在階層式網站下具回饋機制之使用者瀏覽行為預測模型

李朱慧　傅昱翔
朝陽科技大學　資訊管理系
chlee@cyut.edu.tw　s9514611@cyut.edu.tw

摘要

使用者瀏覽行為預測在電子商務上的應用是一項重要的技術，預測的結果可應用在建立個人化、適合的網站、促銷活動、取得相關的市場情報、預測市場趨勢、…等，然而網頁上資料是隨時快速地更新或是新增，如何有效的進行預測並且要能適時的偵測到行為模式的改變，是一項值得探討的課題。在本研究中針對階層式網站架構之本質，設計了一個兩階層式預測模型，同時還加入隱含式的回饋機制，以有效的修正預測的模型。本研究所提出的兩階層式預測模型，除了能夠減少在預測時計算的範圍，同時能夠有不錯的精確度，而且在預測模型之中加入了隱含式回饋機制，能有效率的反應使用者瀏覽行為的改變，以達到提高精確度的目的。

關鍵詞：網站使用探勘、預測、相關回饋、馬可夫模型、貝氏定理。

Abstract

Predicting user's browsing behavior is an important technique of E-commerce application. The prediction result can be used for personalization, building adaptive web sites, promotion, getting marketing intelligence, forecasting market trends etc. However, the information of web pages updates and produces so quickly in time The issues of predicting effectively and detecting the behavior pattern change opportunely are worth to investigate. In this paper, we focus on the property of hierarchy web site and propose a prediction model with relevant feedback mechanism. The prediction model has the advantage of reducing the computing complexity without losing accuracy, and responds the user's behavior which is changed timely by implicit relevant feedback to increase predicting accuracy.

Keywords: Web Usage Mining, prediction, Markov models, Bayesian theorem.

1. 緒論

全球資訊網隨著網際網路的快速成長也日益蓬勃，許多企業也因此更積極的提供有效率的線上交易方式，直接地促進電子商務(electronic commerce) 快速的發展，同時也使得網站使用探勘(web usage mining)技術的發展日漸重要，網站使用探勘是網站探勘(web mining)技術的其中一個種類，主要是將資料探勘(data mining)的相關技術應用到網際網路或網路服務(web service)上，並且能夠自動地發現並取出有用的知識(knowledge)或模式(pattern)，例如：關聯規則(association rule)、分群演算法(clustering algorithm)、序列模式分析(sequential pattern analysis)、…等[7]。

網站使用探勘最主要的目的在於，如何從網站資料(web data)或記錄檔(logs)中找出有用的資訊，並且增加網站資訊的可用性(usability)，將所找出來的資訊進一步應用在網站的預先讀取(pre-fetching)或快取(caching)、個人化(personalization)、目標行銷(target advertisement)、改善電子商務網站的設計、改善顧客滿意度及引導企業利用相關策略及技術作市場分析、… 等[2][7][8][14]。

網站使用探勘也被應用在目前相當熱門的顧客關係管理(CRM，customer relationship management)的領域上，例如：金融業中的銀行。基於線上交易(on-line transactions)的成本遠比傳統交易模式來的低，則我們可以進一步導入網站使用探勘的技術，並將傳統交易模式下的顧客移轉到線上虛擬銀行(visual bank)之上。相同地，我們也可以將顧客關係管理及網站使用探勘應用到電子商務網站上，則可以達到吸引顧客、保留顧客、交叉銷售、…等目的[1]。

為了能夠達到上述的目的，我們必需了解使用者或顧客在網站上的瀏覽行為(browsing behavior)，從網站資料或記錄檔中利用網站使用探勘的技術，對使用者或顧客的瀏覽行為進行預測(prediction)。我們利用過去有用的資訊進行分析，對未來的使用者或顧客在瀏覽網站時，在針對符合的瀏覽行為時，計算出最有可能下一步為何。實行預測的方式有許多的優點，例如：個人化、建立適合的網站，改善行銷策略、促銷活動、產品的提供、取得相關的市場情報(market intelligence)、預測市場趨(prediction)，藉此提升企業的競爭優勢、…等。

馬可夫模型(Markov model)被認為在預測使用者瀏覽行為中是一個合適的機率(probability)模型[9][15]。馬可夫模型能夠表示出使用者在瀏覽時的瀏覽行為，當目前正在瀏覽的狀態(state)亦或者是網頁發生時，都是依賴前一次或 n 次的狀態，則得到一個轉換機率(transition probability)並且進一步建立一個轉換矩陣(transition matrix)[4][5][9][12][13]。相同地貝氏定理[17]亦能表達出使用者的瀏覽行為，以貝氏定理的方式來表示使用者在瀏覽網頁時，可以使用者的前 n 個網頁為依據，對使用者下一步可能瀏覽的網頁做機率的推論。在本研究中針對階層式網站架構[11]之本質(如圖1)，結合了馬可夫模型與貝氏定理，並設計了一個兩階層式預測模型，透過兩階層的方式逐層縮小預測網頁的範圍。

