結合關聯規則與深度優先搜尋於找尋使用者瀏覽路徑

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

摘要

網頁探勘在網頁資訊量爆炸的時代裡是一個強大且有用的工具，其中網頁使用探勘用於分析使用者在瀏覽網站上的使用行為，以了解使用者相關的瀏覽資訊是網頁使用探勘的重要目標之一。對企業而言，使用者之最常存取路徑可以用來建立適合的網站連結、進行促銷活動、改善行銷策略、產品宣傳等，藉此提升企業的競爭優勢。在本論文中我們將提出一個結合關聯規則與深度優先搜尋的網頁探勘技術，可以提高單獨使用關聯規則來搜尋使用者之最常存取路徑時的效率。

關鍵詞：網頁使用探勘、使用者瀏覽路徑、關聯規則、深度優先搜尋。

1. 前言

隨著全球資訊網 (World Wide Web)的爆炸性發展，在網際網路(Internet)上有超過數十億的網頁(Web Pages)存在並且持續的成長中，這也意味著將會有愈來愈多的資訊在網路上傳遞。要如何從這如此龐大的資料(Data)中找出有意義的資訊(Information)，並透過有系統的方法將從網際網路中所得到的資訊組織成為有用的知識(Knowledge)將是一個值得研究的議題。

在全球資訊網的興起之下使得企業改變了交易方式，也促進電子商務(Electronic Commerce)的快速發展。對企業而言，網際網路提供了一個交易平台即電子商務網站(EC-Web)，可以直接、快速的與客戶接觸且能夠花費較少的成本。相較於傳統的交易方式，在現今以網際網路為平台的電子商務交易模式下，對企業來說顧客關係管理(CRM，Customer Relationship Management)就顯得愈來愈重要了[4]。

顧客關係管理的重要性在於要如何從電子商務網站中去找出使用者的使用模式(Usage pattern)，例如：使用者的行為(Behavior)、偏愛、使用模式、瀏覽路徑(Navigation path)、……等等。為了能夠找出使用者的使用模式將會面臨到一個重大的挑戰，要如何有效率地且自動地從網站上的龐大資料中找出使用者模式？此時網頁探勘(Web Mining)就會是一個強大且有用的工具。

在網際網路中網站像是一個個的容器，儲存並記錄著使用者相關的資訊，將這些資訊記錄在網站的記錄檔(Log File) 裡，所以利用網頁使用探勘這個方法，先要能夠收集這些網站中的記錄檔，並且可以自動地取出使用模式，再針對使用模式做分群(Clustering)。假設每個使用者會有不同的使用模式，但是對於某一特定產品或主題資訊有相同興趣的使用者，則可以發現不同的使用者之間是具有相似的使用模式(例：瀏覽路徑)。若針對這些使用模式做分群，所得到的資訊將會是一種具有價值的資訊。

當取得了使用者使用模式並加以分析之後所得到的資訊，對企業而言就可以做到大量客製化(Customization)、個人化(Personalization)、建立適合的網站，改善行銷策略、促銷活動、產品的提供。因此，更可以建立良好的顧客關係管理、改善顧客滿意度、保留顧客忠誠度、取得市場情報(Market intelligence)、預測市場趨，藉此提升企業的競爭優勢等。

本論文的架構如下，第二節為為相關研究的目前成果，我們的研究方法在第三節會做詳細的說明，最後第四節是未來工作及結論。

2. 相關工作

全球資訊網隨著網際網路的快速成長也日益增加，網際網路就像是一座金礦山埋藏許多有用的資訊。網頁探勘就是為了能過有效率並自動地取出資訊進而形成知識的一種強大且有用的工具。網頁探勘一辭是在 1996 年由Etzioni所提出之後，主要是應用資料探勘技術

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[6]。

網頁探勘大概可分為三個部分(圖 1) [10]：

(1)網頁內容探勘(Web Content Mining)

(2)網頁結構探勘(Web Structure Mining)

(3)網頁使用探勘(Web Usage Mining)

網頁內容探勘主要能夠從網頁內容中取出有用知識。網頁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。網頁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。

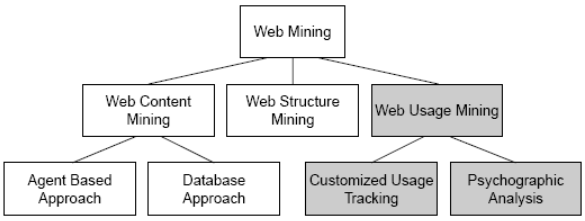

圖 1、網頁探勘的分類[10]

