以改良式網頁搜尋技術結合個人化排序之研究

梁錫卿
朝陽科技大學
資訊管理系
scliang@cyut.edu.tw

羅卓雄
朝陽科技大學
資訊管理系
chlaw@cyut.edu.tw

廖敏妃
朝陽科技大學
資訊管理系
s9514609@cyut.edu.tw

黃釋緯
朝陽科技大學
資訊管理系
s9514605@cyut.edu.tw

摘要

傳統的搜尋過程，通常是依照使用者所輸入的關鍵字來進行搜尋並傳回結果，其結果往往無法滿足使用者的需求。因此，本研究提出一個架構，藉由改善搜尋引擎的運作流程，來滿足使用者的需求，同時能縮短整體的搜尋時間。所提出之方法將利用網頁探勘技術及資料探勘之關聯規則進行分析關鍵字與網頁之間的關係，並且加入個人化機制來記錄使用者的瀏覽行為，最後再針對不同群組的使用者進行分類及分群，以獲得更好的搜尋結果。本研究未來將完成二個部分，第一，整合目前有關於個人化排序與搜尋引擎的發展與問題，並提出改善，且發展出更符合使用者需求的搜尋引擎運作流程；第二，架設搜尋引擎網站及實作個人化機制，來驗證本研究的結果。

關鍵詞：資料探勘、網頁探勘、關聯規則、搜尋引擎、個人化排序

1. 前言

由於網際網路的盛行，人們依賴搜尋引擎來獲得資訊的程度增高。然而，隨著網頁資料的快速增加使得搜尋的複雜度也隨之提高，因此，一個有效的搜尋機制是有其必要性。

網頁探勘是目前在網際網路中擷取資料的重要技術之一，此技術可以讓使用者更快且更有效率的獲得所需的資訊，並可以降低時間的成本。然而，在傳統以關鍵字進行搜尋的結果，往往與使用者的主觀預期有所出入。因此，本研究希望藉由改善搜尋的結果來提高使用者的滿意度，並縮短使用者查找搜尋結果的時間。

本研究提出個人化機制，以改善上述二點。個人化機制的主要功能係針對關鍵字和網頁進行網頁探勘分析的動作，用以尋找及記錄個人化的資訊，在每次使用者查尋相同的關鍵字時，可針對搜尋結果優先列出該使用者曾經點閱過的網頁，藉以減少使用者在搜尋結果中進行過濾所需耗費的時間。當記錄的使用者搜尋資訊數量夠充足時，便可進行使用者分群，並且可針對類似網頁進行分類。當同一群組的使用者進行搜尋時，依據所輸入的關鍵字可提供同一群組內具相關性的使用者曾經使用過的網頁資訊，以輔助個人化的排序結果。進一步可以利用不同群組之間的互相影響讓使用者能擁有更豐富的搜尋結果。

本研究的架構分為四大部分，第一部分為文獻探討，文獻探討的內容會將有關資料探勘、網頁探勘、關聯規則、搜尋引擎以及個人化排序的相關資料做整理介紹；第二部分則是闡述本研究的研究架構；第三部分會針對研究步驟逐一的說明；最後，則是本研究預期的研究成果。

2. 文獻探討

2.1 資料探勘(Data Mining)

資料探勘是指從大量的資料中存取或探勘知識[8]，透過資料的搜尋和分析找出有用的資訊為資料探勘的主要目的，一般而言，資料探勘的技術可分為傳統與改良技術二大項。傳統技術以統計分析為代表，例如機率論、迴歸分析、敘述統計等，而在改良技術方面較普遍的方法有決策樹理論、類神經網路、規則歸納法等，其中類神經網路為非線性的設計，設計的概念是模擬人腦思考的資料分析模式。資料探勘主要可以建立以下六種模式[1]：

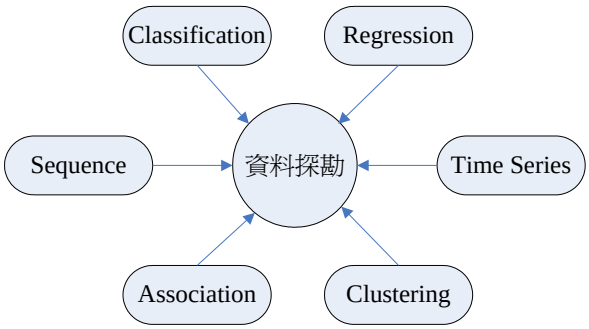

圖 1 資料探勘的模式

(1) Classification

主要是根據一些變數的數值來做為計算的準則，再依照結果進行分類的動作，Classification 常常被用來處理一些已經分類的資料來研究它的特徵，然後再根據這些特徵對其他還沒有分類過或新的資料進行預測分類。