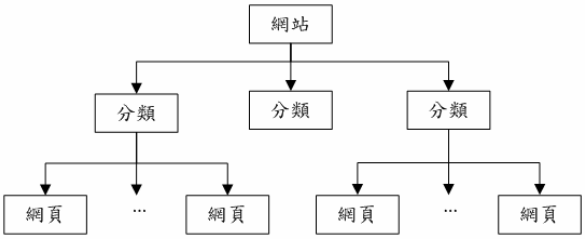

圖 1、階層式網站架構

在現今大多數的入口網站等，會針對不同的目的而將網頁做適當的分類，且該網站的網站架構又大多屬於階層式架構，在階層式架構中在網站架構的最上層是該網站的首頁，接下來按照該網站的網頁資料內容，依相同的網頁內容再分為主要分類，再來該主要分類又會依其內容再細分成其它子分類、…等，最後在網站架構的最底層為屬於該分類或子分類的網頁。我們將使用者的瀏覽行依照階層式網站的本質將預測分成分類及網頁兩個階層，在一個網站中的每個分類都會包含許多的網頁，分類的數量皆小於網頁數量，所以相較之下分類在機率上的轉換會較為穩定。在每個分類底下則是數量眾多的網頁，我們能夠過濾出使用者最可能瀏覽的分類之後，再以貝氏定理做推論，就能夠找出被瀏覽機率最大的網頁。期望透過本研究所提出的兩階層式預測模型，除了能夠減少在預測時的計算範圍，並且能夠有不錯的精確度(precision)。

本論文的架構如下，第二節為相關文獻探討，我們的研究方法在第三節會做詳細的說明，第四節為未來工作及結論。

2. 文獻探討

網際網路就像是一座金礦山(golden mount)埋藏許多有價值的資訊，網站探勘就是一種強大且有用的工具，能夠透過有系統、有效率並自動地取出資訊進而形成有用的知識。網站探勘一辭是在 1996 年由 Etzioni 所提出之後，主要是將資料探勘技術應用到全球資訊網的網頁或服務上，能夠自動地發現並取出有用的知識[7]。

網站探勘可分為三個種類(如圖 2)：1.網站內容探勘(web content mining)、2.網站結構探勘(web structure mining)、3.網站使用探勘(web usage mining)。網站內容探勘主要能夠從網頁內容中取出有用知識。網站結構探勘分析發現在網頁之間連結的結構，藉由這個結構進行知識的推論。網站使用探勘利用到網站的記錄檔中取出與使用者相關的網頁存取資訊。

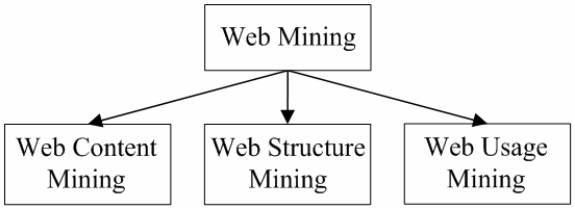

圖 2、網站探勘的分類

2.1 網站記錄檔資料

表 1、網站通用記錄檔內容之描述[18]

Content of log file	Description
163.17.9.84	Address or DNS
-	RFC931
-	Authuser
[08/Nov/2007:20:35: 52 +0800]	TimeStmp
"GET /menu.htm HTTP/1.1"	Http request
200	Status code
1266	Transfer volume
-	*Referrer URL
-	*Usage agent

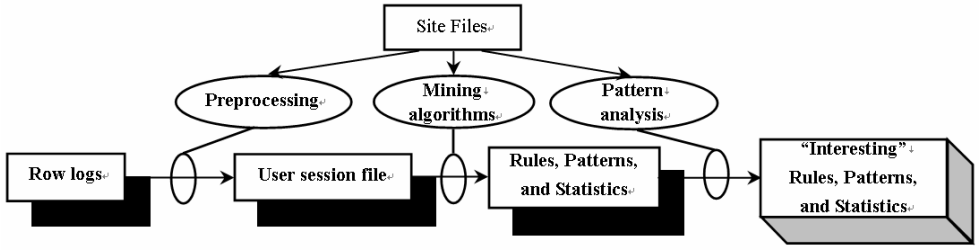

圖 3、網站使用探勘流程

常用的記錄檔格式又可分為主要有兩種：通用記錄檔(common log，如表 1)和擴充記錄檔(extended log)[1][7][16]。通用記錄檔由處理記錄、錯誤記錄所構成，再加上來源記錄、代理人記錄則成為擴充記錄檔。在取得網站記錄檔之後，則需進行網站使用探勘的流程。