2.1 網頁使用探勘

網頁使用探勘從網站的記錄檔中找出使用者存取網頁的相關資訊。當使用者瀏覽網站時的所有使用行為都會被明確的記錄在記錄檔中，包括使用者名稱、IP位址、日期、請求時間、……等等。這些記錄檔主要由四種記錄檔型態所構成(圖 2)[4][10]：

(1)處理記錄(Access Log)

(2)錯誤記錄(Error Log)

(3)來源記錄(Referrer Log)

(4)代理人記錄(Agent Log)

Related part of the line from log file	Its description
172.16.12.127	Address or DNS
–	RFC931 (or identification)
–	Authuser
[10/Apr/2003:12:16:37+0300]	TimeStmp
'GET/home/images/index_on.gif http/1.1'	Http request
404	Status code
304	Transfer volume
'http://www.selcuk.edu.tr/'	Referrer URL
'Mozilla/4.0 (compatible; MSIE 5.0; Windows 98; DigExt)'	User agent

圖 2、網站記錄檔內容之描述[10]

2.2 網頁使用探勘流程

網頁使用探勘之處理流程(圖 3)如下列步驟所述，從網站記錄檔中建立各個使用者的 Session 並辨識出序列模式 (Sequential Patterns)，再對序列模式做關聯規則之搜尋並透過支持度及信賴度所得到的強規則。最後由強規則所建立的圖形則可以從中得到使用者之最常存取路徑，。

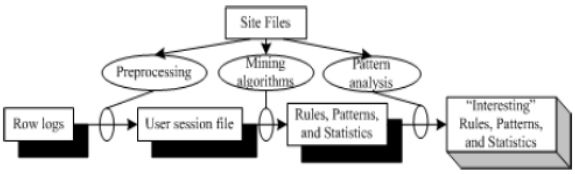

圖 3、網頁使用探勘流程[5]

2.3 關聯規則

關聯規則是基本資料探勘技術也是最多應用在網頁探勘的技術之一，其中Apriori演算法也是最常見的關聯規則演算法，此演算法 在1994 年由R.Agrawal、R.Srikant所提出[3]。關聯規則的意義為在 $X \Rightarrow Y$ 此規則中，X為規則主體經推論得到交易Y的集合 (交易意指為使用者瀏覽行為)。 $X \Rightarrow Y$ 這項規則主要在表示在一個交易中如果包含X則有可能會包含Y。若將關聯規則應用於網頁使用探勘時規則為：

"$A.html \Rightarrow B.html$"

此規則再說明當某一個使用者在瀏覽 A 網頁之後則有可能會瀏覽到 B 網頁，這種關聯規則會發生在每一個瀏覽序列模式上。

support = 2% ， confidence = 60%

在關聯規則中有兩個重要的規則支持度(support)和信賴度(confidence)，這兩個規則可以反應出所發現的規則其有用性和確定性。以規則"$A.html \Rightarrow B.html$"來說，支持度 2%即表示在需要分析的所有網站記錄檔中要有包含2%的使用者瀏覽時會瀏覽A網頁和B網頁。信賴度 60%即表示在使用者瀏覽A網頁時會有60%會從A網頁瀏覽到B網頁。如果所發現到的關聯規則符合支持度與信賴度則表示該規則是有趣的，也表示該規則是強規則(Strong)[2][6]。

所以在上述的使用者 Session 中會存在各個使用者的瀏覽路徑，在每條瀏覽路徑上就會包含許多的關聯規則。Apriori 演算法為基於關聯規則所發展出的演算法，後來則有許多的學者加以修改增進 Apriori 演算法的效能。在本研究中將 Apriori 演算法簡化到產生候選項目集合 C2 和頻繁項目集合 L2 為止，因為我們觀察

出，關聯規則在找出使用者瀏覽路徑中重要的邊是很有用的工具，但是要更進一步得到瀏覽路徑時，如果單獨使用關聯規則在搜尋時的複雜度會較高，因此我們利用搜尋二次關聯規則把重要的邊找出後即結合深度優先搜尋來找出最常存取路徑以提高網頁使用探勘之效率。

