以網站探勘技術爲基礎之電子目錄上推薦服務之研究

A Study of Recommendation Service on E-Catalog Based on Web Mining Technique

林熙禎
Shi-Jen Lin
中央大學資訊管理學系
Department of Information Management, National Central University
許益誠
Yi-Chen Hsu
中央大學資訊管理學系
Department of Information Management, National Central University

摘 要

隨著電子商務的發展，不論是 B2B 或者是 B2C 的商業模式，電子目錄儼然已經成為買賣雙方的主要介面。電子目錄不再只是提供商品的規格，更是提供客戶服務的重要媒介。然而，一個擁有豐富資訊的網站，如何對於不熟悉網站架構的使用者，或對購買商品特性不熟悉的使用者，協助他們做出採購的決策呢？

本文試圖透過網站探勘技術，瞭解使用者的瀏覽目的，並將此探勘所得到的瀏覽樣式，推薦給具有相同需求的使用者做為參考。同時，將推薦分為同類商品、相關產品、以及其他產品資訊說明三種方式推薦，讓使用者更清楚推薦原因。

本文以網站探勘的技術為基礎，提出概念限制型參考（CCR, Conceptual Constrained Reference）的交易識別方式，來確認交易為某類資訊的瀏覽，藉此確認使用者目的。之後，利用改良的混合型順序性（MIXSEQ）相似度比對，對瀏覽路徑做叢集（Clustering）處理，以做為推薦的資料集。在推薦策略方面，先將網頁做分類，區分為內容型網頁與導覽型網頁，推薦時以內容為主要的推薦網頁，以提高推薦實用性。

關鍵字：電子商務、電子目錄、網站探勘、推薦服務、個人化

Abstract

As the E-Commerce prevalence, E-catalog has become an important interface to a company through which customers interact, regardless of B2B or B2C. E-catalog not only provides product information, but becomes the important media of service providing for customers. However, how does an E-catalog assist casual/unfamiliar users with such rich information in their purchasing decision making?

The challenge for Websites is how we know users needs? Thus, we will apply Web Mining technology to understand user's needs and get usage pattern from previous browsing experience.

In this paper, we propose a new transaction identification, Conceptual-Constrained Reference (CCR), for understanding the goal of users in browsing the e-catalog. Furthermore, we use the MixSEQ similarity measure for web-usage clustering to create the recommendation dataset. Finally, we will classify pages into two types-content pages and navigation pages, and recommend those pages which are content types for users when they are browsing.

Keyword：E-commerce、E-catalog、Web Mining、Recommendation service、Personalization

壹、簡介

隨著電子商務的蓬勃發展，越來越多的零售商、入口網站紛紛提供電子目錄及豐富的產品採購資訊，來做爲企業與顧客之間的銷售與行銷管道。然而，不論 B2B、B2C 還是 C2C，電子目錄均從中扮演了一個相當重要的角色。電子目錄是以全球資訊網爲平台的應用，可以當做整個商業服務的虛擬前台(Front-end)19。透過它，顧客可以進行產品資訊的取得、訂購、付款、客服、回饋等，它改變了傳統的廣告、行銷、配送、支援管道。

從顧客的觀點來看，電子目錄提供顧客選擇產品及服務的另外管道 19。電子目錄除了有大量的產品資訊、讓在家購物更方便的特性外，同時也將其他功能整合在一起，例如與合作伙伴進行整合，彼此分享產品資料庫。電子目錄提供者此時要思索的是，如何透過虛擬的介面協助顧客？尤其當顧客一進入店裡發現琳瑯滿目的商品陳列，如何讓顧客快速找到需要的產品？如何協助顧客釐清他的需求？

個人化的服務最常用的方式便是建立個人偏好檔（User Profile），以便推薦使用者感興趣的商品資訊，因此達到客製化的服務。目前的資訊技術可以透過資訊內容過濾資訊（Content-based Filtering），或者是透過合作式的過濾（CF）推薦同好者喜愛的其他項目。然而，這兩種技術均以靜態個人偏好檔的方式預先設定使用者偏好，這對於需求不明確或經常改變需求的顧客就顯得不怎麼合適。

Cooley 將網站探勘（Web Mining）技術技術應用在電子商務上，用它來探勘出使用者的瀏覽樣式，以便獲得虛擬的使用者偏好檔，進而從事目標行銷。更進一步的說，探勘的目的是找出使用者瀏覽網站的使用樣式（Usage Pattern），進而預測使用者的瀏覽路徑，以達到協助使用者瀏覽網站的目的。本文將從網站經營者的角度，探討如何有效地協助使用者獲取決策過程中有關的資訊，並以網站探勘技術爲基礎 23，改善預測使用者行爲的精確度，以及透過推薦網頁分類的方式增進推薦的實用性。

貳、使用者資訊需求分析

一、個人決策資訊需求理論

瞭解消費者行爲模式，有助於提供使用者所需要的功能，協助消費者購買 2。Kalakota 提出由消費者的角度規劃出產品/服務採購時消費者的行爲過程 1。此模型可簡單區分爲採購前決策、採購以及採購後互動。

我們將焦點專注在採購前決策這階段。從消費者的角度來看，任何重要的購買行爲都包含了某種程度的採購前思考，而消費者在思考時會注意任何有助於抉擇的各種新舊資訊，這便是電子目錄在設計時所應該要注意的。因此，電子目錄必需提供不同的服務機制給不同類型的使用者。

