網路瀏覽行為之探勘分析

劉育津
世新大學資訊管理學系
ycliu@cc.shu.edu.tw

王曾甫
世新大學資訊管理學系
tsengpu@gmail.com

江亦瑄
世新大學廣播電視電影學系
yh.chiang@gmail.com

摘要

網站所提供的服務，便利了生活各項活動所需。傳統的行銷活動主要是以客戶人口統計變項，做為區別目標客群之依據；但由於網路科技和日常生活的緊密結合且網頁瀏覽訊息可被完整的收集，若能進一步分析個人興趣的動態行為，將可成為另一種市場區隔的行銷利器。

據此，本研究使用國內創市際市場研究顧問公司的網頁瀏覽資料庫來對顧客建立群集，並進一步探討不同群集間之行為特徵和意涵。而為了驗証模型的適切性，我們參照並比對美國 Yahoo 研究所提之網路消費者行為分類，實驗結果顯示，本研究所提之模型和美國 Yahoo 所提的分類有高達 85.52%的相似性。

關鍵詞：瀏覽行為、使用者分群、網路行為探勘、網路廣告

網路瀏覽行為之探勘分析

壹、緒論

近年來網際網路技術的快速發展，網路使用者已大幅增加。依據資策會 FIND 的追踪調查顯示，截至 2010 年 12 月底止，台灣商用網際網路帳號總數為 2,514 萬，經常上網人口總數已達 1,079 萬人（資策會創新應用服務研究所，2010）。網際網路所擁有的即時性、互動性與多媒體等特點，也使得網站所提供的功能與服務在不斷的進化、創新。同時網路科技帶動了各種類型網站的蓬勃發展，其不僅對人類的行為和生活方式產生相當大的轉變，也對企業之間的商業競爭造成了本質的改變。

然而，網站對現代人來說更重要且更直接的應用不外乎是扮演著資訊蒐尋的最重要利器，這加快了資訊的傳播速度，使得獲取資訊更為平價和容易，大幅拉近過去因為資訊不平衡所造成的機會不均等，讓有為者可藉由網路所提供的資訊開創對自身有利的契機。是故，網際網路也早已成為現代人除了書籍、報章、雜誌、電視、廣播等傳統媒體外，獲取資訊與知識的重要管道，而根據「一心多用的閱聽人—多工程度與媒體使用」（江亦瑄、孫偉珀，2010）研究，受測者對於網路與電視的使用時間每天平均 2-3 小時間，然而每天平均使用網路 12 個小時以上的重度使用者比例，卻有 8.6%之多，遠遠高於其它媒體。所以網路廣告佔企業廣告預算的比例也正與日俱增。

台北市網際網路廣告暨媒體協會（Taipei Internet Advertising and Media Association，簡稱 IAMA）統計了 2010 年台灣整體的網路廣告市場，規模已達到 85.51 億新台幣；且其並進一步預測 2011 年台灣整體網路廣告的市場規模將達到新台幣 99.67 億元，成長率為 16.55%（台北市網際網路廣告暨媒體協會，2010）。由此可見網路廣告在未來仍有很大的成長空間，同時也顯示網路已成為企業主在從事行銷活動時，效果最佳的管道之一。

在越來越多企業主了解到網路影響力的同時，網路廣告也成為企業的主要運用工具，這也導致企業投資在傳統媒體的預算比例逐漸縮減，轉而將預算投資到網路等其他傳播工具（吳美君，2007）。而當企業欲透過不同特性的網站，作為傳達網路廣告的媒介時，如何精準的掌握目標客戶群，以進一步吸引消費者能主動的瀏覽廣告，並有效率、選擇性的吸引好的新顧客，且同時加深消費者的印象，是熱門且重要的網路營銷議題。

在傳統上，企業針對消費者所進行的廣告活動，在選擇媒體投放時，主要是以其收視率、收聽率、發行量或訂閱率來作為判斷標準之一，以及加上媒體所覆蓋用户的基本資訊（人口統計變項）与企業本身設定的潛在用户基本資訊之一致性。而依據市場區隔（market segmentation）概念，細分消費者市場的基礎可分為：地理、人口統計、心理、行為（Wendell R.Smith 1956），但由於人口統計變項的有關數據相對容易取得，且易於衡量，因此經常以它作為目標用户群的細分依據。也因此目前大部分企業，仍是使用人口統計變項去做用户分類進行網路行銷，絶少有使用網站瀏覽行為，去進行目標使用者定位，並以此提供有針對性的行銷方案。

同時，消費者行為學認為消費者行為是一個持續的過程，包括了許多購買前和購買後的行為和反應。另外基於網路科技的特性，使用者瀏覽行為可以被詳實追踪，因此網路廣告領域中發展了「行為行銷(Behavioral Marketing)」（WIKIPEDIA 2011）的研究，他們主要就是依使用者在網路上的行為特徵，企圖提供適性化的廣告建議。

但網路使用者瀏覽行為的追踪，也廣泛引起了侵犯隱私權的顧慮。2008 年美國國會針對行為行銷的隱私影響舉辦了聽証會（LinuxInsider 2008），起因於矽谷廣告公司 NebuAd 與網路服務提供商（ISP）合作，使用特定硬體作「深度的封包探測」（Deep Packet Inspection），觀測其所有上網的使用者的網路瀏覽行為，並提供廣告商針依此資訊對其用户投放精準的廣告。