(2) Regression

Regression 是使用一系列現在的數值來做預測一個連續數值的可能值，目前運用的範圍特別是在類神經網路或決策樹理論等分析工具。

(3) Time Series

Time Series 與上述的 Regression 功能類似，都是利用現有的數值來預測未來的數值，其差異在於 Time Series 所分析的數值都會跟時間相關，換句話說 Time Series 的工具是可以處理有關時間的一些特性，例如季節性、時間週期性等。

(4) Clustering

Clustering 主要的目的在於將資料分群，且可以將群組間的差異分辨出來，同時也可以把同一群組的相似性找出來。Clustering 與 Classification 不同的地方在於，Clustering 在分析資料之前並不知道會以何種方式或根據什麼來進行分類的動作，而 Classification 卻是必須事先知道分類的項目。

(5) Association

Association 的概念是藉由某一事件的發生或是在資料庫中同時會出現的東西。例如：一個顧客買了奶粉後，這個顧客同時也會購買尿布的機率是 80%，也就是當 A 事件發生時，同時 B 事件也一起發生的機率[2]。

(6) Sequence

Sequence 與 Association 的觀念很類似，但是 Sequence 主要是針對時間因素來做為區隔，例如：若 A 股票在某一天上漲了 10%，而且當天的股市加權指數下降，則 B 股票在兩天之內上漲的機率是 70%。

2.2 網頁探勘(Web Mining)

網頁探勘為近年來受到各界關注的議題，利用探勘的原理應用在網頁上，藉由探勘的技術發掘有用資訊並進行分析。網頁探勘的技術一般可分為三種類型[5][9][16]：

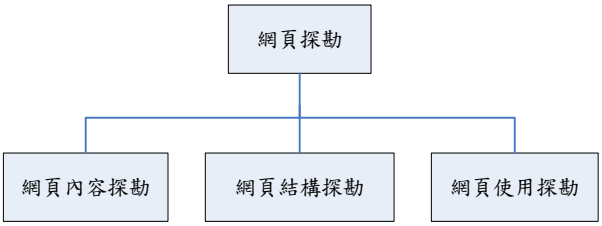

圖 2 網頁探勘的技術

(1) 網頁內容探勘(Web Content Mining)

網頁內容探勘主要是由網頁內容進行探勘的動作，網頁內容探勘是分析網頁內容進而達到資訊擷取、資訊過濾、資訊摘要、及網頁資訊分類等能力。網頁資訊過濾主要是從搜尋引擎所搜尋出來的結果進行過濾，可過濾出個人化的資訊[13]。

(2) 網頁結構探勘(Web Structure Mining)

網頁結構探勘主要是由網頁的結構來進行探勘，利用網頁的結構來改善或突破傳統網頁搜尋引擎的效能。在現今的搜尋引擎中頗具盛名的 Google 即是採用 PageRank 技術，利用全新的設計針對不同的網頁進行排名、分析網頁的重要性，以達到效能提升的目的[10]。

(3) 網頁使用探勘(Web Usage Mining)

網頁使用探勘主要是經由使用者在網路活動的記錄及使用者個人資料這二方面來進行探勘的動作[14][15]。網站的記錄檔(Web Log File) 會記錄使用者參與網路活動中的過程，可經由網頁使用勘的技術得知使用者

瀏覽網頁的順序及行為，便可進一步將有相似行為的使用者叢聚(Cluster)起來。並且可針對不同的群組給予適當的搜尋結果。

2.3 關聯規則(Association Rule)

在資料探勘裡最常用到的技術之一，就是關聯規則，在上面我們已經大略說明過關聯規則，接下來我們將更詳細的說明關聯規則演算法的運作模式。而關聯法則的演算法主要是考慮三個因素，如以下各點[2][3]：

(1) 支持度(Support)

支持度是表示，當 A 產品與 B 產品同時被購買的情況發生時，占全部的交易的機率，其公式如下：

$$P(A \cap B) = \frac{\text{同時包含 A 商品和 B 商品的交易數}}{\text{總交易數}}$$

(2) 可信度(Confidence)