市場研究對於採購行爲做了下列的區分：完全計畫採購、一般計畫採購、記憶採購及完全無計畫採購等。若要讓使用者由瀏覽者變成購買者，也就是使用者由發現到付款(Discovery-to-Payment Cycle)24，必須要能夠提供消費者足夠詳細的內容。因此，除了要提供豐富的內容以及設計良好的網站流程外，還必需協助使用者減少蒐集資訊所需的時間，並可以依據使用者目前的知識呈現，推薦不同的資訊，幫助使用者確認需求。

二、企業採購的資訊需求分析

在 B2B 的電子商務模式下，企業採購是供應鏈中的一個鏈結 5。電子市集的出現是一個關鍵性的應用。電子市集對買方而言，可以降低採購成本、內部運作成本與縮短採購時間；對供應商則可以強化產品推廣效益、確定新價格訂定機制等。

不論是事先規劃的直接採購或是非預先規劃的間接採購，需求認知的錯誤往往是造成不正確採購的主要原因。因此，對於企業採購而言，設計良好的電子目錄應該要提供快速且方便的搜尋功能，讓使用者確認需求，當然也需要顧慮到消費者的行爲模式。

藉由上述的分析，我們試圖設計出一樣技術來協助使用者達到下列目的：

1. 提供相關的參考資訊，幫助使用者確認需求；
2. 提供有效的互動服務，幫助使用者瞭解產品；
3. 記錄及分析消費者的行爲模式，幫助使用者做有效率地資訊搜尋。

參、網站推薦服務之相關研究

在電子目錄中，最常提供的兩種服務便是瀏覽與搜尋。瀏覽是指顧客並不清楚自己的需求，透過逐步地瀏覽以確認自己的需求；搜尋則是消費者清楚自己的需求 4，主要從事更進一步的選擇與比價的動作。瀏覽適合大多數的使用者，具有下列優點：

1. 循序式瀏覽，具有方向性，可逐步找到所需資訊；
2. 結構式瀏覽，階層式分類產品資訊，略去不相關的資訊。

而缺點則有：

1. 超連結過多時，容易造成使用者迷失方向 4；
2. 分類定義不同時，使用者容易找不到資訊。

爲了避免上述缺點，如何更有效地在使用者瀏覽階段提供指引？這將是網站設計上值得努力的方向。

一、個人化瀏覽協助的相關技術

在資訊過濾模式中，我們會先定義使用者偏好檔以記錄使用者的喜好 4。系統透過比對新的資訊與使用者的偏好檔來決定是否要推薦新資訊給使用者 21。

在這項技術中，最困難的階段不在於如何提供相關的資訊，而是偏好檔中如何真實地反應使用者的資訊需求 4。以下，我們將介紹三種的資訊過濾技術：(1)以資訊內容爲主的資訊過濾；(2) 以合作式爲主的個人偏好過濾；(3) 以網站探勘爲主的個人化推薦服務。

（一）以資訊內容爲主的資訊過濾

傳統的 IR，IF 領域的學者，便是利用內容分析的方式，建立個人的偏好檔。通常是利用一個或多個不相關的向量來代表使用者常用的關鍵字或是分類。這種方法有幾個缺點：

1. 靜態式改變偏好檔，可能錯失使用者的潛在需求；
2. 使用者在不同情境可能扮演不同角色，如此靜態的偏好檔就無法做出有效的推薦；
3. 靜態式偏好檔無法隨著使用者的瀏覽行爲自行改變設定；

4. 明示 (Explicit)的回饋機制，容易造成使用者操作上的困擾。

（二）以合作式爲主的個人偏好過濾

合作式過濾的方式(CF)並不是直接針對內容項目做比對，而是找出相同喜好的使用者，推薦他們共同喜好的內容項目 2121。如此，可以針對非文字的資訊作推薦及過濾。

合作式過濾的方式除了要克服規模（Scalability）和預先建立偏好檔兩個問題外，還需要解決另外兩個問題：確認使用者需求以及推薦正確性。若網站提供的商品類別眾多，推薦的項目可能與使用者想要的產品種類差異甚大，沒有分類的推薦可能導致不知所云。因此，推薦必須分別對不同類別產品做進一步處理。

（三）以網站探勘爲主的個人化推薦服務

網站探勘與合作式過濾有著異曲同工之效。不同的是，合作式過濾必須要先知道使用者的偏好檔，而網站探勘乃藉由觀察使用者的瀏覽行爲，利用叢集技術得到使用者的虛擬偏好檔，進而做出推薦。利用網站探勘技術推薦的優點是 16，免去編撰使用者偏好檔的困難與麻煩；至於困難點則是，如何瞭解使用者的瀏覽意圖？

二、探勘網站瀏覽模式的程序

網站探勘的程序包括：(1)資料的前置處理；(2)相似度比對；(3)叢集處理；(4)推薦策略。其中以資料的前置處理最爲麻煩，理由是代理伺服器、瀏覽器快取機制，會無法正確識別單獨的使用者以及 User Session。