因此透過分析使用者的整體網站瀏覽行為，并依此分群使用者，就可以提供企業主在網路上進行品牌經營、市場行銷、廣告投放等等的行銷活動時有所參考，並且規避了隱私風險。

目前針對使用者瀏覽行為所進行的研究，多是運用群集（翁瑞鋒 2001）、分類（方耀白 2000）、關聯法則（紀和村 2010）及序列樣式分析（傅遠榮 2010）等資料探勘技術，而在技術上對於使用者瀏覽行為的記錄可以大致區分為網站伺服器日誌檔、網頁轉換、代理人系統、專用插件等幾種方式。研究者經常藉由資訊分析和資料探勘方式，來分析出使用者在瀏覽行為的特色與模式，但此部分的研究，多集中在特定的單個網站上。

資策會產業情報研究所（MIC）調查了 2011 台灣網路使用者的數位生活型態，發現台灣網路使用者可區分為數位領袖、數位品味、數位 CP、數位觀望四個族群（Market Intelligence & Consulting Institute，2011）。MIC 的研究分析認為瞭解數位族群的消費型態，分析網路族群的採購決策及彼此的關聯性，能夠更有效的經營數位消費族群，且也有助於企業訂定行銷方案。而美國 Yahoo!於 2010 年所做的網路消費者行為分析顯示，使用者在網路上所從事的活動可概括為七類：生活管理、網路購物、資料查詢、新聞取得、維持聯繫、生活娛樂、追求興趣（Yahoo! 2010）。此研究分析則為了掌握消費者的動向，以發揮廣告最大的功效。因此網路使用者的群體分類，對於企業的網路行銷決策有著很大的影響，如果能掌握「網路使用者分群特色脈絡」，將可以助益企業在整體網路市場中，完善整體行銷策略的規劃，進而更有效地提升廣告的效益。

上述此類的研究，多以調查法方式，進行使用者的抽樣調查分析，此將受限於使用者的自我表述情況，和研究者的提問方式有關。然而隨著資訊科技的發展，資料的收集相形容易，使用者的上網瀏覽記錄可被完整的追踪，因此在本篇論文中，我們擬採用創市際市場研究顧問公司所記錄的使用者網路瀏覽行為資料庫（ARO 資料庫）進行分群，企圖使用分群後的集群特色提供企業進行市場區隔與廣告投放之參考依據。期望以此分出有特色的網路使用群集，進而提供企業主在進行網路行銷資源投放規劃時的整體性參考評估。

本文總共分為五章，其架構如下所述，第 2 章是文獻探討，將回顧網站瀏覽行為、使用者分群、網路廣告…等相關研究；第 3 章為研究架構，包含參數上的資訊定義及提出使用者網站瀏覽行為分群模式的解說；第 4 章為實驗設計與預計的結果，以印證本論文所提分群之有用性，最後第 5 章則為結論與未來研究方向。

貳、文獻探討

本研究旨在使用資料探勘技術依網路瀏覽行為對使用者進行分群，以提供企業在網路市場整體行銷策略規劃與廣告投放的參考依據。本章文獻回顧預計分為以下三小節，2.1 節網路廣告的研究；2.2 節是網路瀏覽行為的研究，2.3 節則是結合網路瀏覽行為與網路廣告的相關研究。

2.1 網路廣告

現今台灣民眾接觸的傳播媒體，大部分已經從平面轉變成多媒體與網路了，在 2006 年報紙的閱讀率從 1991 年的 76.3%降低到 39.1%，而電視卻從 85.9%提升到 94.5%，而網路接觸率也高達 52.3%，成為台灣的第二大媒體（徐榮華 2007）。

對網際網路使用者來說，不同的網路媒體在資訊提供上扮演著不同的角色，使用者也能依自身的特定需求來瀏覽適切資訊；是故，這也同時也意謂著企業廣告必須有策略地依此安排針對性的廣告曝光及內容，才能獲得最好的投資報酬。網路廣告吸引企業主的地方，除了表現型式多元化以及豐富的動態多媒體效果以外，還在於廣告能與使用者產生互動性，更可以即時精確取得使用者行為的量化數據，進行深度分析與評估。

Novak & Hoffman（1996）提出三個不同型態的傳播模式，其中超媒體電腦網路媒體模式（Hypermedia Computer Media Environment）即是網際網路媒體的概念，為一種動態散佈模式，其主要特色為人際互動與機器互動，使用者不只是單向被動的接受信息，也會透過網際網路進行信息的搜集與散佈。Novak & Hoffman（1997）將網路廣告的評估方式分為曝光性與互動性兩種，曝光數越高，接觸使用者的範圍越廣，而互動性則指紀錄瀏覽者與網頁互動關係的過程，包含了點擊率、瀏覽時間、瀏覽頁數等等。

蔡佩珊（2004）的研究則認為網路廣告效果評估可分為三個層次：曝光性、互動性、使用者行為。而使用者行為分析的優點在於可以彌補其他指標的不足之處，能完整了解網友對網路廣告的態度；然而其缺點在於它是長期而相對的趨勢觀察，須投入較多的成本始能分析出網友行為背後的動機因素。

