第二章　文獻探討

隨著電腦技術的進步，使得人們可以處理的資料量快速增加，傳統的統計方法與分析技術已不敷使用。因此，資料探勘技術的發展，提供我們一個可以處理大量資料並分析其中所隱含有用資訊的方法。同樣的，網路技術的進步與普及，也使得網路使用者與日遽增，網路上蘊藏無限的商機，也使得電子商務與電子資訊化服務的應用也愈來愈多，致使各式各樣交易資料量快速的增加且資料型態變的更為複雜。所以，近年來，不少的研究試著將資料探勘的技術用來分析網路行為產生的資料。在2.1 節中，我們首先會介紹與本篇論文相關的背景知識，可以幫助我們發掘網路使用者行為的Web Mining 技術與相關的背景、概念，而在2.2~2.4 節中，我們將針對與本篇論文相關的研究理論基礎作整理，在2.5 節中，我們會介紹與本篇論文的相關研究。

2.1 網頁探勘

網頁探勘是資料探勘中的一個技術領域，隨著電腦技術的進步，使得人們可以處理的資料量快速增加，傳統的統計方法與分析技術已不敷使用。因此，資料挖掘技術的發展，提供我們一個可以處理大量資料並分析其中所隱含有用資訊的方法。同樣的，網路技術的進步與普及，也使得網路使用者與日遽增，網路上蘊藏無限的商機，也使得電子商務與電子資訊化服務的應用也愈來愈多，致使各式各樣交易資料量快速的增加且資料型態變的更為複雜。然而，許多有用的資訊往往隱藏在繁忙的網路行為中，若能夠加以有效的分析、瞭解及運用將可以幫助經營決策者取得較佳的優勢地位。

由於網路使用的日益頻繁，網路使用者數量的增加，使得網路上各式各樣的交易資料量大增，而來自使用者個人的獨特性與不確定性等因素，使得資料變的更為複雜且不容易利用傳統統計分析方法進行分析。因此，許多研究都希望藉由使用資料探勘的技術，分析網路上大量看似無意義且多餘的資料，從中取得隱藏在這些龐大的交易資料中有價值及有用的資訊，分析使用者共同的行為模式以及瞭解使用者未來的行為趨勢，並藉此增進網路使用的效能，幫助解決受限於網路頻寬造成的網路壅塞現象，並提供電子商務網站行銷與決策方向的依據，或給予網站管理者在管理與設計網站的指導原則，並期望藉由此一分析使得網站運作的整體效能能夠有效的提升，更進一步得到商業上實質的獲利。圖2-1 是一個Web Mining 系統的流程結構圖。

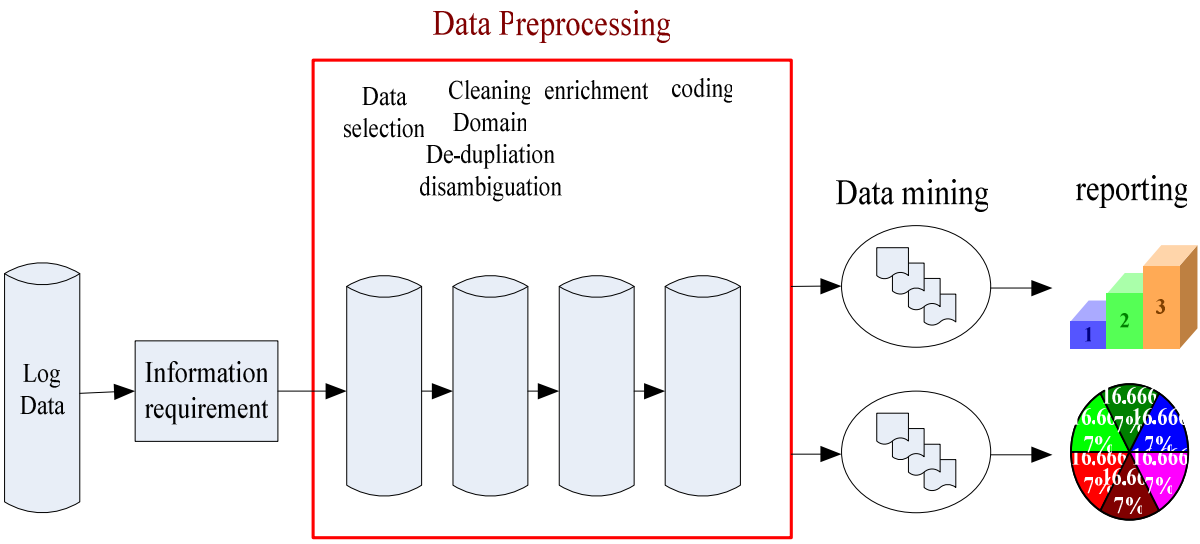

圖2-1 Web Mining 系統流程結構圖

(資料來源:修改自James Pitkow, Peter Pirolli)

