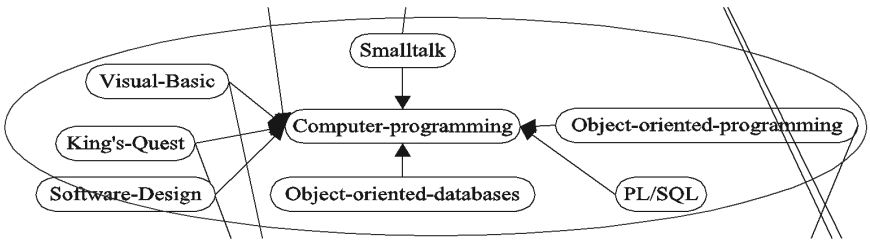


(9.模糊集合社群)

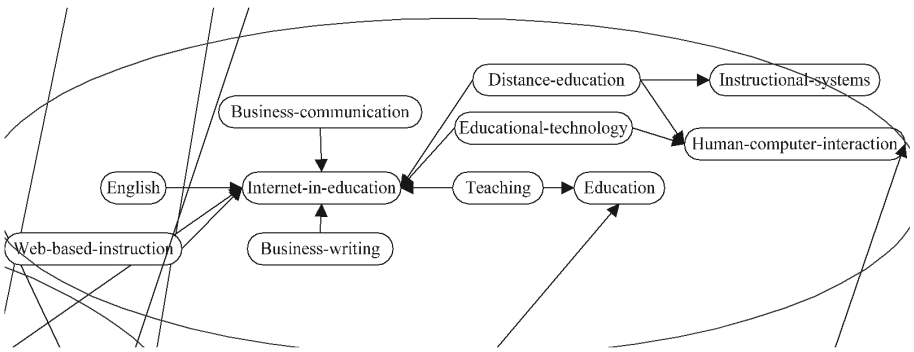
圖10：資訊類書籍關鍵字實例之關聯圖細部圖(續一)



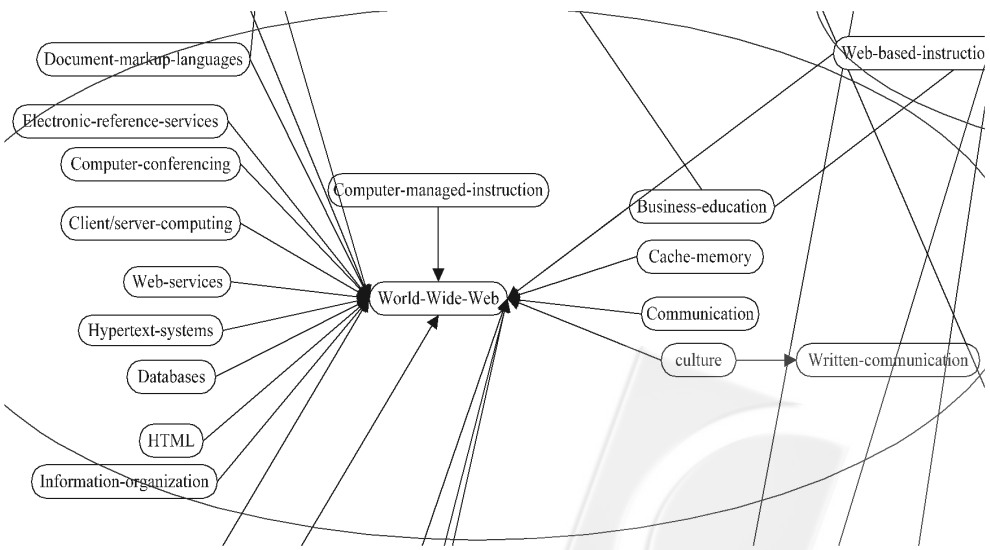
(10. 模糊系統與控制社群)



(11. 電腦程式社群)