由上述研究可得知，以網際網路為廣告媒介，最大的優勢就在於與使用者的互動程度，透過網路互動的特性，使用者可以主動選擇想要看的內容，因此了解使用者瀏覽網站的行為，並據此分析出其背後所隱含的意義，能使網路廣告主或代理商有所依據作出較精準的廣告策略，接觸到最大量的目標對象，取得更好的廣告效果。

2.2 網路瀏覽行為

傳統媒體在研究媒介的使用習慣時，有明顯的實施難度，所以較少有依據媒體本身被瀏覽的行為來進行分群分析的研究。其中郭久蓁（2005）針對各媒體使用行為的研究，探討了報紙、雜誌、廣播、電視與網路的重度人口統計輪廓，以及五大媒體之間的關聯性矩陣，再使用類神經網路與決策樹建構區辨媒體使用行為的模型。而網路媒體的優勢在於可以蒐集精確的瀏覽資料，因為使用者瀏覽網頁的動作皆可以完整的被紀錄下來，

而不像電視、收音機等傳統媒體，只能單向的傳輸資料。因此透過網路媒體，蒐集完整的使用者瀏覽資訊，更有利於分析瀏覽者的實際行為模式。

目前在線電子商務企業經常應用購物車分析（Marketing Basket Analysis; MBA）進行個人化的交叉銷售，此法即利用使用者的過去行為模式，找出使用者瀏覽物品間的關聯性，再採用關聯法則（Association Rule）推導出個性化的建議標的物。

而 Huang, Shen, Chiang, & Lin（2007）的研究則發現，會瀏覽多種類的網站的使用者，相對的會瀏覽同一種類下比較多的網站，也同樣的會在同一個網站下瀏覽較多的頁面，但喜歡使用搜尋引擎的使用者，則不會去瀏覽很多種類的網站，在同一種類下，也不會瀏覽很多網站，而在同一個網站下，也不會瀏覽很多頁面。因此隨著網路使用者的增加，以及網站類型多樣化的發展，網路使用經驗與能力也會影響其網路行為。

資策會產業情報研究所（MIC）的研究分析也認為台灣使用者的數位生活型態相當多元，而透過瞭解數位族群的消費型態，分析網路族群的採購決策及彼此的關聯性，能夠更有效的經營數位消費族群，且也有助於企業訂定行銷方案。美國 Yahoo!於 2010 年所做的網路消費者行為分析，則將使用者在網路上所從事的活動分為七個種類，而又將這七種活動種類歸納為三個目標：達成目的、隨性瀏覽、追求興趣。其研究目的在使廣告商有效的掌握消費者的動向，並發揮廣告最大的功效。

Neale Martin（2009）則認為，人類 95%的行為是由潛意識的習慣控制的，此一研究認為應通過研究顧客行為而不是觀點和意向，去達成企業的營銷目標。而使用者在持續的網頁瀏覽過程中，隱藏著潛在的行為模式，可能代表著使用者的特殊喜好或興趣，這些資訊即提供了解使用者的最好管道。

由上述研究可以得知，使用者的網路瀏覽行為可以提供豐富的訊息供研究者做行銷研究之用。

2.3 網路瀏覽行為與網路廣告的整體研究

由於不同上網目的的網路瀏覽者會有不同的活動與習慣，而網站品牌也會帶給使用者有不同的感覺，其網站功能或目標客群也各有差異。因此網路廣告形式若依據瀏覽者的需求，來顯示特定的廣告，也可以更有效的貼近瀏覽者的目的，並且提高點擊廣告的可能性。

美國 Yahoo!於 2010 年所做的網路消費者行為分析也認為：對於屬於達成目的和追求興趣型的目標，使用者比較不願意或者會忽略廣告的影響，因此廣告效果最有限，但是對於隨性瀏覽型的使用者而言，對於廣告接受程度較大，而且對於廣告的印象也最深，是最容易受到廣告影響的階段。這也使得廣告商可以針對不同類型的使用者，使用不同種類的訊息表現。

Xin Ge, Gerald Häubl, and Terry Elrod（2011）的研究提出了 Delayed Favorable Information 的概念，研究認為消費者選購商品的行為，會經歷兩階段步驟：消費者會先刪除不理想的商品，將注意力放在最後幾個理想商品，最後在從其中選出一個要買的商品。因此若在第二階段時，加强消費者接觸到產品信息，則消費者會調整自己原本設定的重要順序，接受資訊的影響而產生最後的購買行為。

因此針對網路使用者瀏覽行為資訊的收集與監測，進行關聯分析，就能夠知道其偏好以及洞悉使用者的資訊使用模式，再將之提供給企業用以投放精準的網路個性化廣告，就成為很多網路企業發展的方向，但若強勢搜集個人的瀏覽行為，往往也會引發侵犯使用者個人隱私的莫大風險，於此，世界各國政府也積極推動相關的隱私權保護法案。

