行政院國家科學委員會專題研究計畫　成果報告

應用隨機過程時間派翠網路來強化網頁使用模式探勘

計畫類別：個別型計畫
計畫編號：NSC93-2213-E-032-016-
執行期間：93 年 08 月 01 日至 94 年 07 月 31 日
執行單位：淡江大學資訊工程研究所

計畫主持人：陳伯榮

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 94 年 10 月 27 日

行政院國家科學委員會專題研究計畫成果報告

應用隨機過程時間派翠網路來強化網頁使用模式探勘

Using Stochastic Timed Petri Nets to enhance Web Using mining

計畫編號：NSC 93-2213-E-032-016
執行期限：93 年 8 月 1 日至 94 年 7 月 31 日
主持人：陳伯榮　淡江大學資訊工程學系

一、中文摘要

我們將運用大家所熟知的隨機過程時間派翠網路來分析使用者的使用習性；在這篇報告中，我們介紹：
（一）利用隨機過程時間派翠網路來建構網頁結構。
（二）利用隨機過程時間派翠網路永不變結構特性和可到達行為特性來協助資料前置處理。

詞：網頁使用者習性探勘，隨機過程時間派翠網路。

Abstract

The well-known Stochastic Timed Petri Nets (STPN) is introduced to analyze the web usage qualitatively and quantitatively. In this report, we introduce how to:
[1] construct web structure by using STPN,
[2] utilize reachability properties to enhance the web usage mining process in preprocessing phase,

Keywords: Web Usage Mining, Stochastic Timed Petri Nets.

二、緣由與目的

網頁使用者習性探勘是近年來十分熱門的研究領域，主要的作法是從網站服務伺服器的日誌檔中擷取使用者在網站中存取網頁的相關資料，分析得到使用者使用網站的習性模式。第一個有關網頁使用模式探勘的研討會 WebKDD 於 1999 年舉行；Federico[1]於 2003 發表了 2000 年以來超過 150 篇有關網頁使用模式探勘領域的調查研究(Survey paper)。

由於越來越多的網站使用頁框（Frame）技術來進行設計，使得網站結構及使用者使用網站資料的前置處理更形困難，Cooley [1]於 May 2003 中指出：網頁使用模式探勘發展至今，由於網頁內容與架構的相關分析經常被忽略，所以只有少數的特殊案例能夠成功的達成目標。Buchner [2] 、 Cooley [3] 、 Pirolli [4] 、Spiliopoulou [5] 及 Heer [6] 都說明了網頁內容及架構對進行網頁使用者習性探勘能夠有所助益。

另外，網際網路使用過程中由於使用者端的瀏覽器或網際網路連結過程中可能面臨快取伺服器（Proxy server），都可能造成使用者使用網頁日誌不完整，也使得路徑填補益形困難。

派翠網路(Petri Nets, PNs)是一個經常被應用於描述及分析分散式系統的工具。Peterson 在[7]中將 PNs 定義為 C=（P,T,I,O），其中的 P 代表位置所成的集合，T 代表轉移所成的集合，I 代表所有由轉移指向位置的箭頭的集合,O 代表所有由位置指向轉移的箭頭所成的集合。另外又同時定義了可以記錄權重的派翠網路（Marked Petri Net）M=（P,T,I,O,μ），μ 代表了每個位置所記錄的權重（Token），Molloy 在[8]中定義了隨機過程派翠網路（Stochastic PNs）SPN=（P,T,A,M,λ），其中的 P 代表位置所成的集合，T 代表轉移所成的集合，A 代表所有由位置指向轉移或由轉移指向位置的箭頭所成的集合、M 代表每個位置所記錄的權重，λ 代表每一個耗時轉移的指數分配時間。Ajmone Marsan 在 [9]中定義的一般隨