一般而言， Web Mining 一詞是在1996 時，由Oren Etzioni[9]在印度提出， Web Mining 就是利用資料探勘技術，從全球資訊網的文件與所提

供的服務中挖掘或萃取有用的資訊。大體而言，Web Mining上分析的資料型態多不會超過這些範圍，也就是說，Web Mining 是依照研究者的需求所在，選擇所要分析的資料類型，進行分析，得到研究者可能會有興趣的資訊另外，機器學習、統計分析、資料挖掘及模式識別等都廣泛的應用在Web Mining中。

而在今日，普遍提到Web Mining 的分類，指的是將Web Mining 分為以下三個類型[10]:

1. Web Content Mining

Web Content Mining 主要是將資料採掘的技術應用在網站頁面的內容上，挖掘出有意義的資訊，和Web Usage Mining 與Web Structure Mining 不同的地方在於Web Content Mining 主要是將資料挖掘的技術用在分析網頁的主要內容上，而Web Usage Mining 與Web Structure Mining 則是強調網頁與網頁間的關聯性。

2. Web Structure Mining

Web Structure Mining 主要是根據網頁內容的特性所構成的連結架構及超連結，總結出網站與網頁的結構關係，也就是藉由蒐集網頁與網頁之間的相關資訊。某一頁面與直接連結到的頁面，或者與互為鄰居的頁面之間的關係所顯現出來的特性，來挖掘有趣且有意義的資訊的事實，依據此結果來描述網頁之間的連結性，並藉此協助網路管理者作為適時調整網頁架構的一大考量。一般說來，Web Structure Mining 主要分為三個方向與目標：

第一個方向是調整網站整體架構，使得網站內的頁面與網站本身形成一個結構化的形式。與Web Content Mining 的不同處在於，Web Structure Mining 是要找出網頁之間和網站架構連結之間的結

構性，也就是說，我們可以藉由Web Structure Mining 可以將網站內頁面加以分類，並藉以得知網頁間相似度的關係。

第二個方向是發掘網路文件本身的結構性。我們知道當我們分析網頁資料時，如果資料是結構性的將有利於我們的分析，然而如我們前面所提的，大部分的網頁資料其實是呈現非結構性的，因此，如果我們可以找到網頁文件的結構性，便可以用來觀察比較整合網頁的架構，而這樣有結構性的資料，可以方便將資料庫的技術用在資料挖掘上，並藉由索引的方式去存取及分析資料。

第三個方向則是我們希望藉由分析、瞭解網站因為某一範疇的特性所造成的結構性，以及此一特性造成的網路超連結的結構。這可以幫助我們根據這樣的特徵，將網站內容依此一範疇的資訊歸結出一個獨一無二的流向結構，如此一來，當使用者對網站進行查詢動作時，將可以更容易、更有效的獲得有用且有意義的資訊。

3. Web Usage Mining

通常在使用者瀏覽網站的過程中，會留下參觀網站網頁過程的記錄(Log Data)，其實這樣的資料內含有許多有用的資訊，只是一般從網站萃取資訊時，都會把焦點放在網頁內容部分，而忽略了由Log Data 中發掘的使用者瀏覽路徑也是可作為分析重要的指標，我們不但可以根據使用者的Log Data 瞭解該使用者的興趣所在，根據此特性為該使用者提供獨特的個人推薦與服務機制，也可以瞭解到網站的建構是否符合實際上使用者的需求，以提供網站管理者進行維護管理的指標。

總結可歸類出幾個主要的目標：

1. 預測某一個網站使用者的行為。
2. 比較網站期望與實際上的效能是否相同，因為網站規劃者在規劃設計網站之初的構想，未必會符合實際上網路使用者瀏覽網站的需求和興趣，因此，藉由分析log data 瞭解網站使用實際的狀況，可以幫助設計與管理者將該網站的效益大大的提升。
3. 根據使用者的興趣調整網站，如果能夠瞭解網路使用者的獨特性，提 供個人化的服務，一定會提升網站的效能。