2.4 圖形結構

我們將網站視為一個無向圖(Undirected Graph)的圖形結構，因此一個圖形結構G分別由節點(Vertex，即網頁)的集合V和邊(Edge，即超連結)的集合E所構成以G=(V,E)表示[9]，並且在本研究中會將此圖形儲存在關係矩陣中。對於在圖中的方向也必需定義出(1)A→B、(2)B→A雖然為同一個邊但是方向卻不同在本研究中視為相同，雖然在瀏覽時方向不同但是在無向圖中所示都是在由相同的兩個節點所構成的邊上。(3)A→A這種有迴圈(Loop)情況則不列入考慮。經過關聯規則搜尋之後，在通過支持度及信賴度所產生的強規則(Strong)如用來建立形成圖形即為使用者之最常存取路徑(Frequent Access Path，FAP)並將其儲存在矩陣中。透過矩陣將圖形上的路徑轉換成樹狀結構，建立成為一個路徑樹(Paths Tree)(圖 4) [11]。

我們會將網站的圖形結構記錄在關係矩陣中，再利用關係矩陣建立起路徑樹。首先從 A 點這層開始並且搜尋到相鄰節點 B，再由 B 點這層繼續搜尋並且搜尋到相鄰的 A、C、D 三個節點，在此我們為了去除迴圈的部份加入限制條件，因為當由 B 點再搜尋 A 點時則會形成迴圈，所以限制條件為”當搜尋到父節點以上的節點時，判斷形成迴圈則忽略該節點”，最後從 B 點分出 C、D 兩個枝幹。往下到了 C、D 層，則分別搜尋與 C、D 相鄰的節點， 往後依此類推逐層的建立起路徑樹，最後再將路徑樹的起點連接到一個 Root 節點上成為搜尋路徑樹的起點。若結果形成了樹森(Forest)的情況，則將所有樹的起點都連接 Root 節上。

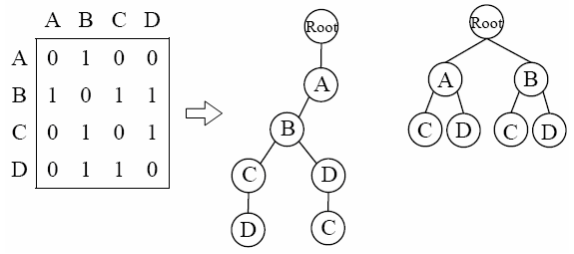

圖 4、路徑樹的建立與樹林

3. 研究方法

我們的研究方法共分成兩大部份，第一是使用關聯規則將最常瀏覽的網站架構圖篩選出來，第二是利用深度優先搜尋(Depth First Search)將每條路徑找出。

假設某網站 W 有 m 個網頁及 n 個超連結則可以得到一個網站結構如下，

有 m 個網頁則得到：

$$WP = \{wp_1, wp_2, wp_3, ..., wp_m\}$$

有 n 個超連結則得到：

$$WL = \{wl_1, wl_2, wl_3, ..., wl_n\}$$

最後網站的圖形結構則得到：

$$W = (WP, WL)$$

我們取得網站 W 的網站記錄檔，經過網頁使用探勘流程之前處理之後，從資料庫中查詢出 i 個使用者交易資料則得到：

$$T = \{t_1, t_2, t_3, ..., t_i\}$$

假設我們得到了 8 個使用者交易(圖 5) ，利用簡化的 Apriori 演算法從使用者交易中搜尋關聯規則。我們將關聯規則中兩個重要的規則支持度及信賴度分別定為 sup = 5% ， conf = 60%。

TID	Content
T1	ACEDCB
T2	BAECD
T3	CEDBA
T4	ACBDC
T5	EDCA
T6	BDE
T7	ABCED
T8	AEDC

圖 5、使用者交易資料

在 Apriori 演算法的過程中首先會產生了 C1 候選項目集合(Candidates)的結果，根據我們所定義出的支持度 sup=5%，在 C1 中所有的項目集合(Itemset)的次數(Count)都必需大於支持度即在 C1 中將全部項目集合的次數之總和與支持度相乘必需要大於 2，則得到 L1 頻繁項目集合的結果(圖 6)。

C1

Itemset	Count
{A}	7
{B}	6
{C}	9
{D}	8
{E}	7

sup

L1

Itemset	Count
{A}	7
{B}	6
{C}	9
{D}	8
{E}	7

圖 6、由 C1 產生 L1

再來會產生了 C2 候選項目集合的結果，根據我們所定義出的支持度 sup=5%，在 C2 中所有的項目集合的次數都必需大於 1，則得到 L2 頻繁項目集合的結果(圖 7)。