2.2 網站使用探勘流程

首先，先對記錄檔做前處理、再來利用探勘演算法找出模式、最後是對探勘出來的模式進行分析。網站使用探勘之處理流程(如圖 3)如下列步驟所述，從網站記錄檔中經過前處理(preprocessing)來建立各個使用者的 *session*，並且辨識出序列模式，再利用探勘演算法(mining algorithm)，例如：關聯規則、分群演算法、序列模式分析、...等方法，從序列模式中找出規則、模式。最後透過模式分析的方法從初步結果中，找出有趣的規則、模式、...等等[2][5][16]。

2.3 相關回饋

相關回饋(relevance feedback；RF)可以被視為是一種直接的策略來針對使用者動作(action)做監督式學習(supervise learning)，透過反覆(iterative)及互動(interactive)的方式，重新組織使用者的查詢(query)，使得檢索的結果可以讓使用者更為滿意並且可以達到使用者模型化(user modeling)的目的。相關回饋可以簡單的分成主動式(active)回饋及被動式(passive)回饋兩種[3]。主動式的機制會自動的分析使用者的變化而修正系統的機制，被動式的則是需要使用者利用回饋表格(feedback form)自行送出回饋資訊。在使用者的回饋中又可以分成正回饋(positive)及負回饋(negative)，正回饋為使用者自己的觀點所認定及被標記(mark)為相關(relevance)的資訊，反之則為不相關(irrelevance)即是負回饋。

在網站搜尋引擎[3][10]方面，在沒有相關回饋機制下，搜尋的結果同樣是依賴搜尋引擎的演算法來判斷那些網頁資訊與使用者所使用的關鍵字(key words)是有相關的，這使得搜尋的結果不一定會讓使用者感到滿意。但是網站搜尋引擎在加入了相關回饋之後，除了演算法的部份外，還能夠有使用者回饋的觀點來輔助搜尋引擎能夠搜尋到更讓使用者滿意的搜尋結果，並且調整更新搜尋引擎演算法中相關的參數及搜尋索引(index)，使得網站搜尋引擎更為可靠，同時也能夠達到個人化搜尋[10]及個人化摘要(summarization) [6]等目的。

相關回饋在網站搜尋引擎的應用上，又被稱為明確回饋(explicit feedback)及隱含回饋(implicit feedback)。明確回饋透過同樣以回饋表格的方式清楚且明確地送出回饋資訊，此作法可以簡單的分析那些檔案資訊的是被使用者認定有相關的。使用者可以透過明確回饋按照使用者本身的觀點，對檔案資訊直接做排序(ranking)或者是直接標記為相關資訊。但是如果太過於要求使用者實行明確回饋，使用者會感到太浪費時間，也有可能得到錯誤的使用者回饋。在隱含回饋方面並沒有像明確回饋中的回饋表格，所以要實行為回饋必需要設計一些方法來分析並取得隱藏在使用者動作中的回饋。在使用者的查詢被回應之前，利用設計出來的方法來判斷及推論出是否有隱含使用者回饋，最後利用學習機制來學習使用者回饋用以了解使用者行為。

3. 具回饋機制之兩階層式預測模型

在本研究中結合了馬可夫模型與貝氏定理，設計了一個適合於階層式網站的兩階層式使用者預測模型(如圖 4)。首先，透過此模型在階層一時，先針對使用者目前瀏覽的狀態(網頁)之分類進行預測。依賴在目前時間 $t-1$ 的使用者瀏覽狀態之分類，及在時間 $t-2$ 的使用者瀏覽狀態之分類，對使用者在時間 t 可能瀏覽的狀態之分類進行預測。

隨後在階層二時，依據階層一時所預測正確的分類，再利用貝氏定理根據使用者在時間 $t-1$ 的狀態，對在時間 t 可能瀏覽的網頁預測。最後將透過兩階層式預測模型的預測結果

輸出，其結果除了可以應用在網站的預先讀取或快取，甚至於可以進一步運用在個人化、目標行銷、改善網站設計、...等。

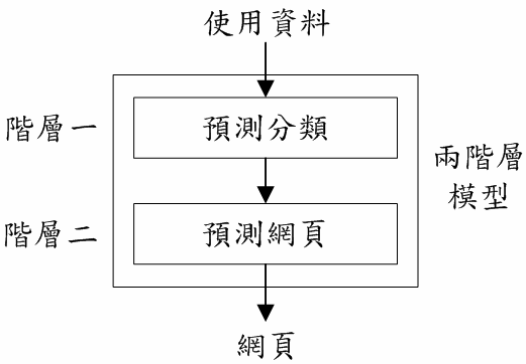

圖 4、兩階層式預測模型

在減少計算範圍方面(如圖 5)，兩階層預測模型的計算範圍在階層一時，會將要預測的範圍縮小到特定 $top-r_1$ 個相關性(relevance)最大的分類，隨後在階層二時再利用貝氏定理對階層一所鎖定預測正確的分類內進行網頁的預測即可，則再一次將其預測的範圍縮小到前 $top-r_2$ 個機率較大的網頁。。