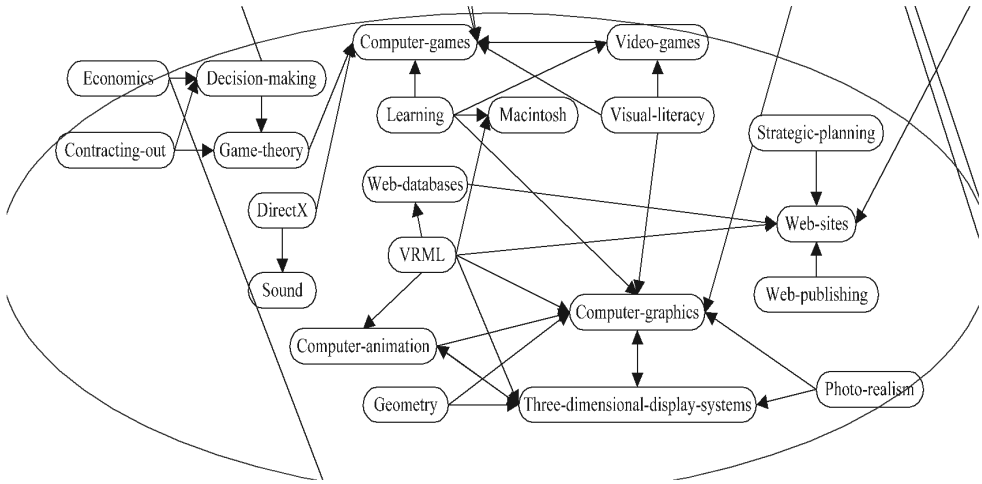


(12.電腦輔助教學社群)

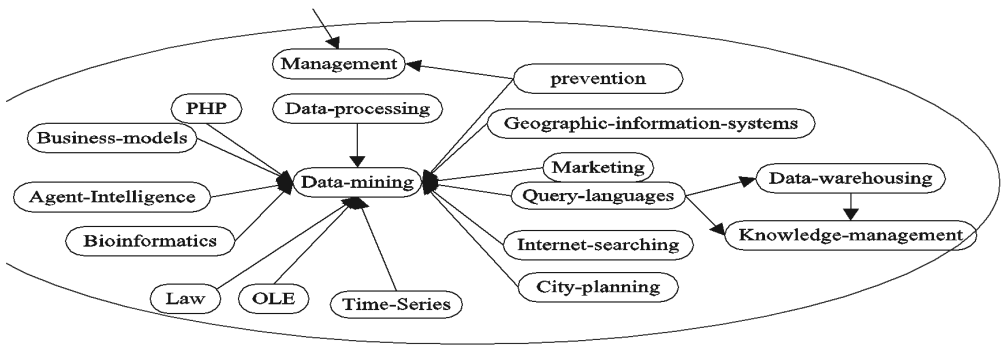


(13.網際網路社群)

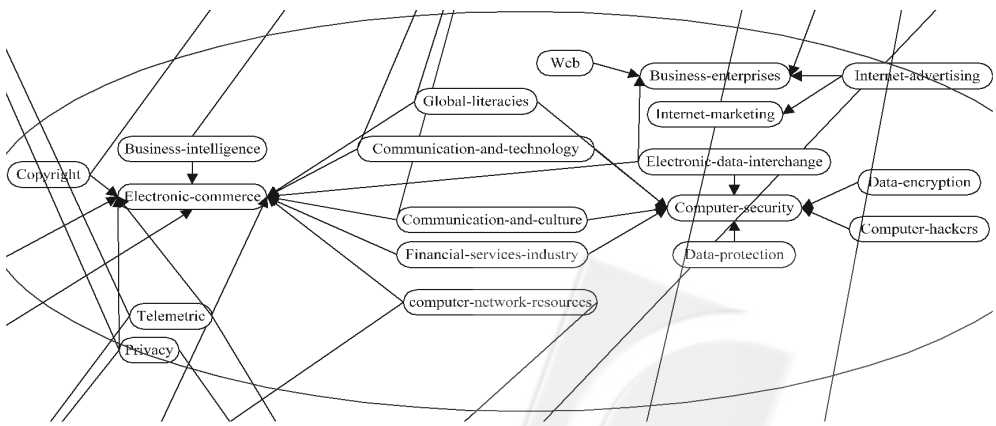
圖10：資訊類書籍關鍵字實例之關聯圖細部圖(續二)