（一）資料的前置處理

Cooley 分別從網站設計者和使用者兩個角度看資料的前置處理工作 9，網站設計者必須將網站資料做分類並配置展現出來。從使用者的角度來看，便是利用使用者的瀏覽記錄，分析使用者的瀏覽行爲並瞭解其意圖，這部分的工作包括了：資料清除、使用者識別、Session 識別、路徑完成以及交易識別。

交易識別的目的是將 User Session 切成多個有意義的路徑，以利後續的探勘工作。Cooley 利用 MFR78 處理交易的切割。

（二）相似度比對

相似度比對能自動分類未知的使用者瀏覽行爲，這裡我們將使用者的瀏覽行爲分爲三類：

1. 以內容爲主的相似度比對：
   依據瀏覽網頁的內容，判斷使用者是否瀏覽相關的資訊。
2. 非順序性的相似度比對：
   僅考慮使用者看過哪些網頁而不考慮瀏覽的先後順序及網頁內容。一個網站若產品資訊多樣化，不考慮瀏覽順序或不把瀏覽路徑做適當切割，推薦時，並無法確認是相關產品推薦，僅能說明是「具有同樣嗜好的推薦」。
3. 順序性的相似度比對：
   同時考慮瀏覽的網頁以及瀏覽的順序。

（三）概念的確定

確定使用者是否瀏覽相同概念（**Concept**）的資訊是件不容易的工作，這個問題與交易的切割、相似度的比對有一定的關係。Cooley 利用 MFR 雖能找出使用者有意義行爲，但還是有以下的缺點：

1. 忽略倒退型網頁對於瀏覽的意義；
2. 無法知道使用者在同一概念中，有哪些連結同時也被瀏覽了。

因此，本文將改良切割法，使它能確定使用者的瀏覽目的，並能解決上述 MFR 的不足。

肆、系統架構

一、系統概觀

圖 1 的系統架構主要參考 WebMiner 系統 1718，，改良 MFR 交易切割法、FM 相似度比對以及推薦策略。本系統同時具備離線及線上兩種處理方式，以符合推薦的需要。離線部分主要是蒐集並處理使用者在網站某段時間內的瀏覽路徑，以及將使用者的瀏覽行爲作分類，產生虛擬的使用者偏好檔。工作包括：網站架構挖掘、網頁分類、存取記錄的前置處理、Session 識別、交易切割、交易叢集。

線上處理主要爲推薦引擎處理網頁推薦的工作。工作包括：追蹤線上使用者的 Session、比對資料庫中的虛擬偏好檔、找出最適合的偏好檔、進行推薦。本文僅針對網站模式化、交易識別、相似度比對以及推薦策略進行說明。

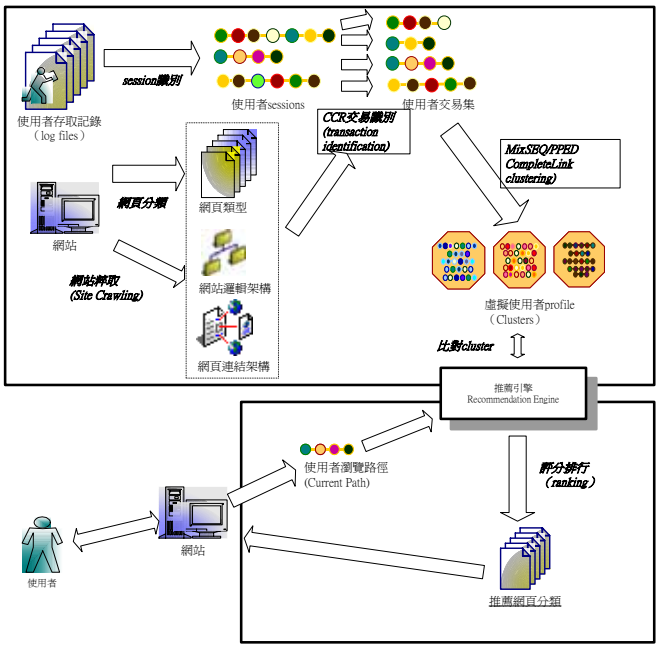

圖 1 系統架構圖

二、網站模式化

廣義而言，我們可以將電子商務網站視爲電子目錄 19。我們可以簡單地視網站的 URL 路徑爲電子目錄的特例。路徑代表了某種程度的分類架構，而每一個網頁則代表分類型錄、產品型錄或產品內容介紹。

三、使用者存取記錄前置處理

我們直接利用網站伺服器的存取記錄作爲使用者的瀏覽記錄。使用者的 Session 識別可以透過 IP 、Agent(瀏覽器)或作業系統來判定，並參考每個網頁的存取時間判斷是否爲另一個 Session 的開始。當使用者兩頁的存取間隔超過一定的時間，就視爲另一個 Session 的開始。網頁補足(Path Completion)可以透過網站的連結架構來判斷是否有連結。此演算法主要是檢查相鄰兩頁間的是否存在連結，若找不到連結則從歷史記錄中尋找最近有連結的網頁。若還是找不到連結，就表示這是另一個 Session 的開始。

四、利用概念限制做交易識別

前一步驟雖然能找出每一個使用者的 Session，但是每一次的使用者 Session 中可能又包含了多個目的或不同概念。在本文中，我們採用以概念限制參考的交易識別法(Conceptual Constrained Reference ,CCR)，希望能夠找出具有共同瀏覽目的的瀏覽行爲叢集，來作爲往後使用者瀏覽時的推薦指引。