可信度又稱為信賴度主要是用來表示，當 A 產品已被購買的情況下，B 產品也被購買的機率，其公式如下：

$$P(B \mid A) = \frac{\text{同時包含 A 商品和 B 商品的交易}}{\text{包含 B 商品的交易數}}$$

(3) 相關分析(Lift or Improvement)

相關分析則用來衡量兩產品間的相關程度，其公式如下：

$$Corr_{A,B} = \frac{P(A \cap B)}{P(A) \times P(B)} = \frac{P(B \mid A)}{P(B)}$$

2.4 k-平均叢集演算法(k-means Clustering Algorithm)

K 平均叢集演算法是最簡單且普遍應用在平方差的一種演算法，下面說明此方法執行的步驟[6]：

(1) 假設有一組特徵向量以座標 x_1 及 x_2 來表示，如圖 3：

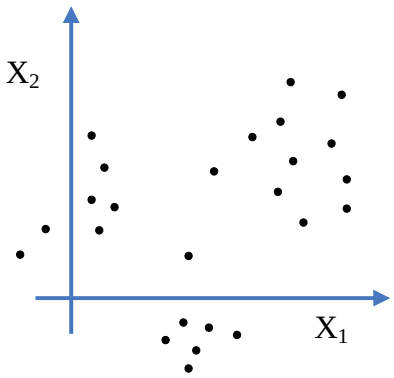

圖 3

(2) 任意選擇 k 個「種子」悉按亂數設定作群集重心，假定 k = 3，如圖 4：

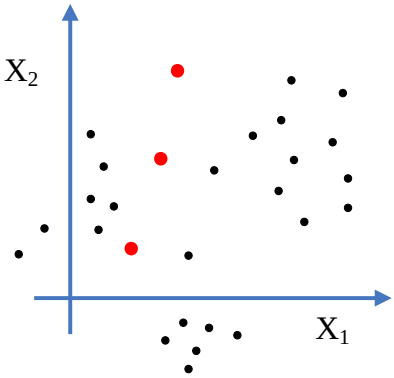

圖 4

(3) 將每一資料點分配到重心最接近的群集中，計算每一個群集的重心，如圖 5，左圖為計算重心前，右圖為計算重心後。

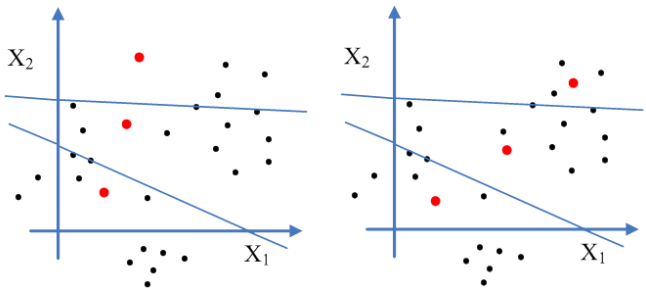

圖 5

(4) 將群集中每一個點的位置加以平均，找出新群集，每一點再次被分配到重心最接近的群集中，如圖 6，左圖為計算重心前，右圖為計算重心後。

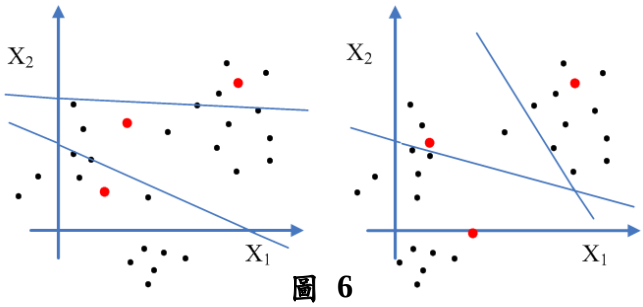

圖 6

(5) 重複進行直到群集邊界不再變動為止，右圖為最後結果，如圖 7：

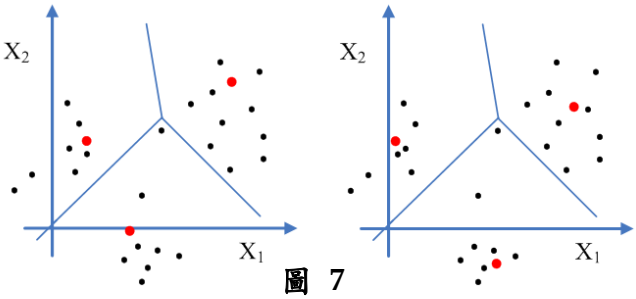

圖 7

2.5 K- 最 臨 近 分 類 (K-nearest Neighbor Classifiers)

k-最臨近分類的原理是利用訓練樣本的方式來進行分類。訓練樣本用 n 維數值屬性來描述，而每個樣本代表 n 維空間的一個點，也就是說所有的訓練樣本都存放在 n 維模式空間中。k-最臨近分類法搜尋模式空間，是找出最接近未知樣本的 k 個訓練樣本，這 k 個訓練樣本是未知樣本的 k 個近鄰。而近鄰是用歐基里德距離來定義的，其中兩個點 $X = (x_1, x_2, \ldots, x_n)$ 和 $Y = (y_1, y_2, \ldots, y_n)$ 的歐基里德距離為[8]：

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

最臨近分類演算法將會存放所有的訓練樣本，並且等到新的樣本需要分類時才建立分類。

2.6 搜尋引擎(Search Engine)