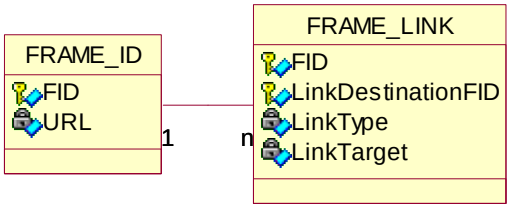
機過程派翠網路(Generalized SPNs)將SPN上加入立即轉移,而立即轉移觸發過程不需耗費時間。

三、結果與討論

我們提出了應用 GSPNs 來建構網頁結構，以 GSPN 中的 P 代表網站中的頁框（frame），T 代表網站中的立即轉移(Immediate Transition)（例如：頁框設定標籤(Frameset Tag)），以 Ts 代表網站中的耗時轉移（例如：超連結標籤(Hyper-Link Tag)），以 F 代表網站使用者所在網頁位置的轉移，以 W 代表位置轉移時分配的權重，以 M 代表系統的狀態，以 E0 代表相對應於耗時轉移動作的指數分配時間。

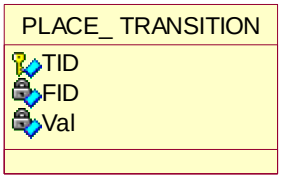
我們先建立包含頁框編號與頁框網址的頁框編號表（FRAME_ID Table）及包含頁框及其超連結資訊的頁框連結表（FRAME_LINK Table）。

其中 FRAME_ID Table 包含兩個欄位，分別是位置編號及對應的網頁網址（URL）。FRAME_LINK Table 包含四個欄位；分別是頁框編號（FID）、頁框所包含的連結的目的頁框編號（LinkDestinationFID）、超連結型別（LinkType）及代表執行超連結時會出現在哪個頁框位置的超連結視窗目標名稱（LinkTarget）。FRAME_ID Table 與 FRAME_LINK Table 的關聯如下圖一所示：



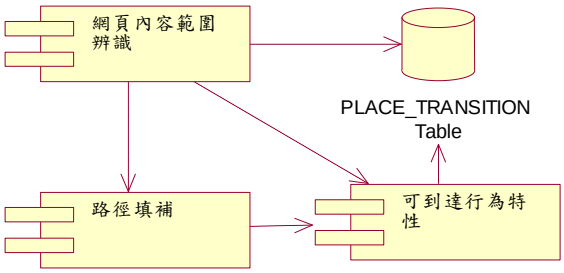
圖一:FRAME_ID Table 與 FRAME_LINK Table 的關聯

我們並進一步分析並建構代表GSPN關聯矩陣的位置與轉移關係表（PLACE_TRANSITION Table），PLACE_TRANSITION Table 中包含三個欄位，分別是頁框編號（FID）、轉移編號表（TID）及代表轉移動作觸發時位置內權重變化（VAL），如下圖二。這個 PLACE_TRANSITION Table 是用來代表 GSPN 中轉移發生時位置內權重變動情形的關聯矩陣 C[i, j]，關聯矩陣中元素 C[i, j]的值代表第 j 個轉移動作觸發時，第 i 個位置得到或失去的標記個數。



圖二：PLACE_TRANSITION Table

由於關聯矩陣中記錄了所有轉移發生時位置中權重變動的情形，我們可以藉由這個特性找出網頁內容範圍也可以利用這個關聯矩陣進行可到達行為特性的運算。有關 PLACE_TRANSITION Table、網頁內容範圍辨識、可到達行為特性及路徑填補的關係如下圖三所示。



圖三：網頁內容範圍辨識及路徑填補系統元件圖

在進行使用者習性探勘過程中，當完成區段辨識後我們將建立使用者區段紀錄(UserSession)，這個區段資料中包含了使用者 IP(UserIp)、需求檔案名稱(FileRequest)、推薦者(Refer)、存取時間(AccessTime)、區段編號(SessionId)等五筆欄位的資料。每個區段紀錄代表單一使用者在一次上網的過程中所瀏覽的網頁內容或瀏覽路徑。