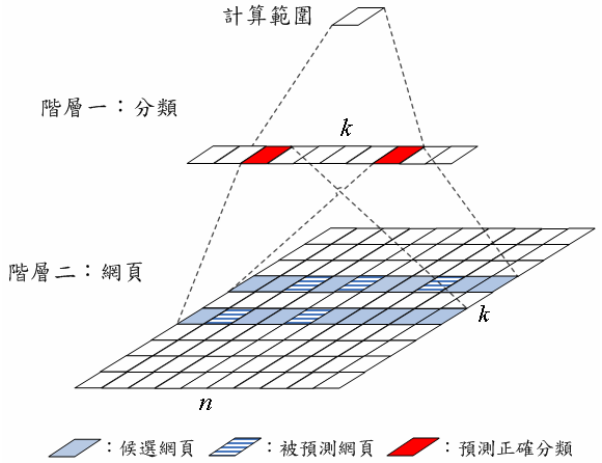

圖 5、計算範圍示意圖

3.1 研究架構

在本研究中的研究架構(如圖 6)之步驟為，在第一步驟為建立各個網頁分類之間的相似度矩陣 S 的作法為從使用者的瀏覽記錄檔中統計並分析使用者的瀏覽行為，進而建立一個以使用者瀏覽行為為基礎的網頁分類相似度矩陣。

在第二步驟為，建立馬可夫模型之轉換矩陣 P 與相似度矩陣相同的地方是，我們同樣以使用者為瀏覽行為作為建立的轉換矩陣的基礎，所建立的矩陣為馬可夫模型之一階轉換矩陣。

在第三步驟為，從相似度矩陣及馬可夫模型之一階轉換矩陣中，相對應位置相乘來建立相關性矩陣。由於在本研究中的預測部份，是將兩個網頁之間的相關性，視為影響瀏覽行為的重要因素。

在第四步驟為，由於在本研究中的預測部份，是將兩個網頁之間的相關性，視為影響瀏覽行為的重要因素。隨後，透過本研究所提出的兩階層式預測模型，對使用者行為進行預測。透過階層一來過濾出使用者最有可能瀏覽的分類，在階層二時利用貝氏定理針對階層一所得到的分類集合，並且預測正確的分類進行網頁的預測。

在第五步驟為，當在階層一預測分類的錯誤率達到所設定的門檻值時，透過相關回饋的方式，對馬可夫模型之轉換矩陣作更新調整及重新計算一階至 n 階轉換矩陣。藉由相關回饋的方式來降低階層一預測分類的錯誤率及提高精確率。

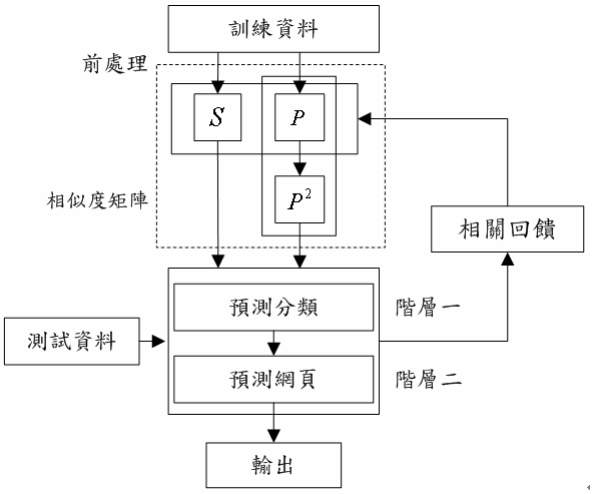

圖 6、研究架構

假設一個使用者瀏覽記錄資料庫 D，該資料庫有 m 筆使用者記錄，即使用者 $session$，則得到 $D=\{session_1, session_2, ..., session_m\}$。每一個使用者 $session$ 可被視為是一個序列模式，並由 n 個網頁依照時間順序被瀏覽，則得到 $session^p=\{page_1, page_2, ..., page_n\}$，若該網站有 k 個分類(category)，則我們可將其轉成 $session^c=\{c_1, c_2, ..., c_k\}$ 的資料，$session^c$ 定義為所瀏覽過的網頁分類，$c_i=1$ 代表瀏覽過分類 i，$c_i=0$ 則代表此 $session$ 未瀏覽到分類 i，依照各網頁所屬的分類來替代在 $session$ 中原有的 n 個網頁，即得到與時間順序無關的