（一）概念的定義

首先，我們必須先將網站分爲多個概念，交易識別時，我們可以依據瀏覽的概念切割使用者 Session。概念的劃分由網站設計者依據網站特性來決定，爲了能讓網站設計者，依據網站的特性與實際的探勘需求決定概念的範圍，本文透過共同路徑 CLevel 的定義來決定概念區域。

共同路徑

$CLevel(U_i, U_j)$ ≡ 兩網頁具有相同路徑的數目

我們將兩個網頁的 URL 路徑相比較，由根目錄開始算起相同的目錄數目，就是共同的路徑數。

最小的共同路徑數目：minCLevel

minCLevel 表示每一個概念內的網頁至少需要的最小相同路徑數目。圖 2 表示以 minCLevel=2 的網站架構圖，每一個節點代表一個 URL 路徑。虛線框則表示所形成的概念。虛線框內的節點表示在此概念的目錄，此框內所有目錄內的網頁均爲同一個概念。

（二）CCR 交易定義

確定了如何劃分網站概念後，就可以藉此概念的劃分進行交易的識別。使用者在某特定概念瀏覽網頁的集合稱爲一個交易。在相同概念下瀏覽具有相同的目的，點選的網頁與點選的順序越相近，其瀏覽行爲越相似。交易 T 是依據存取時間排列的網頁集合：

$$T：x_1 \to x_2 \to x_3 \to ... \to x_n$$

必須滿足下列條件：

1. 交易內任意兩個網頁 x_i, x_j，均在同一概念中而且要：

$$CLevel(x_i, x_j) >= \min CLevel$$

2. 網頁的類型可以是導覽型網頁或內容型網頁。

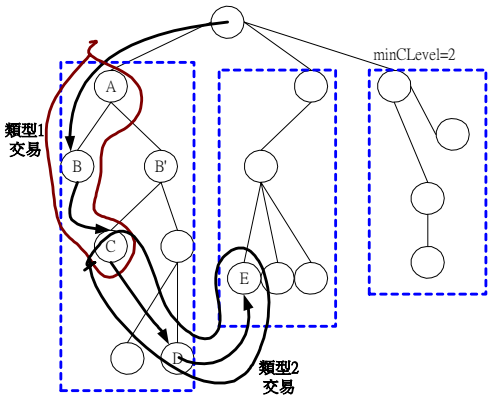

圖 2：以 CCR 處理的交易識別

交易類型：

為了瞭解使用者瀏覽網頁之間的關連性，並且避免交易過大，我們將 CCR 的交易類型分為兩種：

- 類型 1

  主要是記錄使用者「第一次」或「重新」進入此概念時，瀏覽的次概念網頁；

- 類型 2

  同一概念下，使用者除了第一次瀏覽的次概念外，瀏覽其他次概念時，都會被視為類型 2 的交易。

（三）跨概念處理

當使用者離開目前概念進入另一個概念時，會形成另外一個類型 1 的交易，此時使用者的瀏覽目的與之前會有所差異。然而，使用者在瀏覽網站時，也有可能在不同的概念中，發現相關訊息而跨概念瀏覽。為了避免相關資訊可能跨不同概念，交易時會記錄到新類型交易的第一頁網頁，這樣可以記錄使用者除了看此分支，還會進行其他類的交易瀏覽，以便用來推薦使用者「相關產品」或「其他概念資訊」。

圖 2 說明兩種不同交易的形成，使用者一進入網站先瀏覽 A、B 概念下的網頁，之後，瀏覽 B'次概念下的網頁。此時，A,B,C 會形成類型 1 的交易。接著，瀏覽 C、D 後，又瀏覽另一個概念中的 E 網頁。此時，C,D,E 會形成類型二的交易。使用者若繼續瀏覽的話，則又會再形成另一個類型 1 的交易。

（四）CCR 交易識別優點

相較於 MFR，CCR 具有下列的優點：

1. 確定使用者在某特定的概念下瀏覽的相關連結：
   在一個交易中，可以瞭解使用者在哪一個網址後做反覆的瀏覽以及分支，瞭解使用者看了此產品，選擇了哪些同類型的連結，避免無法瞭解倒退（Back Reference）的瀏覽行為。
2. 減少叢集處理所需要的時間：
   將交易類型區分為類別 1 及 類別 2 兩種，通常類型 1 的交易數目較類型 2 少，因此我們可以分開對這兩種交易類型的瀏覽行為做分析，提供推薦服務。

五、Session 模型

Session 是由許多不同的特徵矩陣所組合而成，在此我們使用 20

- 點選矩陣(Hit Matrix)
  計算相鄰網頁依序被瀏覽的次數，並加總。
- 序列矩陣(Sequence Matrix)
  找出兩網頁出現的相對位置並以算數平均數的方式平均。

使用者可能透過不同的路徑觀看相同的內容，若完全將瀏覽的順序列入，則會使得不相似度提高。因此，我們在決定序列矩陣時，會將網頁類型納入考慮，用以降低被提高的不相似度。本研究所提出的混合型順序性（MIXSEQ）比對，乃藉由考慮網頁類型，給予不同的順序編碼。