(14. 電腦繪圖社群)

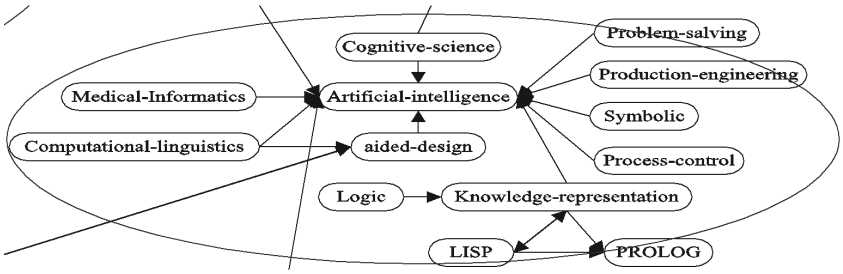


(15. 資料探勘社群)

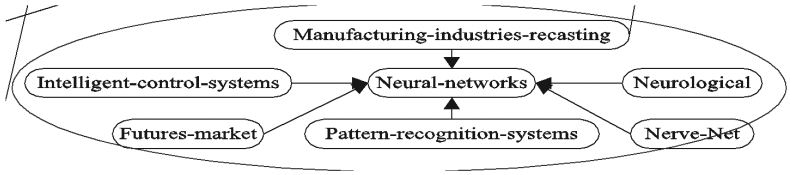


(16.電子商務與資訊安全社群)

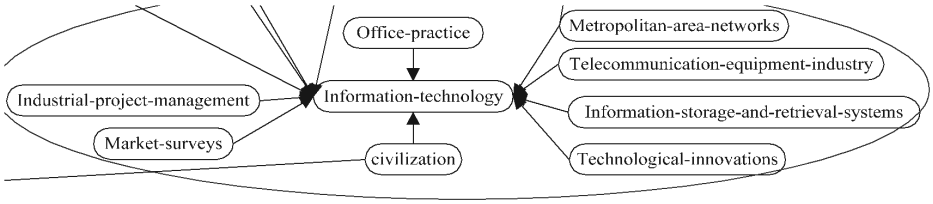
圖10：資訊類書籍關鍵字實例之關聯圖細部圖(續三)



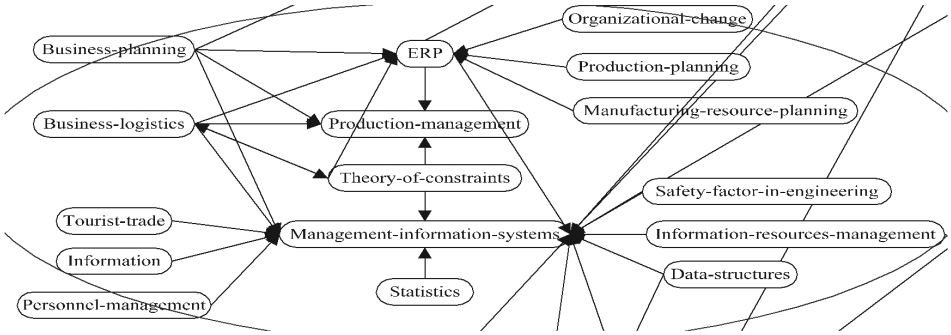
(17.人工智慧社群)