$session^c = \{c_1, c_2, ..., c_k\}$。在做好資料的相關定義之後，則可繼續建立本研究中的預測模型。

3.1.1 網頁分類之相似度矩陣

在研究架構的第一步驟為依據使用者瀏覽記錄檔，建立各個網頁分類之間的相似度矩陣。首先需了解在每一個使用者 *session* 中各個分類被瀏覽的情形(如表 2)，其中每一列為 $session^c = \{c_1, c_2, ..., c_k\}$，代表一個使用者的瀏覽各分類的狀態，表 2 中的每一行則為分類 i 被各使用者流覽的狀態，$vector_i = < v_{1,i}, ..., v_{h,i}, ..., v_{m,i} >$ 代表第 i 行的資料，其中 $v_{h,i} = 1$ 則表示分類 i 被使用者 h 瀏覽過，若沒有被瀏覽過則 $v_{h,i} = 0$。

表 2、使用者 Sessions

	C_1	C_2	…	C_k
$session_1$	1	0	…	1
$session_1$	0	1	…	0
⋮	⋮	⋮	⋱	⋮
$session_m$	1	1	…	1

任兩個分類都可以利用下列兩個公式來計算其相似度，分別為集合相似度(set similarity)公式(1)及歐幾里德距離 (Euclidean distance)公式(2)來計算任兩個分類之間的相似度，在歐幾里德距離公式的部分，我們進一步將歐幾里德公式做標準化則得到公式(3)，並將公式(1)及公式(3) 的計算結果分別給予合適的權重再加總則可得到公式(4)。

集合相似度(set similarity)：

$$SetSim(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

歐幾里德距離(Euclidean distance)：

$$D(A,B) = \sqrt{\sum_{i=1}^{m} (A_i - B_i)^2} \quad (2)$$

標準化(Normalization)：

$$N(D(A,B)) = 1 - \sqrt{\frac{\sum_{i=1}^{m} (A_i - B_i)^2}{m}} \quad (3)$$

總合權重相似度：

$$S(A,B) = SetSim(A,B) \cdot W_{SS} + N(D(A,B)) \cdot W_D \quad (4)$$

且 $W_{SS} + W_D = 1$、$W_D = 1 - W_{SS}$。

在計算出任兩個分類的相似度之後，則可以依序上述的步驟來建立一個 k x k 的網頁分類之相似度矩陣 S，其 S 矩陣表示如下：

$$S = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1k} \\ S_{21} & S_{22} & \cdots & S_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ S_{k1} & S_{k2} & \cdots & S_{kk} \end{bmatrix} \end{array}$$

在 S 矩陣中每一個元素都表示為某兩個分類之間的相似度，此為一個對稱矩陣，例如，S_{ij} 為分類 C_i 與分類 C_j 之間透過公式(4)計算所得到的相似度，其相似度計算的基礎皆基於使用者瀏覽記錄資料中，所記錄的使用者瀏覽行為。

3.1.2 馬可夫模型之轉換矩陣

在研究架構的第二步驟為依據使用者瀏覽記錄資料，建立馬可夫模型之轉換矩陣 P 和相似度矩陣 S 相同的是，轉換矩陣的內容也是根據用者瀏覽記錄檔統計之後的結果，其矩陣為一階轉換矩陣。在本研究預測的部份，由依賴時間 $t-2$、$t-1$ 的狀態對時間 t 的狀態作預測，除了轉換矩陣 P 還需建立 P^2 等轉換矩陣，一階轉換矩陣 P 的表示方式如下：

$$P = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1k} \\ P_{21} & P_{22} & \cdots & P_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ P_{k1} & P_{k2} & \cdots & P_{kk} \end{bmatrix} \end{array}$$

且 $$P_{ij} = \frac{Number(i,j)}{\sum_{j=1}^{k} TotalNumber(i,j)} \quad (5)$$

在 P 矩陣中每一個元素都表示任兩個分類之間狀態的轉換機率，例如，P_{ij} 即表示由分類 C_i 轉換到分類 C_j 的機率，機率計算的方式利用公式(5)，公式(5)為計算由分類 i 轉換到分類 j 的機率，其中分子為所有從分類 i 轉換