然而，由於越來越多的網頁利用頁框(Frame)技術來進行設計，使得一個網頁內容範圍中往往包含一個以上的頁框，也因此，每一筆或多筆紀錄究竟代表使用者是進入哪一個網頁內容範圍往往無法直接由紀錄本身得知，而必須利用網頁架構來協助判斷。

由於 PLACE_TRANSITIONT Table 記錄了每個轉移發生時相關位置內權重的改

變，因此也代表著與當超連結被使用者點選而觸發一次轉移時，相關位置內權重的改變，也可以說是代表著網站使用者。藉由這個特性可以協助我們進行網頁內容範圍辨識。

有關網頁內容範圍辨識處理程序說明如下：

我們將針對每一筆未處理的 UserSession (LL):

1.以 Refer 為 key,查詢 FRAME-ID Table 中的 URL 欄位，找到對應的 FID（LL-RFID）。

2.以 LL-RFID，”-1” 為 key，查詢 PLACE_TRANSITION Table 中的 FID 及 VAL 欄位，讀取 TID 集合（RTID_Set）。 這個觸發轉移動作構成的集合代表由此 FID 所組成的可能網頁內容範圍中的連結。

3.針對 RTID_Set 中的每一個 TID，以 TID，”-1”為 key，查詢 PLACE_TRANSITION Table 中的 TID 及 VAL 欄位,找到包含 Refer 的可能網頁內容範圍集合（Refer_Pv_Set），並藉此集合決定起始網頁內容範圍集合(Source_Pv_Set）中「網頁內容範圍的優先處理順序」。

4.以 FileRequest 為 key，查詢 FRAME-ID table 中的 URL 欄位，找到對應的 FID（LL-FID）。

5.以 LL-FID，”-1” 為 key，查詢 PLACE_TRANSITION Table 中的 FID 及 VAL 欄位,讀取 TID 集合(TID_Set)。這個觸發轉移動作構成的集合代表由此 FID 所組成的所有可能網頁內容範圍集合中的連結。

6.針對 TID_Set 中的每一個 TID，以 TID，”-1”為 key，查詢 PLACE_TRANSITION Table 中的 TID 及 VAL 欄位,找到可能的目標網頁內容範圍集合(Destination_Pv_set)，他也是下一次內容範圍辨識時的 Source_Pv_Set 。

7.利用 GSPN 的可到達行為特性，並依上述步驟 3 中「網頁內容範圍的優先處理順序」來協助決定 Destination_Pv_set 中的每一個 Pv 是否為可到達的網頁內容範圍，並依需要進行路徑填補。有關可到達行為特性與路徑填補的相關細節我們會進一步說明。

Murata 於 [10]中提到了可到達性問題的必要條件及充分條件，其中必要條件是：假使目標狀態(Md)可以經由初始狀態(M0)來到達，則由觸發順序所構成的集合可利用線性算式來求解，而且存在至少一組解。亦即在派翠網路中，其狀態的轉移可由狀態方程式 (state equation) 來表示，其運算式如下：

此線性算式中 n 表示位置(place)的總個數,m 表示轉移動作(transition)的總個數， 這個一維向量代表在狀態 時位置 P0 到位置 Pn-1 中標記(token)的個數。

表示初始狀態， 表示目標狀態，關聯矩陣 ，表示第 j 個轉移動作觸發時，第 i 個位置得到或失去的標記個數。

這一 的一維向量就是線性算式的解，也就是可到達性問題所要求的答案。這組方程式解表示初始狀態 經由一連串轉移動作的觸發後，產生了目標狀態 。如此即可大量地除去可到達解以外不必要的狀態擴展與搜尋。

在利用 GSPN 的可到達行為性來協助完成路徑填補部份，我們利用可到達行為特性來協助進行路徑填補的工作，填補的程序說明如下：

1. 利用 GSPN 的可到達行為特性來找到相對於起始網頁內容範圍的狀態[Ms]到達相對於目標網頁內容範圍的狀態[Md]的動作轉移觸發順序。

[Ms]→tj1→[Ms1] →tj2→[Ms2] →……→tjk→[Msk]=[Md]