搜尋引擎是現今人們最常使用的一種網路服務，而搜尋引擎是一種架設在全球資訊網上的網站，做為網路上查詢服務的機制[4]。通常搜尋引擎會提供一個查詢的畫面，讓使用者可輸入要查詢的字，再經由搜尋引擎的機制將搜尋結果呈現在網頁上，使用者可以根據自己的需求輸入不同的關鍵字組合，而搜尋引擎也會因不同的關鍵字組合，搜尋出相對應的搜尋結果。目前搜尋引擎的發展依照不同的設計方式可分為下列二大類[4]：

(1) 主題目錄式

「主題目錄式」主要是由網站的管理者事先建立分類目錄，依照主題來進行分類，最後即形成樹狀的目錄結構，其中，如何分類到適當的分類目錄中這個步驟是由人工的方式來完成，利用人工的方式把相關網頁的網址或網站記錄到適當的分類目錄裡。

主題目錄式的搜尋引擎，其最大的特點是不收集網頁，只儲存了簡單的網址相關的描述。此特點節省了儲存的成本，然而當使用者利用關鍵字查詢時，卻會因為沒有記載足夠的資訊導致難以判斷使用者的需求，因此又發展出搜尋機式之搜尋引擎。

(2) 搜尋機式

「搜尋機式」主要是利用 Web Crawler 的工具到全球資訊網蒐集所有的網頁資料，其原理會將網頁取回，並且儲存在本地的資料庫，因此有足夠的資料可以判斷使用者的需求，故其優點則為當使用者輸入關鍵字查詢可做較符合的判斷，但是，因為蒐集了所有的網頁資料，所以資料量非常的龐大而造成了儲存成本的提高。

因此，為了滿足使用者的需求，搜尋的技術不斷的創新，故提出了一種混合式的搜尋引擎。「混合式搜尋引擎」主要是結合主題目錄式及搜尋機式的優點所產生的搜尋引擎，如 Yahoo! 就是一個典型的例子。除了混合式搜尋引擎外，另外還有一種新興的搜尋引擎的服務，就是將多種不同的搜尋引擎進行整合，此類的搜尋引擎稱為整合式的搜尋，或是匯總式搜尋引擎，希望藉由整合式的搜尋來提供更多元化的搜尋結果。

2.7 個人化排序

所謂的個人化排序主要是希望使用者在利用搜尋引擎時，可以經由個人化的方式，讓使用者在搜尋的過程中，更快速且更正確的得

到搜尋結果。在傳統的搜尋結果中，相同的關鍵字，極有可能尋找出相同的結果，然而不同的使用者想獲得的搜尋結果不盡相同，因此希望經由個人化的方式，記錄使用者的習慣、偏好。

利用使用者的偏好針對搜尋結果進行排序藉以達到快速且符合使用者需求的搜尋結果。根據 Y. Ke 等學者[12]提出了個人化因子在搜尋引擎中可分三大類，重要性評量、相關性評量、排序函式。下面將針對重要性評量及相關性評量進行說明[7][10][11]。

(1) 重要性評量 (Importance Measure)

「重要性評量」主要是評估網頁的重要性，重要性評量是利用使用者偏好的網頁中超連結的關係來進行評量。換句話說，當某一網頁被越多網頁鏈結時，該網頁的重要程度也會隨著提高，因此，此方法又稱為以鏈結為基礎的個人化，在這類的演算法中，往往受限於使用者的偏好，所以新的網頁或不容易被發現的網頁很難表現出其重要性，將會導致有些較好的網頁不一定具有較好的排名。