基於真實的網路使用者行為資料不易取得，所以據此進行使用者分群的相關研究相當少，大部分都僅針對特定網站或模擬實驗環境下，研究使用者的網路行為及認知態度。而從使用者上網的瀏覽行為來觀察，如能依據訪問的網站類別行為進行使用者分群，定義出特色的用户群，就能提供廣告主在投放網路廣告時，一個大面向的網站組合，且能確實覆蓋到各種不同購買決策階段的目標用户。如此一來，不但能保證廣告覆蓋人數，亦能精準的定位目標客群，以提高目標客群對該廣告留下深刻的印象，業者可在成本、影響深度、速度等獲得最大的廣告投放結果。此外，亦能藉由洞察用户網站瀏覽行為模式的分群，幫助網路媒體經營者定位網站價值，進而提高其廣告收入及展示效果。

緣此，有別於過去的文獻，本研究欲從真實網路使用者行為資料中，從使用者實際已發生的瀏覽行為，依其差異探尋使用者分群的特色，進而找出網路使用者瀏覽偏好相似的群集，以提供企業主廣告投放以及後續研究者進行網路行為研究時有更完整的參考。

參、研究方法

本研究以使用者實際已發生過的瀏覽行為為基礎，從真實網路使用者的整體網站之間的瀏覽動作中，探討使用者分群的特徵，進而找出網路使用者瀏覽偏好相似的群集。本章將詳細說明所提之分群模型，內文節次安排如下：3.1 節將先針對本研究所提之研究架構做說明；3.2 節將說明研究之流程，并依研究流程詳加說明：3.3 節對本研究資料來源及清整條件說明；3.4 節對網路活動種類進行說明；而 3.5 節對網路使用行為及網站瀏覽率分群模型說明-即參數的資訊定義及本文所提網站瀏覽率的量測模型解說。

3.1 研究架構

本研究將焦點放在使用者網際網路間的整體瀏覽活動，針對其網際網路瀏覽活動的記錄資料，分別計算出每位使用者的網路使用行為及網站瀏覽率資訊，並進一步將這些變項，使用資料探勘中的資料分群（Data Clustering）做為分析算法，進行使用者瀏覽行為分群，再提取及對照使用者群體特徵，研究架構如下圖 1 所示。

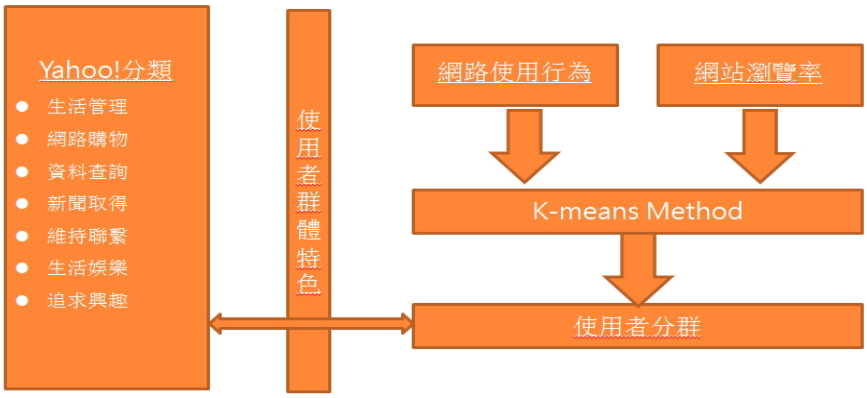

圖 1 研究架構

研究期望使用 k-means 方法進行使用者分群後，所提取出的使用者群體特徵，和美國 Yahool！所採用調查法研究分析之分類特徵做對照。據此，驗証網路使用行為及網站瀏覽率兩者的綜合變項，分辨出有參考價值的使用者群體，其中網路使用行為變項包含有寬度、長度、深度，網站瀏覽率則為排名前十大使用者網站接觸重要度的網站。

3.2 研究流程

本研究所使用的資料來源為創市際市場研究顧問公司所記錄的使用者網路瀏覽行為資料庫（ARO 資料庫），其中包含了使用者在網際網路上的各種瀏覽行為記錄，因此須依研究的目的對來源資料進行清整，篩選出具有研究對象價值的瀏覽行為歷程記錄，以便後續模型建立及變項計算，計算使用者的各項網路資訊行為衡量指標，整個研究的流程如圖 2 所示，並於後續小節依此研究流程與步驟詳加說明。