到分類 j 的次數，分母為所以由分類 i 轉換到其它分類的總次數。在建立 P 矩陣再根據公式依續建立 P^2 的轉換矩陣。

3.1.3 相關性矩陣

在研究架構的第三步驟為從相似度矩陣 S 及馬可模型之一階轉換矩陣 P 中，相對應位置相乘得到一個相關性矩陣 R^n。由於在本研究中我們將兩個網頁之間的相關性，視為影響瀏覽行為的重要因素。則相關性矩陣的表示方式如下：

$$R^n = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} R^n_{11} & R^n_{12} & \cdots & R^n_{1k} \\ R^n_{21} & R^n_{22} & \cdots & R^n_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ R^n_{k1} & R^n_{k2} & \cdots & R^n_{kk} \end{bmatrix} \end{array}$$

且 $R^n_{ij} = S_{ij} \cdot P^n_{ij}$ (6)。

在 R^n 矩陣中每一個元素都表示某兩個分類在兩個時間點之間的相關性程度，例如，R^n_{ij} 即表示由分類 C_i 在時間 $t-n$ 與分類 C_j 在時間 t 之間相關性的程度，在我們的模式中我們會使用到 $n=1$ 與 $n=2$ 的相關性；相關性的計算方式是利用公式(6)，公式(6)將矩陣 S 及矩陣 P 中相對應位置相乘，即 $S_{ij} \cdot P^{n=1}_{ij}$ 得到 R^1_{ij}，相同的由 $S_{ij} \cdot P^{n=2}_{ij}$ 可得到 R^2_{ij}，R^n_{ij} 表示相似度高的分類之間，應當會存在著較高的轉換機率，即在相似度與轉換機率相乘之後得到高的相關性。計算相關性之後，其結果可作為在本研究之預測部份中一個相當重要的參考依據。

3.2 兩階層式預測模型

在研究架構的第四步驟為兩階層式預測模型，利用透過馬可夫模型之一階($n=1$)及二階轉換矩陣($n=2$)所得到相關性矩陣，來對使用者瀏覽行為作預測。在本研究的預測機制中(如圖 7)，對未知的 C 作預測會依賴目前所在的位置為 B 及過去的位置 A。在本研究中提出兩階層式預測模型，先後分別進行階層一的預測分類及階層二的預測網頁。其主要目的為大幅減少預測時的候選項目(網頁)，可以減少預測時的計算範圍。

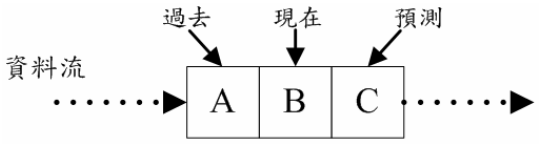

圖 7、預測示意圖

3.2.1 階層一：預測分類

在階層一中最主要目的為預測分類(如圖 8)，依賴前一個時間點的分類狀態 C_{t-1} 或前二個時間點的分類狀態 C_{t-2}，來找出最有可能的分類 C_t 則可得到一個分類集合 θ。在得到分類集合 θ 之後再進入到階層二進行網頁的預測，在階層二中只有屬於分類集合 θ 的網頁才被考慮作為預測的候選項目。如此一來在透過階層一的過濾之後，則可以大幅度減少在階層二預測時的計算範圍。

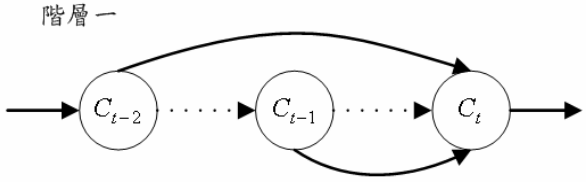

圖 8、階層一預測分類

在階層一中，我們利用相關性矩陣來過濾出可能的分類，$R^n_{C_{t-n}}$ 代表從 R^n 矩陣中取出的第 C_{t-n} 列，表示為 $[C_{t-n,1}, C_{t-n,2}, ..., C_{t-n,k}]$，為一個屬於分類 C_{t-n} 的預測向量，該向量的元素值愈大者代表下一步走到所對應到的分類的可能越大。若我們選擇 $n=1$ 及 $n=2$ 時，則可以得到 $R^1_{C_{t-1}}$ 及 $R^2_{C_{t-2}}$ 兩個向量，並且這兩個向量中取前 $top\text{-}r_1$ 的值所對應到的分類形成一個分類集合 θ。此分類集合 θ 被定義如下：

$\theta = \{C_t \mid$ 為列向量 $R^1_{C_{t-1}}$ 及 $R^2_{C_{t-2}}$ 中前 $top\text{-}r_1$ 的元素值所對應到的分類 $C_t\}$

3.2.2 階層二：預測網頁

在階層二中利用貝氏定理從分類集合 θ 中，計算後繼者 $page_b$ 並取得前 $top\text{-}r_2$ 個機率最大，即最有可能被使用者所瀏覽的網頁(如圖十一)，則可得到一個網頁集合 τ，並輸出做為兩階層式預測模型所預測的結果。