(2) 相關性評量 (Relevance Measure)

「相關性評量」主要是根據使用者查詢或過去偏好的相似度及網頁內容做為主要的判斷。這類的演算法主要的概念是分析網頁內容和使用者的偏好是否相符，因此，此方法又稱為以內容為基礎的個人化，通常相關性評量都需要參考一些相關的使用者特徵。

3.研究架構與方法

本研究中，主要是利用網頁探勘及資料探勘之關聯規則來達到個人化的目的，透過個人 web log 檔的剖析與所有使用者的瀏覽行為的分類、分群及相關性，最後期望能夠提高使用者對於搜尋結果的滿意度。本研究所提出的系統架構如圖 8 所示：

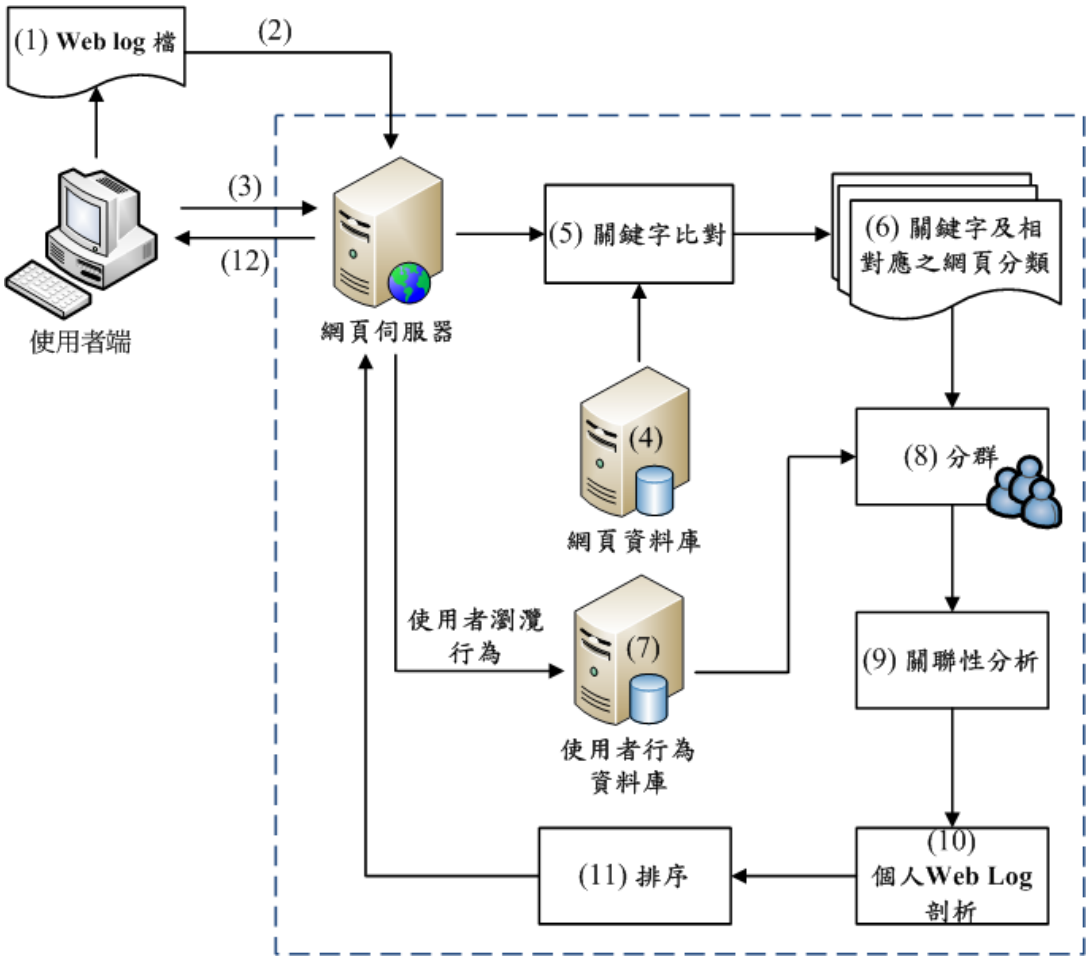

圖 8 系統架構

(1) Web log 檔記錄使用者端的資訊，例如：使用者的 IP 位址、需求日期及時間、參考到網頁的 URL 等。
(2) 網頁伺服器將記錄部分的Web log檔資訊。
(3) 使用者輸入關鍵字進行查詢的動作。
(4) 網頁資料庫事先收集所有的網頁，並儲存在資料庫。
(5) 結合使用者輸入的關鍵字與網頁資料庫進行比對的動作，找出該關鍵字所對應到的網頁。
(6) 首先，必須先建立好類別，再將上一步所得的結果，根據分類演算法，針對網頁進行分類，而在不同類別中可能會出現相同的關鍵字，但所相關連的網頁卻不會重覆。例如："蘋果"，這個關鍵字可能出現在我們預先分好的類別，水果和電腦這二類，雖然蘋果出現在二種不同的類別中，但所對的網頁會因為類別的不同，而有所不同，所以不會有重覆的情況發生。
(7) 使用者行為資料庫會記錄，每次使用者輸入關鍵字及點選網頁的行為，這個部分是接下來分群方式的基礎。
(8) 將使用者行資料庫的資料結合(6)和(7)的資料，進行分群的動作。
(9) 延續上一步驟，再利用關聯規則分析各群組之間的相關性。
(10) 根據個人 Web log 剖析，藉此得知個人瀏覽網頁的偏好。
(11) 搜尋結果會根據個人 Web log 檔剖析後，進行搜尋結果的排序。
(12) 最後將排序後的搜尋結果回傳給使用者端。