混合型順序性編號最大的差異在於某一瀏覽頁 Hub Page，反覆瀏覽此頁的內容型網頁連結時，這些內容型網頁的順序均相同。內容頁之後的網頁，順序值會繼續增加，直到重返 Hub Page，再從 Hub Page 的順序繼續編號。圖 3 便說明了混合型編號後的結果。

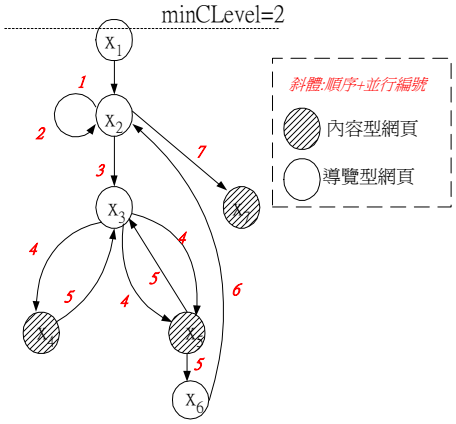

圖 3：MIXSEQ 編號後的結果

$(x_1,x_2)=1$ ， $(x_2,x_2)=2$ ， $(x_2,x_3)=3$ ，

$(x_3,x_4)=4$ ， $(x_4,x_3)=5$ ，

$(x_3,x_5)=\frac{4+4}{2}=4$ ， $(x_5,x_3)=5$ ，

$(x_5,x_6)=5$ ， $(x_6,x_2)=6$ ， $(x_2,x_7)=7$

之後，我們會產生兩個矩陣，分別爲點選矩陣 $M_{Z^2}^{H}$ 、序列矩陣 $M_{Z^2}^{S}$ ，其中z爲此網站的所有網頁數目。假設範例中的 $Z=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$。 $M_{7^2}^{H}$ 矩陣中的 x_{ij} 元素代表網頁 x_i 到網頁 x_j 的瀏覽次數，而 $M_{7^2}^{S}$ 矩陣中的 x_{ij} 元素則代表網頁 x_i 到網頁 x_j 的瀏覽編號。底下爲圖三的點選矩陣值和序列矩陣值：

$$M_{7^2}^{H}=\begin{bmatrix}0&1&0&0&0&0&0\\0&1&1&0&0&0&1\\0&0&0&1&2&0&0\\0&0&1&0&0&0&0\\0&0&1&0&0&1&0\\0&1&0&0&0&0&0\\0&0&0&0&0&0&0\end{bmatrix}$$

$$M_{7^2}^{S}=\begin{bmatrix}0&1&0&0&0&0&0\\0&2&3&0&0&0&7\\0&0&0&4&4&0&0\\0&0&5&0&0&0&0\\0&0&5&0&0&5&0\\0&6&0&0&0&0&0\\0&0&0&0&0&0&0\end{bmatrix}$$

六、叢集模型

將交易進行群聚後，我們使用「算數平均數」算出叢集的質心，最後的結果，一個

叢集是由多個特徵矩陣的總和：

$$C = \left\{ M^H, M^S \right\}$$

每一個特徵矩陣爲交易的特徵矩陣之算數平均：

$$M^F = \frac{1}{N} \sum_{i=1}^{N} M_i^F$$

N 爲叢集所包含的數量(Cardinality)。

七、相似度比對

相似度比對是用來衡量兩使用者行爲的相似度，透過此方式，將「具有相似行爲」的 Sessions 叢集處理在一起。本文採用 Shahabi 等人提出的 FM 定義 20 將 Session 及叢集用特徵矩陣做不相似度的比對，不相似度的總和爲個別不相似度的加權平均值：

$$D^F = D^H \times w_h + D^S \times w_s$$

在本研究中，僅使用網頁點選、網頁順序兩種特徵，因此，我們需要爲兩矩陣分別設定不相似度的權重值W_s, W_h。

不相似度測量，我們採用 Project Pure Euclidean Distance (PPED)，以避免 Session 比對叢集時，過度衡量不相似度，其公式爲：

$$PPED\left(\vec{A}, \vec{B}\right) = \left(\sum_{i=1, a_i \neq 0}^{N} (a_i - b_i)^2 \right)^{\frac{1}{2}}$$

八、叢集處理

本研究使用 Hierarchical Method 中的 Complete-Link Clustering 建立群集。此演算法的優點是兩叢集間物件的距離均相差最大。進行階層式叢集處理後，我們必須選擇某一個門檻值(TDC)值，將叢集合併成較大的群組，減少叢集數，以方便之後的應用。

九、推薦方法

推薦引擎負責利用離線時建立的樣式，進行線上的推薦作業。它會產生一組排序過的推薦資料列表供使用者參考。推薦引擎記錄線上使用者 Session s 並與叢集 C 進行比對，得到與每一個叢集的比對值。由於我們想得到相似度值而非不相似度值，因此我們要定一個上限 maxPPED 使得

$$match(s,c) = MaxPPED - PPED(s,c), c \in C$$

十、推薦得分計算

在網站中，導覽型的網頁僅作爲瀏覽相關網頁的指引，因此，直接推薦內容型網頁，比推薦導覽頁網頁更具有意義。推薦叢集中，叢集的資料結構亦爲 FM，包含點選矩陣與序列矩陣，因此我們要推薦矩陣中，每一個 segment(u1,u2)的 u1 或 u2 網頁。首先，我們定義 DF (Distance Factor)，用它來計算推薦網頁與 Session 的距離：

$$DF(u) = \begin{cases} \log(\text{maxWeight} - |SeqWeight(u_1,u_2) - n|) + 1, \text{if } u_1 = u \vee u_2 = u \wedge u \notin S \\ 0, \text{if } u \in S \end{cases}$$