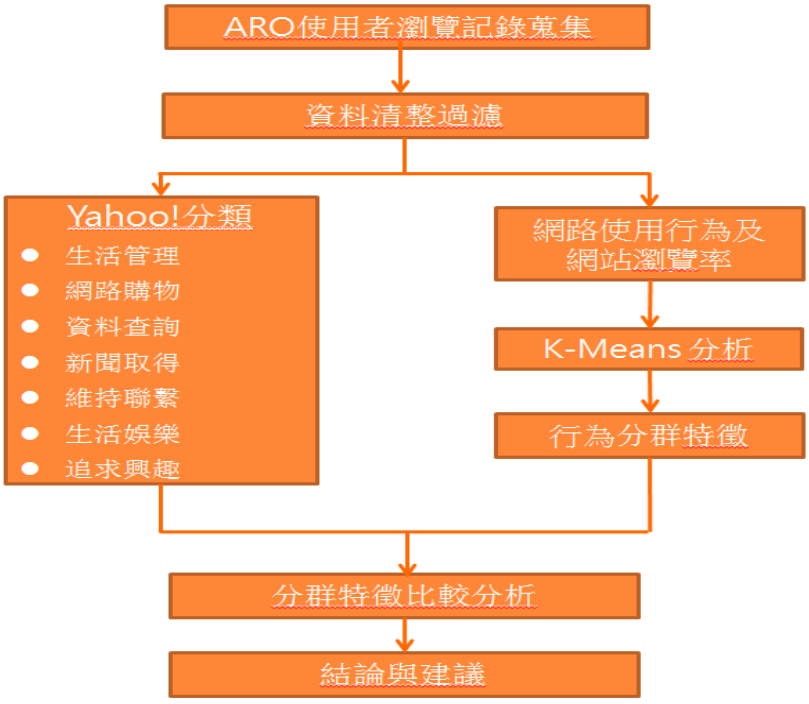

圖 2 研究流程圖

3.3 研究資料及清整條件

創市際市場研究顧問公司是目前台灣網路市場測量產業的領導企業之一，其也是台灣首家針對網際網路使用行為與網際網路產業相關研究的研究機構，其主要客戶為電子商務經營者、網站營運者、網路廣告代理商、網路行銷業者等，同時也是觀注網際網路趨勢分析研究的學者的重要參考來源資料之一。

Access Rating Online（ARO）是一種「網路收視率調查」的研究工具（http://www.insightxplorer.com/product/ARO01.html），其蒐集資料的方式類似過去電視收視率調查使用黑盒子裝置方式，使用該公司的樣本點選流向監控記錄軟體 NetRover™，可以在幾乎不影響會員的狀態下，精確的記錄使用者瀏覽網頁流向的歷程，並即時回傳至創市際伺服器再加以記錄至資料庫中。紀錄資料經過前置轉換與處理之後，即可分析出國內相關網站的使用現況。其記錄的資料包含有使用者 ID、連線 ID、瀏覽器種類、造訪的網域、子網域、目錄、停留時間等等完整的瀏覽資訊，而本研究所使用的資料欄位如表 1 所示。

表 1 研究使用的 ARO 資料格式

使用者 ID	連線 ID	日期時間	造訪的網域	停留的時間
CustomerID	SessionID	UserDateTime	DomainName	TimeSpend

在過去即有運用創市際 ARO 網路資料庫所提供之資料進行網際網路行為方面的研究，如：探討橫幅廣告點擊效果的影響因素（俞帛宏 2008），探索網站縱剖面的瀏覽行為（黃莉雯 2008），探索網路廣告點擊率與點擊廣告後行為（江岱衛 2007）等等使用者行為方面相關議題的研究。

依據從 ARO 資料庫所擷取出的 2011 年 4 月份的使用者瀏覽記錄之資料內容特性，可藉由定義資料取用的限制條件，過濾出有效的資料。在本研究中，定義了以下的限制條件，說明如下。

一、使用者限制：

本研究為了提高使用者瀏覽行為的真實度，經交叉比對後刪除未有完整基本資料的使用者瀏覽記錄，而此部分的使用者其瀏覽筆數約僅佔整體筆數的 0.09%。

二、瀏覽次數限制：

瀏覽次數指在同一個 session 之中的瀏覽行為，而 Session 指同一瀏覽時間內，使用者瀏覽網頁的區隔時間之內。本研究定義了一個最低的瀏覽次數門檻值，以過濾出瀏覽次數高於門檻值的使用者瀏覽資料，其中若使用者的瀏覽次數少於等於 5 次的將略過不計。

三、訪問目標限制：

受測使用者訪問的目標網站，以有域名（Domain Name）監測的資料為主，因為域名代表著一個網站在網路上的身份，且其也做為使用者識別網站地址的基礎條件，因此沒有域名型式資料的瀏覽記錄，將予以刪除。

3.4 網路活動種類

本研究預計從使用者的網路瀏覽行為中，進行使用者分群探討，故擬對照美國 Yahoo！於 2010 年針對網路消費者行為分析的研究結果。

據此，本研究預計將使用者行為分為 7 個群，群特徵則採用美國 Yahoo！研究得出的七個使用者網路活動種類為依據。因此將 ARO 值前三十名的網站，先使用 TF-IDF 計算出瀏覽行為具有使用者獨特性特徵的前十名網站，再根據其具有的活動種類特徵加以權重值歸類（如表 2 所示），從而計算使用者瀏覽行為中包含此部分網站的比率，並依值權重分類使用者所屬的種類。

表 2 網站所屬活動種類範例

	Manage	Shop	Research	Inform	Connect	Entertain	Passion	total
facebook.com	0	0	0	10	70	20	0	100
yahoo.com.tw	10	30	10	20	10	10	10	100
wretch.cc	0	0	0	0	80	20	0	100

3.5 網路使用行為及網站瀏覽率分群模型

本小節預計依研究架構分成兩個小節進行說明，第一節說明網路使用行為的測量計算，第二節則加入網站瀏覽率的計算說明，最後則將此兩二節得出的變項，使用 K-means 進行分群分析研究。