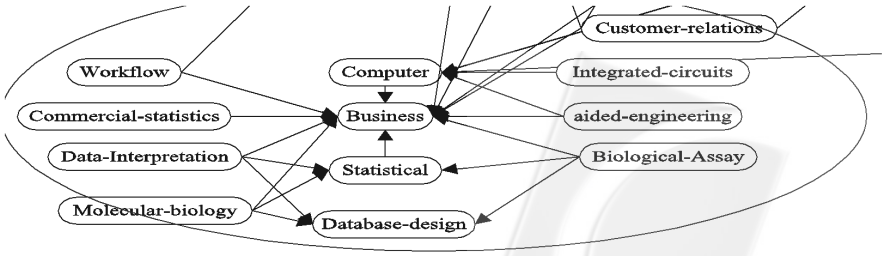
(18.神經網路社群)



(19.科技管理社群)



(20.生產管理與 MIS 社群)



(21.企業管理與統計社群)

圖10：資訊類書籍關鍵字實例之關聯圖細部圖(續四)

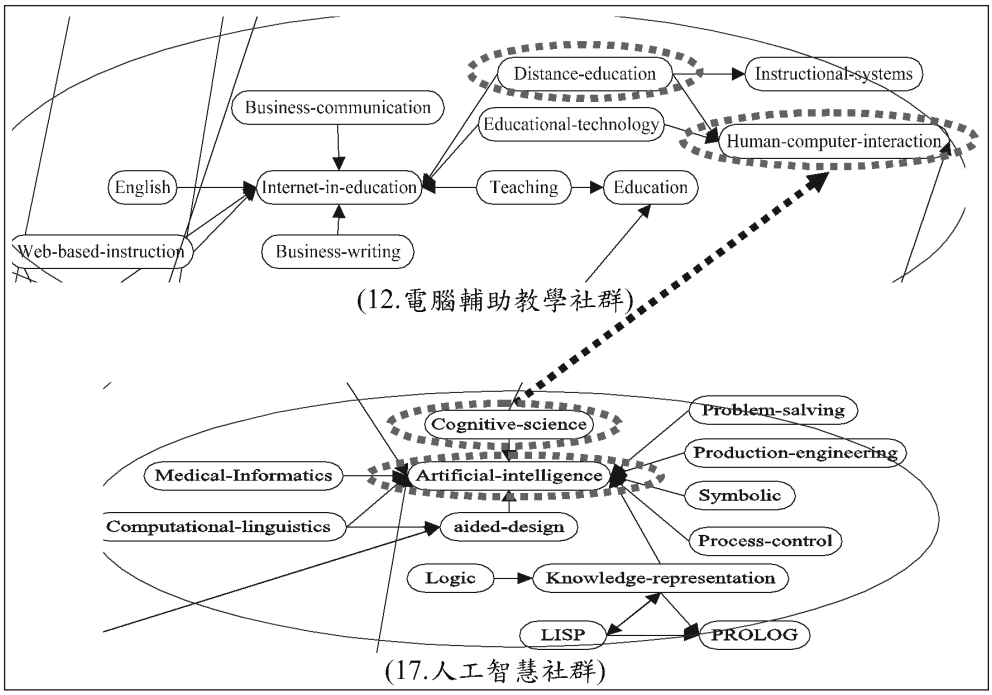


圖11：Artificial Intelligece與Distance-eduction之關聯圖

伍、結論

經由上述的例題結果可以歸納出以下結論：

1. ARNN的推論輸出值與資料中的信賴度大約相等。
2. 當ARNN的隱藏單元數減少時，檢出率、準確率會跟著變小。其原因可能是當隱藏單元數目變小時，ARNN記憶規則的能力也變差。
3. 當ARNN的隱藏單元數減少時，支持度低的關聯規則的推論輸出值會降低。
4. ARNN可以 (1) 利用提高法則門檻值來減少關聯規則的數目。(2) 利用減少隱藏單元數來抑制信賴度較低的關聯規則的產生。
5. ARNN產生的規則與由傳統關聯分析法所找到的規則重疊性很高，且能找出一些被傳統關聯分析法忽略的關聯規則。