C2

Itemset	Count
{A,B}	3
{A,C}	3
{A,D}	0
{A,E}	2
{B,C}	3
{B,D}	2
{B,E}	0
{C,D}	5
{C,E}	3
{D,E}	6

sup

L2

Itemset	Count
{A,B}	3
{A,C}	3
{A,E}	2
{B,C}	3
{B,D}	2
{C,D}	5
{C,E}	3
{D,E}	6

圖 7、由 C2 產生 L2

在本研究中我們利用搜尋二次關聯規則，有別於以往會繼續搜尋到 Lk 次才將最常存取路徑的頻繁項目集合找出，即將 Apriori 演算法簡化到產生 L2 頻繁項目集合為止。再來根據所設定的信賴度 conf = 60%得到強規則(圖 8)即為使用者之最常存取路徑。

L2

Itemset	Count
{A,B}	3
{A,C}	3
{A,E}	2
{B,C}	3
{B,D}	2
{C,D}	5
{C,E}	3
{D,E}	6

conf

R2

Rule	%
CD	62.5
DE	85.7

圖 8、由 L2 產生 R2

由於我將網站結構視為一個無向圖，在計算每一個項目集合是否符合信賴度時必需考慮到兩種狀況。以 L2 中的{C,D}為例可以了解到規則 $C \Rightarrow D$，由於因為無向圖的關係必需也要考慮到 $D \Rightarrow C$ 這個規則，所以在計算信賴度時就要兩種狀況都要考慮。

信賴度計算的公式如下：

$$confidence(X \Rightarrow Y) = P(X \mid Y) = \frac{\sup\_count(X \cup Y)}{\sup\_count(X)} \quad (2)$$

兩種規則計算結果的比較：

$$Max(confidence(X \Rightarrow Y), confidence(Y \Rightarrow X)) \quad (3)$$

所以在 L2 中{C,D}的信賴度結果為 $Max(confidence(C \Rightarrow D), confidence(D \Rightarrow C))$ 則得到 62.5%。

我們將在 R2 中由強規則所得到的項目集合儲存到關係矩陣中，接下來是我們的第二步驟，用深度優先搜尋將使用者之最常存取路徑找出，如圖 9 所示。

	A	B	C	D	E
A	0	0	0	0	0
B	0	0	0	0	0
C	0	0	0	5	0
D	0	0	5	0	6
E	0	0	0	6	0

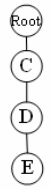

圖 9、R2 之關係矩陣與路徑樹

4. 未來工作及結論

我們利用搜尋二次關聯規則與 DFS 演算法直接將路徑找出，取代以往的單獨使用關聯規則搜尋的方式，期望能提昇網頁使用探勘在搜尋使用者之最常存取路徑的效率，未來工作中我們將進實驗的部份來驗證我們的方法可行。由於在本論文中我們只考慮無向圖的表示法，這無法完整呈現使用者行為，只能將瀏覽頻率高的路徑找出，未來我們也會加入方向性等因素，能夠更精確的透過網頁使用探勘來搜尋出使用者瀏覽行為。

參考文獻

[1] 張峻彬、傅昱翔，"主題知識管理與查詢 – 以溫泉相關資料為例"， ***2006 電子商務與數位生活研討會***，2006。

[2] 曾龍 譯，***資料探礦 概念與技術***，維科圖書有限公司，2003。

[3] R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules", ***Proceedings of the 20th VLDB Conference Santiago***, 1994.

[4] S. Araya, M. Silva, R. Weber, "A methodlogy

for web usage mining and its application to target group identification", ***Fuzzy Sets and Systems 148***, pp. 139-152, 2004.

[5] W. Bin, L. Zhijing, "Web Mining Research", ***ICCIMA'03 IEEE***, 2003.

[6] F. M. Facca, P. Luca Lanzi, "Mining interesting knowledge from weblogs: a survey", ***Data & Knowledge Engineering 53***, pp. 225-241, 2005.

[7] L. Liu, J. Chen, H. Song, "The Research of Web Mining", ***Proceeding s of the 4th WCCICA IEEE***, pp. 2333-2337, 2001.

[8] J. Srivastava, R. Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining:Discovery and Applications of Usage Patterns form Web Data", ***SIGKDD Explorations***, Vol. 1, Issue 2, pp. 12-23, Jan 2000.

[9] Q. Song, M. Shepperd, "Mining web browsing patterns for E-commerce", C***omputer in Industry 57***, pp. 622-630, 2006.

[10] E. Tug, M. Sakiroglu, A. Arslan, "Automatic discovery of the sequential accesses form web log data files via a genetic algorithm", ***Knowledge-Based System 19***, pp. 180-186, 2006.

[11] X. Wang, Y. Ouyang, X. Hu, Y. Zhang, "Discovery of User Frequent Access Patterns on Web Usage Mining", ***The 8th ICCSCWD IEEE***, pp.765-769, 2003.