2.2 資料倉儲系統

資料探勘的演進，自 1960 年代電腦發明以來，利用電腦進行資料的管理即是電腦最主要的應用之一。資料探勘系統的演進可分為四個階段[11]：檔案系統、資料庫系統、資料倉儲、以及資料探勘系統。隨著電腦科技的進步，每一個階段都扮演著下一個階段的基礎，其所使用的技術、系統特性如表 2-1 所示。

表 2-1 資料探勘的演進過程

演進步驟	應用技術	系統特性
檔案系統 (1960 年代)	電腦、磁帶、磁碟	傳遞歷史性的靜態資料
資料庫系統 (1970 年代)	階層式資料庫（Hierarchical）、網路式資料庫(Network Database)、關聯式資料庫(Relational Database)、結構化查詢語言(SQL)、開放性資料庫聯結協定(ODBC)	傳遞即時性的單層次動態資料
資料倉儲系統 (1990 年代)	線上分析處理(OLAP)、多維度資料模型(Multidimensional Data Model)、資料倉儲(Data Warehouse)	傳遞歷史性的單層次動態資料
資料探勘系統 (現今)	進階演算法、多處理器電腦系統、大量資料儲存技術、人工智慧	傳遞預知的、鑑往知來的資訊

(資料來源:本研究自行整理)

資料庫的概念是起源於共享共用的資料資源，是為了滿足多使用者的資訊需
求而設計，用以載入資料、取得程式和使用者所需求的資料、格式化所擷取的資料以符合程式或使用者所期望的形式，以及隱藏某些資料不被特定的使用者存取及更新。

資料倉儲是利用儲存大量歷史資料的資料庫，提供匯總或是統計的資訊，以支援決策的使用，而資料倉儲的誕生即是為了所關切的決策問題。建構過程首先是收集資料，經過資料清理、資料轉換、資料整合、資料載入和定期資料更新，最後便能建置一套資料倉儲系統。而資料倉儲基本上只是一個存放大量匯總資料的後端儲存體，還必須配合前端的運用才能顯示出它的價值，而線上分析處理(OLAP：On-Line Analytical Processing)與資料探勘則是兩個最常使用的應用，資料探勘與線上分析處理的不同之處，在於線上分析處理主要是原本的呈現出使用者查詢的結果，而結果的解讀將由使用者自行加以判斷。資料探勘則能夠進一步利用統計、機器學習等方法將資料再分析，探勘出新且有用的知識，在資料的運用更勝於資料倉儲。

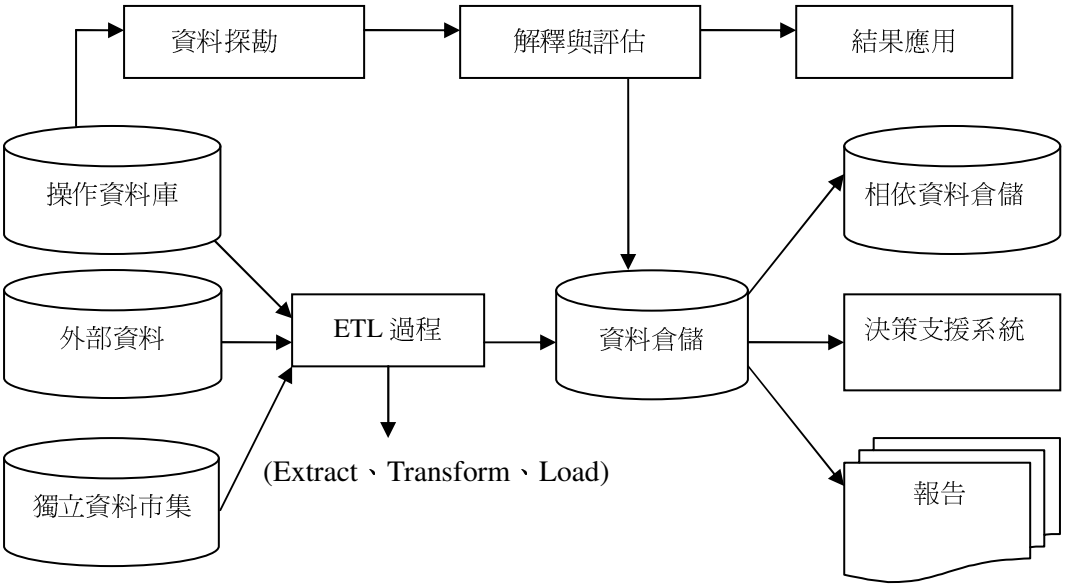

圖2-2 資料倉儲系統流程圖

(資料來源:本研究自行整理)