ARNN所找出的部份關聯規則因其支持度較低，無法被傳統關聯分析所接受。一個支持度較低的large itemset，ARNN卻可以找出其規則，究其因可能是ARNN具有隱藏層的架構，一個輸入單元並不會直接「觸發」一個輸出單元，而是透過「觸發」部份隱藏單元，再由隱藏單元「觸發」輸出單元。例如在圖12中，假設A=>B有很強的關聯，則經過學習後，A輸入單元可能透過某隱藏單元去影響B輸出單元，形成圖中黑色粗線的連結；又假設B=>C有很強的關聯，則B輸入單元可能透過某隱藏單元去影響C輸出單元，形成圖中灰色粗線的連結。因為要記憶的關聯規則可能多於隱藏單元數，上述兩條關聯

規則可能共用同一個隱藏單元(圖12中的1號隱藏單元)。因此，當A輸入單元「觸發」該隱藏單元時，因為該隱藏單元與C輸出單元有很強的連結強度(大的網路權值)，可能也觸發了C輸出單元，因而產生了A=>C的關聯。

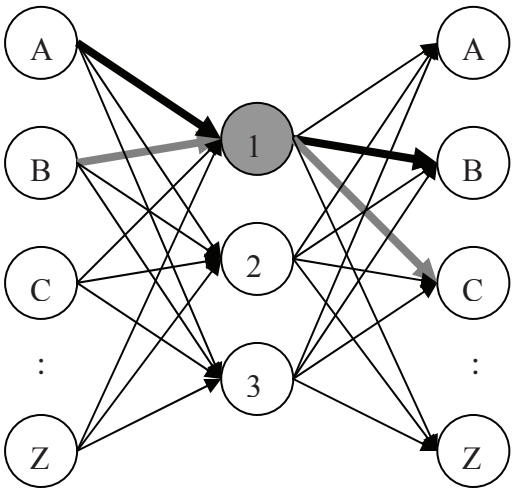


圖12：A=>B，B=>C與A=>C在ARNN上的解釋

未來研究的建議如下：

1. 由於神經網路本身即可接受0到1之間的數值作為輸入，因此ARNN亦有此能力。但這並不在本研究的範疇之內，是未來可以研究的方向。
2. 雖然ARNN有能力產生二階以上的關聯規則，但本研究大都探討一階關聯規則以及和傳統關聯之間的比較，未來可探討二階以上的關聯規則。
3. 能夠預測個案的關聯項目是ARNN的特色，因此未來的研究可以往這方面的應用去發展或更深入的探討。例如探討是否可以針對一位已購買過數本書的讀者，根據其書的關鍵字，預測其關聯的關鍵字，以向此讀者推薦他可能有興趣的書籍。

誌謝

本研究承蒙國科會贊助(計畫編號96-2221-E-216-032)，特此致謝。

參考文獻

1. 王錫中，2002，運用關聯法則技於產品開發設計之研究，元智大學工業工程與管理研究所碩士論文。
2. 林瑞山，2003，類神經網路於預測晶圓測試良率之應用，國立成功大學工學院工程