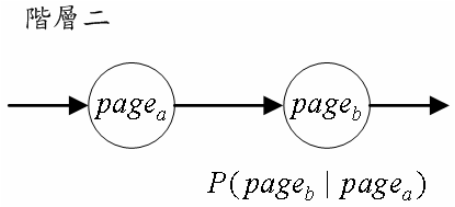

圖 8、階層二預測網頁

在階層二中，我們利用貝氏定理來預測。在圖中 $page_a$ 為目前所瀏覽的網頁，$page_b$ 為下一個使用者可能瀏覽的網頁，並且 $page_b$ 的可能範圍被限制為只屬於階層一的預測結果，即分類集合 θ。此時可以假設 $page_b$ 有可能候選網頁可以被定義成集合 P_θ，則 $P_\theta = \{page_{b_1}, page_{b_2}, ..., page_{b_r}\}$ 且所有可能候選網頁的分類皆屬於分類集合 θ。貝氏定理的公式如下：

$$P(page_{b_i} \mid page_a) = \frac{P(page_a \mid page_{b_i})P(page_{b_i})}{\sum_{j=1}^{r} P(page_a \mid page_{b_j})P(page_{b_j})} \quad (7)$$

網頁集合 τ 被定義如下：

$\tau = \{Page_{b_i} \mid$ 在 P_θ 中利用 $P(page_{b_i} \mid page_a)$ 計算後前 $top\text{-}r_2$ 個之網頁$\}$

3.3 隱含式回饋機制

在研究架構的第五步驟為在模型訓練完之後，在對未來使用作預測時，若在階層一預測分類的錯誤率到達所設定的門檻值時，透過隱含回饋的方式(如圖 9)對相似度矩陣及馬可夫模型之轉換矩陣作更新調整及重新計算一階至 n 階轉換矩陣，藉此由相關回饋的方式來降低階層一預測分類的錯誤率。

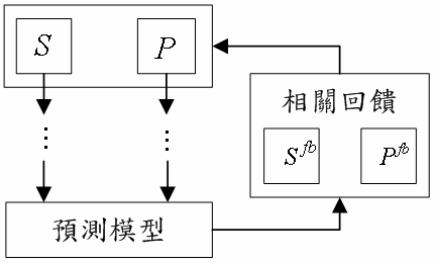

圖 9、隱含式回饋

隱含式回饋的流程(如圖 10)為在進入階層一預測分類時，假若階層一的某一分類 C_t 之預測錯誤率到達我們所自行設定的門檻值時，此時可利用指數平滑法，利用相關回饋對轉換矩陣中此一分類 C_t (即某一列)之轉換機率 $P_{t,1}, P_{t,2}, ..., P_{t,k}$ 與 $S_{t,1}, S_{t,2}, ..., S_{t,k}$ 作修正。在修正轉換矩陣中特定分類 C_t 的轉換機率之後，再重新計算 n 階相似度矩陣。若沒有超過門檻值時，則繼續進入階層二作網頁的預測。

3.3.1 指數平滑法

在作業研究等研究領域，指數平滑法經常被用來作為預測的方法，在此我們利用其為回饋的機制。依照所設定的平滑數 α 乘上目前的實際轉換機率值與欲修正值之差，再加上欲修正值則得到下一期所要修正的轉換機率值。指數平滑法的公式表示如下：