資料倉儲處理模組如圖2-2所示，蒐集操作資料庫、外部資料、獨立資料市集，組合這些所蒐集的資料來分析，接著將這些資料丟到資料探勘的工具，解釋結果與評估預測，對新的問題或狀況產生時，運用資料倉儲的結果得到支援決策系統得以解決問題。

在資料進入資料倉儲之前，還必須經過ETL〔擷取(Extract)、轉換(Transform)、載入(Load)〕。ETL的主要任務就是處理這些問題：從一個或多個輸入來源中提取資料，改正或轉換所擷取的資料為所需求的格式，載入資料到倉儲之中。儲存在資料倉儲內部的用戶提供統一、協調和整合式的資訊環境，支援決策過程和作深入的綜合分析，資料能從操作型環境中轉換到只有對決策支援有用的資料，才會從操作環境中擷取出來並存到倉儲資料庫中。而資料倉儲與資料探勘之間有著密切關係，將資料探勘擴充

到它的資料倉儲系統環境中，可以增強決策支援能力，資料探勘是建立在資料倉儲上的決策支援技術。

而資料探勘的目的是從大量的交易資料中擷取潛在有用的資訊與知識，對產
品之行銷分析可提供企業非常有用的資訊。而在整個資料探勘的過程中，其實也是一直不斷地重複以下幾個步驟：

1. 辨識正確的問題

   問題的定義是成功的資料探勘裡，最重要的部份，因為如果問題的定義不明確，將導致探勘結果不符合使用者需求。

2. 資料選取

   自資料庫或資料倉儲中選出資料。包括過濾掉不必要的資料與無法探勘確認的資料。

3. 將資料轉換到可以採取行動的結果

   這個階段是資料探勘的核心，其目的是要建立一個最佳的探勘模型和可執行的探勘結果。

4. 解釋結果與採取行動

   資料探勘的目的是要對結果採取行動，如果我們不依照模型之結果來行事，則資料探勘便沒有提供任何效果。

5. 評估結果

   評估結果是資料探勘循環步驟的最後階段，這些評估將會提出更多問題及資料，以便後續資料探勘的進行。資料探勘的技術在最近幾年被廣泛地研究，大量的資料探勘方法已被成功地發展出來。

2.3 Petri-net

Petri-net理論的發展源自1962年Dr.Petri的博士論文，Petri-net為一具有圖形特性與數學理論基礎的系統建構工具[12]，藉由分析完成所建構出來的Petri-net，不僅能表示系統的同步與互斥行為，而且可以對系統做定性與定量的分析。此外隨著系統複雜度的增加，基本的Petri-net也陸續延伸出許多更富有特性的高階Petri-net理論，以其延伸的特性來加強對系統塑模的能力。Petri-net可以建立系統的狀態方程式、代數方程式以及其他可以管理系統行為的數學模組。

Petri-net是一個具有方向性的加權圖形，其構成元素包含兩種節點所組成：一種稱之為位置節點（Place），另一種稱之為轉換節點（Transition），狀態節點在圖形的呈現上以圓形表示，轉移節點則以長條形或方形表示。而狀態節點與轉移節點之間，則由方向性連線（Arc）的方式直接連接，用來表示節點之間的關係，在箭頭上可以標示其加權值，表示起動轉移節點所需的標記（Token）數量。另外Petri-net也同時包含標記的概念標記被配置在狀態節點中，以黑色的圓點來表示，一個派翠網路是由五個部分所組成(5-tuple)， PN=(P,T,F,W_0)正式的定義如下[13]：