3.5.1 網路使用行為

Huang, Shen, Chiang, and Lin（2007）的研究定義了網路資訊行為的衡量方式，共包含六項衡量指標：（1）寬度（width）指使用者造訪的網站類型數量；（2）長度（length）為使用者在某網站類別中所造訪的平均網站個數；（3）深度（depth）係使用者在一個網站中所瀏覽的總頁數；（4）網站停留時間（duration）意指使用者於一個網站中平均停留了多少時間；（5）搜尋引擎使用比率（propensity）是搜尋引擎之瀏覽總頁數佔總瀏覽頁數的比率；（6）造訪網站關聯性（relatedness）乃使用者所造訪的網站彼此間的關聯程度。而藉由寬度、長度以及深度三項衡量指標，可描繪出足以代表各別使用者資訊行為的空間示意圖。

1.寬度：即使用者造訪網站的類別數總和，此處的網站類別定義為 ARO 資料庫對每一個網站所訂定之類別，共分為十大類，分別是：成人、文化藝術、通信、電腦、經濟、網際網路、知識、休閒、媒體、社會等十大類，而總計有 69 種網站類別。

$$寬度 = \sum 使用者造訪網站的類別數 \tag{1}$$

2.長度：使用者於每種不同類別的網站中，平均造訪的網站數量，而網站則以子網域來做區分，主要原因在於此能夠更細緻地呈現網路行為的真實情況。

$$長度 = \frac{使用者造訪的網站數量}{使用者造訪網站的類別數} \tag{2}$$

3.深度：使用者造訪每個網站之平均瀏覽的總頁數。

$$深度 = \frac{所有網站的瀏覽頁數總和}{造訪網站的總數} \tag{3}$$

3.5.2 網站瀏覽率

此部分擬採用資訊檢索常見的加權技術 TF-IDF（Salton, G. and C. Buckley, 1988），來統計評估某一網站在使用者瀏覽行為過程中的重要性。我們將使用者的整體瀏覽行為記錄，視為一份文件，而使用者所瀏覽的網站域名，則視為某一字詞。

TF-IDF（Term Frequency-Inverse Document Frequency）用以評估一字詞對於一個文件集中的其中一份文件的重要程度。字詞的重要性隨著它在文件中出現的次數成正比增加，但同時會隨著它在語料庫中出現的頻率成反比下降。

詞頻（term frequency，TF）指的是某一個給定的詞語在該文件中出現的次數。對於在某一特定文件裡的詞語 ti 來說，它的重要性可表示為：

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{4}$$

其中 ni,j 是該詞在文件 dj 中的出現次數，而分母則是在文件 dj 中所有字詞的出現次數之和。

逆向文件頻率（inverse document frequency，IDF）是一個詞語普遍重要性的度量。某一特定詞語的 IDF，可以由總文件數目除以包含該詞語之文件的數目，再將得到的商取對數得到：

$$\mathrm{idf_i} = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{5}$$

其中|D|為語料庫中的文件總數，$|\{j : t_i \in d_j\}|$則包含詞語 ti 的文件數目。

$$\mathrm{tfidf_{i,j}} = \mathrm{tf_{i,j}} \times \mathrm{idf_i} \tag{6}$$

某一特定文件內的高詞語頻率，以及該詞語在整個文件集合中的低文件頻率，可以產生出高權重的 TF-IDF。

ARO 值前二名的網站為 Facebook 及 yahoo，而其 ARO 值各為第三名網站的 5 倍、4 倍左右。因此使用 TF-IDF 加權技術，可以過濾掉使用者大多會經常訪問的網站，而保留具有重要意義的網站，從而得到瀏覽網站行為的獨特性特徵。

肆、實驗設計與結果

本章將應用網路使用行為及網站瀏覽率的變項數據進行實驗，相關的實驗步驟與結果將在如下節次說明：4.1 節資料分析，對於使用者網路瀏覽資料解析處理過程予以說明；4.2 節則依實驗流程與步驟加以說明。

4.1 資料分析

本研究資料為從 ARO 網路資料庫中擷取出的使用者瀏覽區段記錄，期間為 2011 年 4 月 1 日至 2011 年 4 月 30 日的網站瀏覽記錄，并以此過濾出有效的資料來進行研究分析。而我們將研究目標對象定義為 4 月份的瀏覽次數大於 5 次的使用者，並且其瀏覽目標為具有域名的網站記錄。

經清整後，整體的瀏覽行為記錄次數為 162596 次（session），而合計的總瀏覽網頁次數為 7367925 次，整體使用者人數則共為 1637 人。

為將資料進一步收斂以利計算：網站瀏覽率變項，故將 4 月份所有瀏覽網站取 ARO 排名前三十名為計算目標，計算此三十名網站的 TF-IDF 值，並取最高的前十名網站，做為此變項的來源。

而另以 ARO 所建立的網站分類，並依照 3.5.1 小節所述，分別標識出每個網站所屬的類別，并統計出每個使用者的網站瀏覽個數、網頁瀏覽個數、網站瀏覽類別數，再分別計算出網路使用行為的寬度、長度、深度。

4.2 實驗流程與步驟

本論文針對網路使用行為及網站瀏覽率兩個變項進行計算，并使用 K-means 進行分群，再將分群結果進行分析、對比。

實驗首先邀請了二位長期在網路產業服務的行業專家，請專家對於本研究所提出的 ARO 網站分類對應於 YAHOO！的 7 大分類的對應權重數據進行複核，再依據數據來計算研究對象的所屬歸類，以進一步驗証 K-means 使用各不同變項所分群出來的效果。