2. 將每一個[Msi]， i=1..k ，轉換成[M]nx1 對應的網頁內容範圍所包含的頁框編號。

3. 依需要利用找到的 URL 來協助路

徑填補。

可到達行為特性的處理程序如下

1. 將起始網頁內容範圍轉換成 GSPN 中的狀態[Ms]。

2. 將目標網頁內容範圍轉換成 GSPN 中的狀態[Md]。

3. 求得下列公式的線性代數解[Tj]mx1:

[Ms]nx1+[Cij]nxm[Tj]mx1 = [Md]nx1

（此線性算式中 n 表示位置(place)的總個數,m 表示轉移動作(transition)的總個數， 這個一維向量代表在狀態 時位置 P0 到位置 Pn-1 中標記(token)的個數。）

4. 若無此解，則從起始網頁內容範圍無法到達目標網頁內容範圍。

5. 若有解，則可由[Tj]中找出由狀態[Ms]到達狀態[Md]的轉移動作觸發順序。

[Ms] →tj1→ [Ms1] →tj2→ [Ms2] ……→tjk→ [Msk]=[Md]

四、計劃成果自評

我們應用隨機過程時間派翠網路來建構網站的結構模型，並進一步應用隨機過程派翠網路的行為特性來協助資料前置處理。

我們成功的利用隨機過程時間派翠網路中的 PLACE_TRANSITIONT Table 來協助進行區段辨識中的網頁內容範圍辨識，使得區段辨識的過程可以解決網站因為使用頁框設計所產生的內容範圍辨識問題。另外我們也利用隨機過程派翠網路的可到達行為特別解決了快取問題所造成的路徑填補問題。這些成果都有助於成功進行網頁模式探勘中的模式發掘。

在未來的研究方面我們將持續探討如何透過漸進的方式來調整網頁結構以便進行使用者習性探勘並加強網頁結構特性的分析。

五、參考文

[1] Robert Cooley “The Use of Web Structure and Content to Identify Subjectively Interesting Web Usage Patterns”, ACM Transactions on Internet Technoloey, Vol.3, No.2, ppP.93-116, May 2003.

[2] A. Buchner, M. Mulvenna, “Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining”, SIGMOD Record, Vol.27, No.4, pp.54-61, Dec.1998.

[3] Robert Cooley, Pang-Ning Tan, Jaideep Srivastava, ”Discovery of Interesting Usage Patterns from Web Data”, Lecture Notes in Computer Science, 2000.

[4] Peter Pirolli, James Pitkow, Ramana Rao , “Silk from a Sow’s Ear:Extracting Usable Structures from the Web”, Conference on Human Factors in Computing Systems, CHI-96, 1996.

[5] Myra Spiliopoulou, Carsten Pohle, Lukas C. Faulstich, “Improving the effectiveness of a web site with web usage mining”, WEBKDD, 1999.

[6] Jeffrey Heer, Ed H. Chi, “Identification of Web User Traffic Composition using Multi-Modal Clustering and Information “, In Proceedings of the 1st SIAM International Conference on Data Mining Workshop on Web Mining, pp.51-58, 2001.

[7] James L. Peterson, Petri Net Theory and the Modeling of Systems, Englewood Cliffs, NJ:Prentice-Hall Inc., 1981.

[8] M. K. Molloy, “Performance analysis using stochastic Petri nets,” IEEE Trans. Computers, vol. C-31, no. 9, pp. 913-917, Sep. 1982.

[9] M. Ajmone Marsan, “A Class of Generalized Stochastic Petri Nets for the Performance Evaluation of Multiprocessor Systems,” ACM Tran. on Computer Systems, Vol. 2, No. 2, pp. 93-122, May 1984.

[10] Tadao Murata, “Petri Nets: Properties, Analysis and Applications,”Proceedingsof the IEEE, Vol. 77, No. 4, 1989