$P= \{p_1, p_2,..., p_m\}$ 是一組有限(finite)的places 集合(set)；

$T= \{t_1, t_2,..., t_n\}$ 是一組有限的transitions 集合；

$F \subseteq (P \times T) \cup (T \times P)$ 是一組arcs 集合（set）（流程關係）；

$W：F \rightarrow \{1,2,3,...\}$ 是一個權重函數（weight function）；

$M_0：P \rightarrow \{0,1,2,3,...\}$ 初始(initial) P_N呈現的（marking）；

I為輸入函數，I : T→P∞或I : P→T∞，表示由轉變輸入位置的集合或由位置輸入轉變的集合；O 為輸出函數O：T→P∞或O：P→T∞，表示由轉變輸出位置的集合或由位置輸出轉變的集合；(P∞為位置群組；T∞為轉變群

組)。其圖形呈現結果如圖2-3.，傳統Petri-net之簡單數學定義如圖2-3及表2-2 所示[20]。

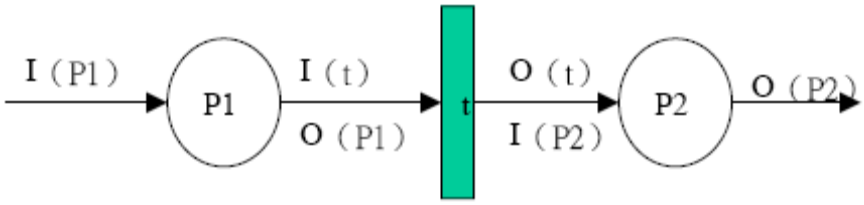

圖 2-3 Petri-net 基本架構圖

(資料來源:參考M. Crovella and A. Bestavros)

表2-2 Petri-net的數學定義

$P= \{p_1, p_2,..., p_m\}$	是一組有限的places 集合
$T= \{t_1, t_2,..., t_n\}$	是一組有限的transitions 集合
$F \subseteq (P \times T) \cup (T \times P)$	是一組arcs 流程關係集合
$W：F \rightarrow \{1,2,3,...\}$	是一個權重函數
$M：P \rightarrow \{0,1,2,3,...\}$	初始P_N 呈現的 marking
其中$P \cap T=\varnothing$ and $P \cup T \neq \varnothing$。 一個不含任何初始marking 的Petri-net 定義為PN=(P,T,F,W)。 一個含有給定的初始marking 的Petri-net 定義為(N, M_0)。	

(資料來源:參考M. Crovella and A. Bestavros)

而Petri-net特性可分做兩大類：結構特性、行為特性，一般研究中常探討的特性包括可達性、限制性、活性，細項說明如表2-3[14]。

表2-3 Petri-net性質彙整表

提出者	**做法**	**目的**
Buhler& Vidal(2003)	BPEL4WS 標籤和 Petri-net模型的 轉換方式。	在多代理人的工作流程系統中建立網路服務和網路服務之間的

		關係。
Hamadi& Benatallah(2003)	利用Petri-net塑模網路服務組合。	確保網路服務組合的可靠性，方便建立複雜的網路服務組合，提高網路服務的再利用性。
Yi& Kochut(2004)	以Color Petri-net為基礎的網路服務組合設計及驗證架構。透過BPEL4WS 文件與CP-net 模型的轉換	在組合服務設計階段透過驗證盡早發現並更正網路服務組合的錯誤。

(資料來源:參考戚玉樑碩士論文)

Petri-net模型以數學矩陣的方式表達,再透過方程式的計算來得知狀態的可達性。首先將Petri-net模型以輸入矩陣D^-和輸出矩陣D^+ 的方式表達，輸入矩陣和輸出矩陣後再以$D= D^+-D^-$ 運算式求得代表整個Petri-net系統模型的矩陣D。Petri-net利用矩陣的表達和方程式的運算即可利用預測的方式來獲知兩狀態之間是否可以彼此轉移，無論兩狀態是否為相鄰的關係。此外對於較複雜的系統採用矩陣方式也能夠輕易的表達。

Petri-net如同圖形化的工具一樣，可以被運用在視覺傳達的教學工具上，類似於流程圖、方塊示意圖及網路。除標記（Token）可以在Petri-net上運用於模擬動態及同步動作特性外，就數學化工具而言，Petri-net也可以建立狀態方程式(State Equations)、代數方程式(Algebraic Equations)或其他數學模型，透過Petri-net了解模型的運作行為，故Petri-net是可以被應用在實際與理論之間。

Petri-net的應用領域相當廣泛，只要具有流程觀念的資料，都可以藉由Petri-net來呈現。如效能評估 (Performance Evaluation)、溝通協定(Communication Protocols)、彈性製造的控制系統(Flexible Manufacturing/ Industrial Control Systems)、同步與平行程式語言 (Concurrent and Parallel