實驗使用了網路使用行為及網站瀏覽率兩個變項的不同組合，進行 K-means 的分群計算，最後得出了依據前面第三章研究方法所提的方式，以 K-means 分群出來的結果，有 85%左右能正確被歸類到 YAHOO！的 7 大分類。

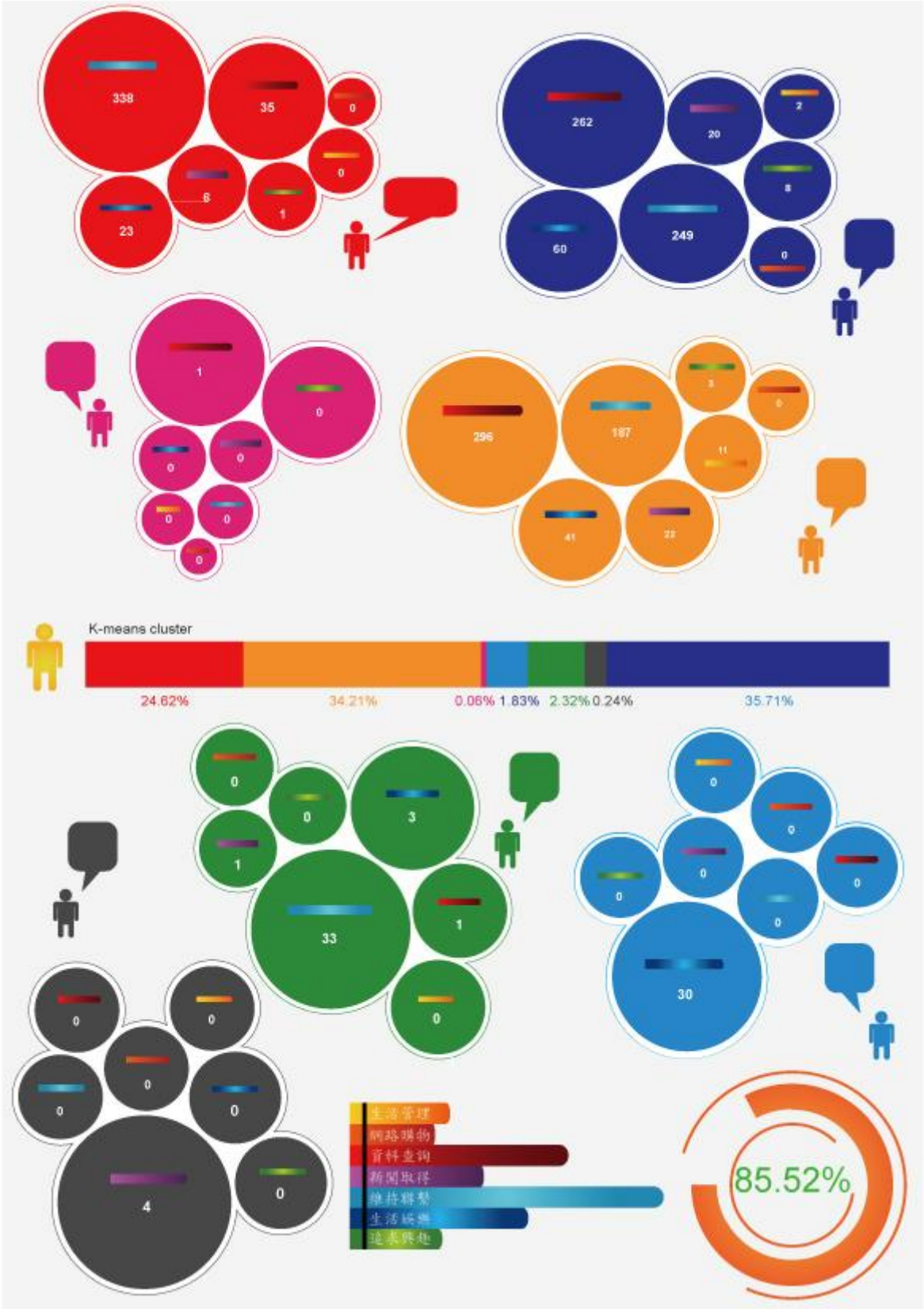

圖 3 分群對照率資訊圖表

伍、結論與未來研究方向

使用者群體分類一直是商務從業者急欲探知的範疇，因為若能依據使用者不同的群體提供不同的服務或分析其行為模型，可以極大化其商業目的。而使用者群體分類一直以來多半都是使用人口統計變項來進行分析主軸，而依據近幾年的研究也能發現，依據行為特徵的分類，更能清晰找出商業目標用户。

本研究旨在利用實際網路瀏覽行為數據，使用 K-means 的分群技術，來找出使用者群體分類的可行方式。而實驗結果亦佐證，搭配美國 Yahoo 的七大分類，本研究所提的研究方法在使用者群體分類上，能有效地追踪網路使用者之群集脈絡，可大幅降低調查法的受訪者主觀性表述問題。

儘管如此，本研究仍未臻完美，後續研究可對分類對照的權重值採用更加客觀的研究方法進行查訪，以期實驗的驗証步驟更為嚴謹。另外，有關網站的分類架構，亦可考慮使用其它的網站類別架構進行比照，以更廣泛地檢驗本研究之適用性。