管理碩士專班碩士論文，。
3. 陳建銘，2001，類神經網路於Web Mining之應用，國立台北技大學商業自動化與管理研究所碩士論文。
4. 黃南傑，2004，高效率解之關聯規則探勘，南台科技大學資訊管理學系碩士論文。
5. 葉怡成，2006，應用類神經網路，台北：儒林圖書有限公司。
6. 葉怡成、王逸芸，2006，『以關聯探勘分析資訊類教科書關鍵字之關聯』，2006年資訊管理暨電子商務經營管理研討會，中華大學，新竹市。
7. 劉向陽、王如雲，2006，『基於新型三層誤差反向傳播網路的圖像壓縮』，電腦工程與應用，第42卷．第13期：33～35頁。
8. Abdel-Wahhab, O. and Fahmy, M. M.“Image Compression Using Multilayer Neural Networks,”*IEE Proceedings of Vision, Image and Signal Processing* (144:5) 1997, pp:307-312.
9. Abidi, M.A., Yasuki, S., and Crilly, P.B.“Image Compression Using Hybrid Neural Networks Combining the Auto-associative Multi-layer Perceptron and the Self-organizing Feature Map,”*IEEE Transactions on Consumer Electronics* (40:4) 1994, pp:796-811.
10. Agrawal, R., Imielinski, T. and Swami, A.“Mining Association Rules between Sets of Iems in Large Databases,”*Proceedings of the ACM SIGMOD Conference on Management of Data,* 1993, pp:207-216.
11. Agrawal, R. and Srikant, R.,“Fast Algorithms for Mining Association Rules,”*Proceedings of the 20th VLDB Conference, Santiago,* Chile, 1994, pp:487-499.
12. Arozullah, M. and Namphol, A.“A Data Compression System Using Neural Network Based Architecture,”IJCNN 1990 *International Joint Conference on Neural Networks,* Washington, DC, 1990, pp:531-536.
13. Cai, C. H., Fu, W. C., Cheng, C. H. and Kwong, W. W.“Mining Association Rules with Weighted Items,”*The International Database Engineering and Applications Symposium,* 1998, pp:68-77.
14. Chan, K. C. C. and Au, W. H.“Mining Fuzzy Association Rules,”*The Sixth ACM International Conference on Information and Knowledge Management,* Las Vegas, Nevada, 1997, pp:10-14.
15. Hipp, J., Güntzer, U. and Nakhaeizadeh, G.“Algorithms for Association Rule Mining - A General Survey and Comparison,”*ACM SIGKDD Explorations* (2:1) 2000, pp:58-63.
16. Huang, S.J., Koh, S.N., and Tang, H.K.“Image Data Compression and Generalization Capabilities of Back-propagation and Recirculation Networks,”*IEEE 1991 International Sympoisum on Circuits and Systems,* Singapore (3) 1991, pp:1613-1616.
17. Liu, B., Hsu, W. and Ma, Y.“Mining Association Rules with Multiple Minimum Supports,”*Proceedings of 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* San Diego 1999, pp:337-341.
18. Paik, J.K. and Katsaggelos, A.K.“ Image Restoration Using a Modified Hopfield

Network,” *IEEE Transactions on Image Processing* (1:1) 1992 pp:49-63.

19. Sarawagi, S., Thomas, S., and Agrawal, R. “Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications,” *Data Mining and Knowledge Discovery* (4:2) 2000, pp:89-125.
20. Setiono R. and Lu, G. “A Neural Network Construction Algorithm with Application to Image Compression,” *Neural Computing & Applications* (2:2) 1994, pp:61-68.
21. Setiono, R. and Lu G.. “Image Compression Using a Feedforward Neural Network,” *IEEE 1994 International Conference on Neural Networks* (7) 1994, pp:4761-4765.
22. Sonehara, N., Kawato, M., Miyake, S., and Nakane, K. “Image Data Compression Using a Neural Network Model,” *IJCNN 1989 International Joint Conference on Neural Networks, Washington,* DC, (2) 1989, pp:35-41.
23. Wang, L. Y. and Oja, E. “Image Compression by Neural Networks: A Comparison Study,” *IEEE 1993 Winter Workshop on Nonlinear Digital Signal Processing,* Finland (7:2) 1993, pp:1-6.
24. Wenger, E., McDermott R., and Snyder, W.M. *Cultivating Communities of Practice*. Harvard Business School Press, Boston, 2002.
25. Zheng, Z., Kohavi, R., and Mason, L. “Real World Performance of Association Rule Algorithms,” *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining,* San Francisco, 2001, pp:401-406.