Programs)、錯誤容忍系統 (Fault-Tolerant Systems) 、程式邏輯與積體電路 (Programmable Logic and VLSI Arrays) 、分散事件系統 (Discrete Event Systems)、決策模型 (Decision Models)等 [15]。

應用Petri-net來分析工作流程有下列優點：

1. 圖形化表達：透過Petri-net模型可以模擬系統在運作時的動態行為。
2. 正規的語意：Petri-net除了提供圖形化的表達之外，也提供正規的語意以數學的方式表示。
3. 分析技術：Petri-net提供許多分析的行為屬性(Behavioral Property)以驗證Petri-net模型。

在Petri-net中，有以下兩種典型的方法進行分析上述的性質[16]：

1. 可達樹(Reach Ability Tree)

   可達樹首先要建立如同樹狀的結構，其中每一個結點(Node)表示此時的狀態，上一層的節點經過轉換的觸發就可以到達下一個節點。利用可達樹可以用來分析Petri-net是否具有安全性、有限性。但是當系統過大時，其圖形將變得很複雜而無法分析，故其只適用於小系統。圖2-4為一可達樹的例子。在可達樹中，括號內的值分別代表P1，P2，P3，P4 之狀態， P1及P2各有一個初使狀態，故在可達樹中標記成(1,1,0,0)。初始狀態(1,1,0,0)中，T1及T2皆可被觸發，若T1被觸發，其標記變成(0,1,1,0)，此時觸發T2，狀態轉變成(0,0,1,1)，最後當T3 觸發之後，狀態回復到初始狀態。透過可達樹的分析，可以確認在整個過程中有沒有發生死鎖現象。

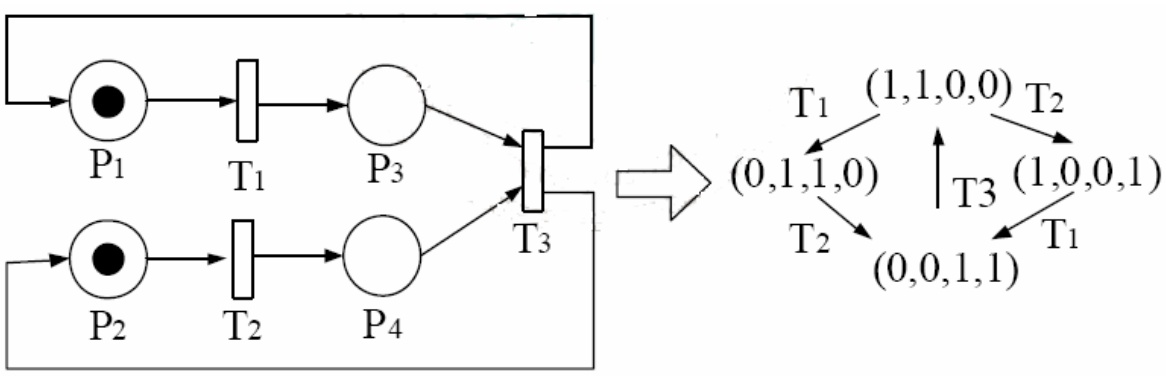

圖2-4 Petri-net使用可達樹表示方式
(資料來源:參考Ying-Yen Hsu)

2. 矩陣方程式

透過數學矩陣的計算，推理出Petri-net的行為。首先定義出一個系統矩陣，並利用此矩陣來表示位置與轉換間的流向關係，經過數學公式推算轉換之後的狀態。

2.4 轉換控制矩陣

要描述空間中一個物體的位置與方向，必須要先有一個參考座標，在機構分析中我們通常定義一個固定座標作為參考座標，然後我們會在物體再附上一個座標，稱之為隨動座標（ local coordinate），之後便可以利用一個4×4 的矩陣也就是所謂的轉換矩陣，來表示固定座標的位置與方向[17]。

1. 位置的描述

一旦座標系統建立好了之後，我們可以用一個3×1 的矩陣來表示空間中的位置向量（position vector）。例如我們用向量來表示空間中P點相對於座標系A的位置。