因此，內容型網頁 u 的推薦得分計算如下：

$$\text{Rec}(u^c) = match(s,c) \times \frac{1}{N_u} \sum \sqrt{HitWeight(u_1,u_2) \times DF(u)},$$
$$\forall u_1 = u \vee u_2 = u, u \in c$$

十一、推薦資訊分類

若是資訊系統提供的服務回應，是讓使用者難以解釋的，將會造成顧客的反感 3。因此，爲了讓使用者瞭解推薦的目的，我們將推薦分爲三類：

- 同類產品推薦

$$\text{REC}_2(s) = \left\{ u_j^c \middle| c \in C \wedge typeof(u_j^c) = content \wedge CLevel(s,u_j) > \min CLevel \right\}$$

供使用者做產品選擇、比較時的參考，

可以推薦這類產品中，使用者最常瀏覽的產品資訊。或者是使用者瀏覽此類產品看了哪些資訊。

- 相關產品推薦（同概念產品推薦）

$$\mathrm{REC}_2(s)=\left\{u_j^c \middle| c\in C \wedge typeof(u_j^c)=content \wedge CLevel(s,u_j)=\min CLevel\right\}$$

當使用者看了某產品後，推薦其子產品的資訊或相關訊息。

- 其他相關概念資訊推薦

$$\mathrm{REC}_3(s)=\left\{u_k^c \middle| c\in C \wedge typeof(u_k^c)=content \wedge CLevel(s,u_k)<\min CLevel\right\}$$

透過此類推薦，使用者會瀏覽不同概念下的資訊，同時推薦給其他使用者做參考。

將推薦網頁分爲三類之後，分別依照得分排序產生列表，我們可以選擇各類內容網頁推薦的排名門檻值N_1, N_2, N_3，將同類產品網頁排名前N_1名的網頁、相關產品網頁前N_2名的網頁、其他概念網頁前N_3名的網頁都推薦給使用者。

伍、實驗與討論

一、模擬環境設計

本研究爲了使實驗更具真實性，我們採用 taiwan.cnet.com 的網站架構及內容作爲我們的實驗平台。同時，模擬使用者透過瀏覽器模擬網站的情境。在此我們參考 8 的實驗方法，設定參數來模擬瀏覽路徑。

產生一組虛擬的瀏覽路徑作爲之後模擬之資料集，參數設定如表 1。在這次的虛擬瀏覽路徑中，我們將P_d設爲 0，同時，均預設使用者先進入硬體專區，以避免因爲網站架構的不完整，導致產生的瀏覽路徑有誤。在不相似度衡量方面，設點選矩陣權重W_h=0.8，序列矩陣權重W_s=0.2。

表 1　實驗參數說明及設定

參數名稱	說明	設定值
minCLevel	決定網站的概念範圍。	2
P_a	使用者瀏覽產品比較、詳細資訊的機率值。	0.6
P_c	使用者瀏覽相關產品資訊的機率。	0.2
P_d	使用者瀏覽其他概念資訊的機率。	0
P_b	使用者退回前一頁的機率。	0.2
t1, t2	同時開啓視窗的最少、最多值。	1, 1
L,V	User Session 的長度：$L \pm V\%$	20, 0.5
Session 數		1000

二、實驗設計與方法

本實驗要測試不同的交易識別法、相似度比對，對預測品質的影響以及統計叢集分析後的相關統計。因此，我們要評估的項目主要爲預測的品質。包含：(1) MFR,CCR 交易識別法的預測品質比較；(2)不同交易識別法的與門檻值的關係；(3)CCR 交易識別法與 MIXSEQ 的預測品質比較。

我們將使用者的瀏覽記錄分爲兩個資料集，前 80%爲訓練資料集，後 20%測試資料集 6。訓練資料集的目的是建立使用者行爲叢集，而在測試資料集的使用者則視爲線上使用者。接著，我們將每位使用者的（後續）瀏覽行爲分爲觀察階段以及預測之後的行爲。

三、評估標準

在這裡我們改良 IR 領域常用的評估方

法：

- 回收率（Recall）

$$recall = \frac{\text{目前提供的相關網頁}(|RA|)}{\text{之後點選的網頁}(|AnsSet|)}$$

- 精確率（Precision）

$$precision = \frac{\text{目前提供的相關網頁}(|RA|)}{\text{所有提供的網頁}(|RelSet|)}$$

- 單一的評量值 HM

$$HM = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

HM 值越高，表示回收率與精確率同時也很高 4。

四、實驗結果比較與分析

（一）MFR 與 CCR 之比較

以概念爲限制的交易切割方式，經過叢集分析後預測使用者瀏覽行爲，在較低的門檻表現很出色，有較高的預測品質。門檻越高，兩者的表現越趨近（如圖 4）。這說明了 CCR 此種識別方式在低門檻時，預測品質較高且更能確定使用者的瀏覽意圖。