參考文獻

1. MIC，台灣網友分四類「酋長、自魅、喜比、慢熱」，http://mic.iii.org.tw/aisp/pressroom/press01_pop.asp?sno=267&type1=2，取得日期：2011/06/23。
2. IAMA，2010 網路廣告量統計報告，http://www.iama.org.tw/Resource/Trend，取得日期：2011/05/15。
3. 方耀白，2000，應用 SOM 和資料探勘模型建置智慧型網頁探勘系統，大同大學資訊工程研究所碩士論文。
4. 江亦瑄、孫偉珀，2010，一心多用的閱聽人——多工程度與媒體使用，創市際市場研究顧問公司報告。
5. 吳美君，2007，台灣地區廣告代理商經營策略之研究-以台灣奧美廣告為例，臺灣大學商學研究所碩士論文。
6. 吳書婷，2011，網路廣告之訴求與互動性對廣告效果的影響，世新大學公共關係暨廣告學研究所碩士論文。
7. 紀和村，2010，基於使用者瀏覽行為之適性化推薦系統，臺中技術學院資訊工程系碩士論文。
8. 徐榮華，2007，台灣報業經營困境與因應策略，國立政治大學傳播學院碩士論文。
9. 翁瑞鋒，2001，網頁瀏覽者行為之泛化分群分析，國立交通大學資訊科學系碩士論文。
10. 許順富，2000，網路廣告特性類型與廣告效果之探討，國立臺灣大學商學研究所碩士論文。

11. 黃聿清、莊春發，台灣網際網路市場多樣化之初探--以網站類型為例，http://ccs.nccu.edu.tw/history_paper_content.php?P_ID=1257&P_YEAR=2010，取得日期：2011/07/16。
12. 傅遠榮，2010，資料探勘技術在分析網站使用者瀏覽模式之研究，大同大學資訊工程研究所碩士論文。
13. 郭久蓁，2005，應用資料探勘技術於媒體使用行為之研究－以 2004 世新傳播資料庫為例，世新大學傳播管理學系碩士論文。
14. 蔡佩珊，2004，網路廣告效果評估方式之探討，國立政治大學廣播電視研究所碩士論文。
15. Dominique R. Shelton, Will Congress Squelch Behavioral Marketing?, http://www.linuxinsider.com/story/security/64301.html, Retrieval：2011/09/21.
16. Huang, Shen, Chiang, & Lin (2007) , "Characterizing Web User's Online Information Behavior" , Journal of the American Society for Information Science and Technology, Vol. 58, NO. 13, PP. 1988-1997.
17. IAB,IAB Internet Advertising Revenue Report Conducted by PricewaterhouseCoopers(PWC) ,http://www.iab.net/insights_research/1357, Retrieval：2011/10/14.
18. Jansen, Bernard J. and Marc Resnick(2006) , " An Examination of Searcher's Perceptions of Nonsponsored and Sponsored Links During Ecommerce Web Searching" , Journal of the American Society for Information Science and Technology, Vol. 57, No. 14, PP. 1949-1961.
19. Martin, Neale（2009）, " Habit: The 95% of Behavior Marketers Ignore" , USA, Pearson PTR.
20. Novak, Thomas P. and Donna L. Hoffman(1996) , " Marketing in Hypermedia Computer-Mediated Environments: Conceptual Model" , Journal of Marketing, Vol.60, No.3, PP. 50-68.
21. Novak, Thomas P. and Donna L. Hoffman(1997) , "New Metrics for New Media: Toward the Development of Web Measurement Standards" , World Wide Web Journal, Vol. 2, No. 1, PP. 213-246.
22. Wikipedia, Behavioral targeting, http://en.wikipedia.org/wiki/Behavioral_targeting, Retrieval：2011/09/21.
23. Xin Ge, Gerald Häubl, and Terry Elrod, Is it best to withhold favorable information about products?, http://www.eurekalert.org/pub_releases/2011-10/uocp-iib102111.php, Retrieval：2011/10/26.
24. Yahoo!, Advertising by Mindset: Optimizing Ad Receptivity Based on Activity, http://advertising.yahoo.com/article/advertising-by-mindset-optimizing-ad-receptivity-based-activity.html, Retrieval：2011/06/23.

Analyzing Customer Clusters Based on Web-Browsing Logs

Yu-Chin Liu
Department of Information Management, Shih Hsin University
ycliu@cc.shu.edu.tw

Tseng-Pu Wang
Department of Information Management, Shih Hsin University
tsengpu@gmail.com

YI,I-HSUAN CHIANG
Department of Radio, TV, & Film, Shih Hsin University
yh.chiang@gmail.com

Abstract

As the advent of web applications prevail, user browsing activities can be collected effectively. Traditionally, corporate marketing strategies are decided according to customers' demographic attributes; however since customer behavior become reachable, clustering customers with web browsing logs would facilitate corporate target right customers more precisely.

Therefore, this paper proposes a model to cluster customers with their web-browsing activities. In order to verify the model, a real database collected from InsightXplorer is used to cluster customers, further the clustering results are compared to the taxonomy announced by US Yahoo!. The experimental results show our proposed method agree with YAHOO taxonomy in 85.52% degree.

Keywords: Browsing behavior, User clustering, Network behavior mining, Internet advertising