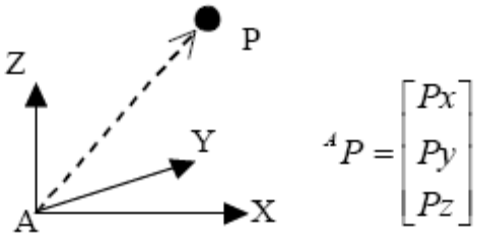

圖2-5 位置向量之描述
(資料來源:參考宋長耀碩士論文)

2. 方向的描述

但是對於一個物體來說，光是描述它在空間中的位置是不夠的，還必須說明該物體在空間中的方向（orientation）。要描述一個物體在空間中的方向時，我們可以用一個3×3 的矩陣來表示該物體之隨動座標的三個軸向量在固定座標上的分量。也就是說，我們以隨動座標來代表物體在空間中的方向，並且可以由沿著不同軸向的旋轉來得到。假設座標B 為物體之隨動座標，而座標A 為固定座標，然後我們用矩陣來描述座標B 相對於座標A 的關係。

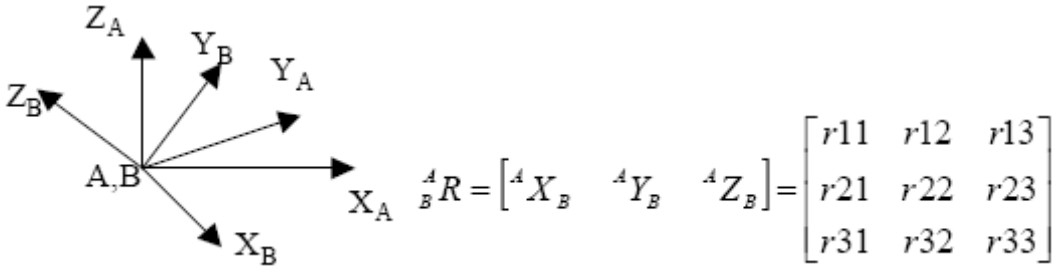

圖2-6 座標方向之描述
(資料來源:參考宋長耀碩士論文)

2.5 相關研究

預測網路使用者未來瀏覽行為的研究大致上有以馬可夫鏈[17]、Self Simi -larity[18]、類神經網路、或者Web Mining 這幾種較為常見。以馬可夫鍊做預測主要是依循著網站架構及歷史的使用者瀏覽行為，去統計出每

一個連結發生的機率，當新使用者的瀏覽行為發生時，會根據該使用者一連串的瀏覽行為找出發生機率最大的連結作為預測的結果。以類神經網路建立一個預測模型的話，是利用類神經網路可以自我學習的特性，調整系統，以適應多變的網路使用者瀏覽行為。

至於以Web Mining 分析使用者的行為更是不勝枚舉[19][20]，但是，大部分的論文都必須有一個額外輔助的資訊，如：網頁的架構、頁面與頁面的相似度，來輔助預測系統。利用這些事先調查好的輔助資訊的確有利於預測系統的運作，然而，網路使用者瀏覽行為的多變性或不確定性卻降低了真正使用上的可行性。舉例來說，我們知道網站的特點就是可以隨時更新資訊，並隨時根據使用者的需求調整網站，因此，一個網站架構及包含的頁面並不是恆久不變的，所以，當網站更新或擴充時，該類輔助資訊往往需要增加許多額外的工作量。很不幸的，這樣的情況在網路使用者瀏覽行為以及實際的網站建置維護中屢見不鮮，造成此類方法的致命傷。

大部分的研究都只著重在如何利用使用者瀏覽行為或者使用者瀏覽網頁興趣的相似度有效的做分群，並且利用分群結果建立使用者的行為模型來進行預測，而本篇論文主要是應用Petri-net預測方式，提供一個完整的使用者行為預測系統架構，對使用者的行為進行預測，並以此特性適應於多變的使用者瀏覽行為。此外，大部分利用分群做預測的情形，都只考慮使用者屬於哪一群，但我們認為在實際情況下，由於使用者大部分都具有獨特的特性，就僅以此一群提供的資訊做預測，未免太過牽強，並不符合真實情形。因此在本篇論文中，我們除了希望就僅知網路使用者過去的瀏覽Log Database 紀錄，去分析建立我們的預測模型，找到滿足門檻值的參考模式去預測網路使用者未來的瀏覽路徑。也由於採用這樣的預測概念，很有可能能夠提供做為我們預測的參考模式不只一個，而這樣的特

性可以讓我們更容易適應多變的使用者行為。