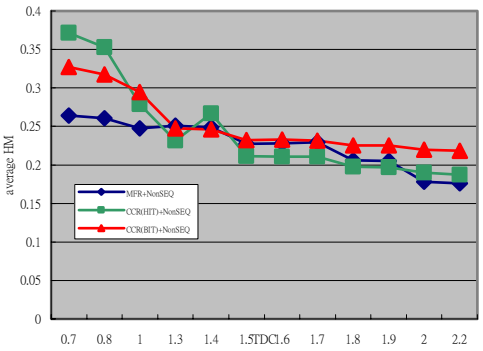

圖 4 MFR 與 CCR 在 HM 上的比較

另一個實驗中，我們並不加總 CCR 交易內網頁的點選次數，只看是否有出現（稱爲 CCR-BIT），然後再進行一次叢集處理。我們發現，在低門檻值的時候，CCR-BIT 的預測品質依舊比 MFR 來的好，但是比 CCR-HIT 來的差。然而在高門檻值的時候，HM 值降低的速度比 MFR 及 CCR-HIT 均來的慢。會導致這樣的結果，我們推論是由於 CCR 交易識別計算點選次數（HIT 數）擴大了不相似度所致。

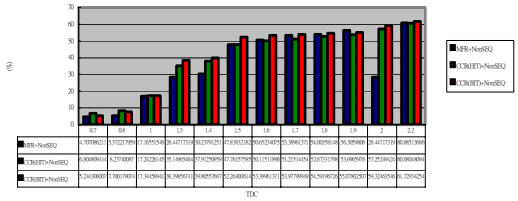

圖 5 MFR 與 CCR 在成功查詢率上的比較

此次實驗共做了 3909 次。但並不是每一次的查詢都能夠找到適合的叢集，因此我們比對兩者的成功查詢率。大致上兩者的成功查詢率相差不遠，不過 CCR 在 HM 較高時，成功查詢率也較高（如圖 5）。

（二）順序型編號與混合型編號之比較

我們均以 CCR 方式切割，並使用順序性編號與混合性編號做比較。如圖 6 所示，利用混合型編號會使得預測的 HM 略下降，但在另一方面，查詢的成功率在特定門檻值下均高出許多（如圖 7）。

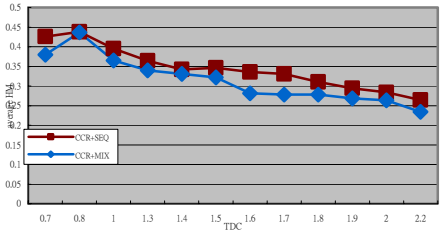

圖 6 CCR+SEQ 與 CCR+MIX 對 HM 的影響

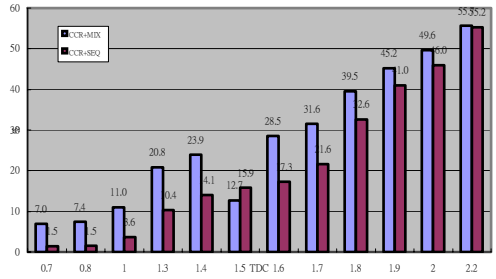

圖 7 CCR+SEQ 與 CCR+MIX 對成功查詢率的影響

再看特定門檻值下的叢集數（圖 8），順序型的叢集數一直都比混合型的叢集數多，這代表用順序型的編號不易使叢集叢聚，雖然 HM 較高，但是所要付出的比對成本也就相對提高。對於更多使用者的商業網站而言，其效率以及規模擴展性均難免會受影響。

綜合而言，我們提出的 CCR 交易識別法，在查詢成功率有較好的效果。不過，HM 值僅在門檻較低的時候，表現較好。而混合型順序性的相似度比對，對於降低相似度有達到其效果，使得叢集數減少。不過，也因爲如此，當門檻值大時，容易形成較大的叢集，因此預測的品質也隨之降低。這在未來的研究中，有相當的空間可以繼續做探討與改進。

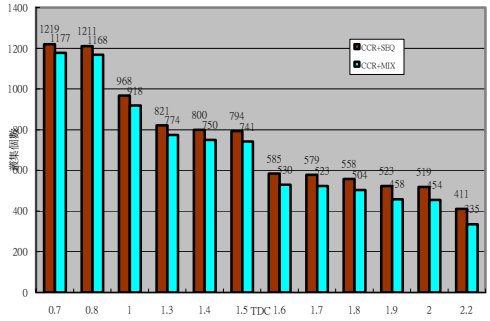

圖 8 CCR+SEQ 與 CCR+MIX 對叢集數的影響

陸、結論與未來方向

使用者個人化服務將是電子商務中最重要的一環，透過推薦系統的協助以及網站探勘的技術，將有助於提昇網站服務的品質，同時，也是電子商務網站達到差異化的競爭策略。

我們利用簡單的參數 minCLevel，讓網站管理者依據電子目錄特性訂定概念範圍，以了解使用者的瀏覽目的。並透過 CCR 爲依據的概念限制，識別出主要產品的瀏覽路徑以及相關產品的瀏覽路徑，同時避免因交易太長而無法確定使用者瀏覽目的。相似度測量方面，結合順序性與非順序性比對提出 MIXSEQ 比對，藉此減少因爲瀏覽先後順序而增加的不相似度。模擬結果發現，利用此方法計算瀏覽順序，能有效減少叢集的數