$$S_t^{new} = S_t^{current} + \alpha(S_t^{fb} - S_t^{current}) \quad (8)$$

$$P_t^{new} = P_t^{current} + \alpha(P_t^{fb} - S_t^{current}) \quad (9)$$

S_t^{new}、P_t^{new}：所要修正成的新轉換第 t 列的機率值。

S_t^{fb}、P_t^{fb}：目前實際的第 t 列轉換機率值。

$S_t^{current}$、$P_t^{current}$：修正前第 t 列的轉換機率值。

α：$0 \sim 1$，$\alpha \to 0$：誤差的調節能力小(平滑度高)。$\alpha \to 1$：誤差的調節能力大(敏感度高)。

在預測模式中我們將用新的 S_t^{new} 與 P_t^{new} 值取代對應的值。

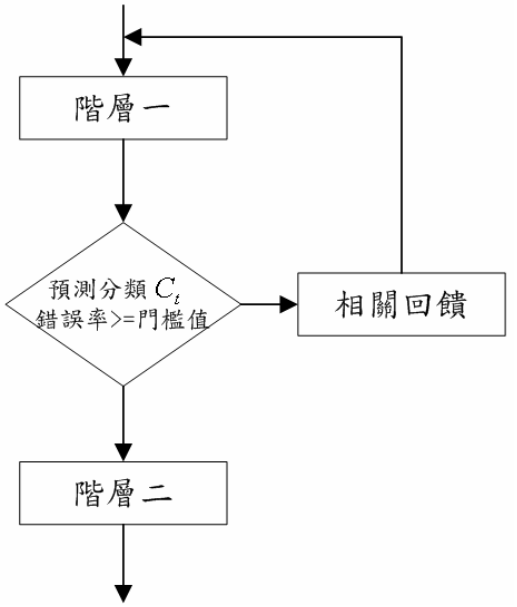

圖 10、隱含式回饋流程

4. 結論及未來工作

在許多的入口網站中，由於網頁資料量的龐大，為收搜尋方便都會將其依照階層式架構加以分類整理。在本研究中，希望能藉由階層式網站架構的特色進行為兩階層的預測，並提出了兩階層式預測模型。在階層一即是預測分類，可先將不必要的分類排除在計算之外，先大幅度縮小計算範圍，再進行階層二的貝氏計算，期望能有效減少預測時的計算範圍且準確地預測使用者行為。假若在階層一的預測分類的錯誤率到達到門檻值時，利用相關回饋來修正新的相似度矩陣。在未來工作的部份，我們將要進行實驗的步驟，透過實驗來證明兩階層式預測模型，能夠達到所預期的目標。除此之外，針對非階層式架構的網站，進一步加入分群的方法來修改兩階層預測模型。

參考文獻

[1] S. Araya, M. Silva and R. Weber, “A Methodology for Web Usage Mining and Its Application to Target Group Identification,” ***Fuzzy Sets and Systems 148***, pp. 139-152, 2004.

[2] W. Bin and L. Zhijing, “Web Mining Research,” I***CCIMA’03 IEEE***, pp. 84-89, 2003.

[3] R.I Chang and Y.Y. Wu, ” Internet Search by Active Feedback,” ***ISCIT'07 IEEE***, pp.1524-1529, 2007.

[4] D. Dhyani, W. K. Ng and , S. S. Bhowmick, ”A Survey of Web Metrics,” ***ACM Computing Surveys***, Vol. 34 2002.

[5] D. Dhyani, S. S. Bhowmick and W. K. Ng, “Modeling and Predicting Web Page Accesses Using Markov Processes,” ***DEXA’03 IEEE***, 2003.

[6] A. Dl´az, and P. Gerva´s, “User-model based personalized summarization,” ***Information Processing and Management 43***, pp.1715–1734, 2007.

[7] F. M. Facca and P. Luca Lanzi, “Mining Interesting Kknowledge from Weblogs: A Survey,” ***Data and Knowledge Engineering 53***, pp. 225-241, 2005.

[8] R. Kosala, H. Blockeel, “Web Mining Research: A Survey,” ***SIGKDD Explorations***, Volume 2, Issue 2, pp. 115.

[9] S. Jespersen, T. B. Pedersen and J. Thorhauge, “Evaluating the Markov Assumption for Web Usage Mining,” ***WIDM’03 IEEE***, pp.82-89, 2003.

[10] S. Jung, J. L. Herlocker, J. Webster, “Click data as implicit relevance feedback in web search,” ***Information Processing and Management 43***, pp.791–807, 2007.

[11] N. Liu and C. C. Yangi, ”Extracting a Website’s Content Structure From its Link Structure,” ***CIKM’05 IEEE***, pp. 347-348, 2005.

[12] G. Pallis, L. Angelis, A. Vakali, “Vaildation and interpretation of Web users’ sessions clusters,” ***Information Processing and Management 43***, pp. 1348-1367, 2007.

[13] L. Shi, Z. M. Gu, L. Wei and Y. Shi ”Popularity-based Selective Markov Model,” ***WI’04***, pp.504-507, 2004.

[14] J. Srivastava, R. Cooley, M. Deshpande and P. N. Tan,”Web Usage Mining:Discovery and Applications of Usage Patterns form Web Data,” ***SIGKDD Explorations***, Vol. 1, Issue 2, pp. 12-23, Jan 2000.

[15] R. R. Sarukkai, ”Link prediction and path analysis using Markov chains,” ***Computer Networks***, pp. 377-386, 2000.

[16] R. M. Suresh and R. Padmajavalli, ”An Overview of Data Preprocessing in Data and Web Usage Mining,” ***Digital Information Management IEEE***, pp. 193-198, 2006.

[17] R. Walpole, R. Myers, S. Myers and K. Ye, “***Probability and Statistics for Engineers and Scientists***,” in Paperback, 7 ed., Pearson Education, 2002, pp.82-87.

[18] W3C Extended Log File , http://www.w3.org/TR/WD-logfile.html.