目，對於推薦系統的規模變化很有幫助。同時我們也將推薦的網頁予以分類，讓使用者清楚推薦的原因，這有助於提升推薦的可用性。

電子目錄上的推薦服務還有很大的研究空間，如何預測使用者的意圖便是個棘手問題？除此之外，如何過濾無意義瀏覽以及提供更精確的服務？也是未來繼續研究的方向。

1. 參考文獻

1. 陳雪美譯，「電子商務概論（Kalakota, R. and Whinston, A. B 原著）」，和碩科技文化有限公司，民國 88 年。
2. 魏志仲，「電子商務聖經：電子商務概念」， 第三波，民國 89 年。
3. Aberdeen Group, Inc. “Interactive Customer Care： Enriching the Self-Service Experience with Automated Agents,” http://www.Aberdeen.com, November 2000.
4. Baeze-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval*, ACM Press Book, Addison Wesley, 1999.
5. Baron, J. P., Shaw, M. J. and Bailey, Jr. A. D. “Electronic Catalogs in the Web-Based Business-to-Business Procurement Process,” In *Handbook on Electronic Commerce*, M. Shaw (ed.), Springer Press.
6. Breese, J. S., Heckerman, D. and Kadie, C. “Empirical Analysis of Predictive Algorithms for Collaborative Filtering,” *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, July 1998, pp.43-52.
7. Chen, M. S., Park, J. S. and Yu, P. S. “Data Mining for Path Traversal Patterns in a Web Environment,” *Proceedings of the 16th International Conference on Distributed Computing Systems*, May 27-30 1996, pp. 385-392.
8. Chen, M. S., Park, J. S. and Yu, P. S. “Efficient Data Mining for Path Traversal Patterns,” *IEEE Trans. on Knowledge and Data Engineering*, 10 (2), March 1998, pp. 209-221.
9. Cooley, R., Mobasher, B. and Srivastava, J. “Data Preparation for Mining World Wide Web Browsing Patterns,” *Knowledge and Information Systems*, Vol. 1, No.1, 1999, pp. 5-32.
10. Cooley, R., Mobasher, B. and Srivastava, J. “Web Mining： Information and Pattern Discovery on the World Wide Web,” *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, November 1997.
11. Jaideep, S., Cooley, R., Deshpande, M. and Tan, P. N. “Web Usage Mining： Discovery and Applications of Usage

Patterns from Web Data," *SIGKDD Explorations*, Vol.1 No.2, 2000, pp.12-23..

12. Jung, J., Kim, D., Lee, S. G., Wu, C. and Kim, K. "EE-Cat： Extended Electronic Catalog for Dynamic and Flexible Electronic Commerce," 1999.
13. Kappel, G., Retschitzegger, W. and Schwinger, W. "Modeling Customizable Web Applications A Requirement's Perspective, " *Kyoto International Conference on Digital Libraries*, 2000, p.387.
14. Keller, A. M. "Smart Catalog and Virtual Catalogs," *Workshop on Electronic Commerce following CIKM*, December 1994.
15. Kim, J. G., Lee, E. S. K. "Intelligent Information Recommend System on the Internet," *IEEE*, 1999.
16. Kohavi, R. "Mining E-Commerce Data： The Good, the Bad, and the Ugly," *KDD 2001's Industrial Track*, 2001.
17. Mobasher, B., Cooley, R. and Srivastava, J. "Creating Adaptive Web Sites Through Usage-Based Clustering of URLs," . *IEEE Knowledge and Data Engineering Workshop (KDEX'99)*, 2000.
18. Mobasher, B., Cooley, R. and Srivastava, J. "Automatic Personalization Based on Web Usage Mining," *Communications of the ACM*, Vol.43, No.8, 2000,pp.142-151.
19. Segev, A., Wan, D. and Beam, C. "Designing Electronic Catalogs for Business Value： Result of the CommerceNet Pilot," CMIT Working Paper 95-WP-1005, UC Berkeley, 1995.
20. Shahabi, C., Kashani, F. B., Farugue, J. and Faisal, A. "Feature Matrices： A Model for Efficient and Anonymous Mining of Web Navigations," http://www.usc.edu/dept/cs/technical_reports.html, 2000.
21. Shardanand, U. and Maes, P. "Social Information Filtering for Music Recommendation," S.M. Thesis, Program in Media Arts and Sciences, Massachusetts Institute of Technology, 1994.
22. Shardanand, U. and Maes, P. "Social Information Filtering： Algorithms for Automationg "Word of Mouth"," *Proceedings of CHI'95 Conference on Human Factors in Computing Systems*, ACM Press, 1995.
23. Troussse, B. "Evaluation of the Prediction Capability of a User Behaviour Mining Approach for Adaptive Web Sites," *Proceedings of the 6th RIAO Conference - Content-Based Multimedia Information Access*, Paris, France, april, 2000.
24. Wong, W. T. and Keller, A. M.

"Developing an Internet Presence with On-line Electronic Catalogs," CMIT Working Paper 95-WP-1005, UC Berkeley, October 9, 1994.

作者簡介

林熙禎

美國密蘇里大學羅拉分校 (University of Missouri at Rolla)電腦科學博士，現任國立中央大學資訊管理系副教授，研究興趣領域爲企業電腦網路及軟體代理人安全。

許益誠

國立中央大學資訊管理系碩士，目前服國防役中。