根據上述的步驟，可以得知，首先使用者輸入關鍵字來進行查詢，經由系統的網頁資料庫進行比對後，所得出的網頁會根據系統事先所設定的分類規則來進行分類，此時關鍵字會儲存在使用者行為資料庫(7)中，而分類結果(6)會與使用者行為資料庫(7)進行分群，再利用關聯規則分析各群組之間的關係，並結合個人 Web log 剖析(10)，最後，進行排序並回傳給使用者。

在整個系統的流程中，Web log 檔會記錄個人每次所輸入的關鍵字及瀏覽網頁行為，此外，每一位使用者每次所輸入的關鍵字與瀏覽行為也將儲存在資料庫中，二者皆為網頁探勘主要的資料來源。

在進行分群的過程中，群組之間的關係具有影響力，除了可以得到原本的搜尋結果外，還可獲得相關群組的搜尋結果，而搜尋的結果會依個人 Web log 檔剖析的資訊來優先排序，並列出匯整的相關搜尋結果，最後，將得到個人化排序的結果。

4. 預期結果

根據本研究的架構，在搜尋演算法方面，運用分類和分群的特性，可將使用者行為資料庫裡的資料進行分群，並利用關聯規則分析群組之間的相關性，將有助於個人化排序的工作；個人化方面，個人化排序之搜尋結果取代一般的搜尋結果會有以下的特點：

(1) 降低整體搜尋的時間。
(2) 提供其它相關類別領域的搜尋結果。
(3) 提高使用者對於搜尋結果的滿意度。

本研究以提出個人化機制的架構，滿足個人化排序的特性，經由分類、分群及相關性分析的流程，預期在搜尋的過程中，能讓使用者獲得較符合需求的搜尋結果，並且減少整體搜尋的時間。

5. 結論與未來工作

本研究以網頁探勘及搜尋演算法為核心技術，提出個人化的機制，主要是針對個人化的部分透過關聯規則與分類、分群演算法，進行整理、分析使用者資料及網頁資料，最後將搜尋的結果透過個人化排序，呈現給使用者。

未來，我們將深入研究各類的網頁探勘技術與搜尋演算法，建構出系統且進行改善，以及透過比較來篩選出最佳的搜尋演算法，並達到個人化排序的結果，最後，進行評估使用者對系統的滿意度，讓個人化排序之搜尋引擎運作流程更趨完美。

參考文獻

[1] 中華資料採礦協會，***CDMS-Newslette 第十七期***，中華資料採礦協會出版，2006。
[2] 古永嘉、羅元祠，"關聯規則在 Web Mining 的應用研究"，2003。
[3] 王孝熙、吳凱雯，"利用資料挖掘技術提供網際網路使用者個人化服務"，2001。

[4] 何正信、王仲民，"以知識本體支援網路搜尋技術之探討"，2003。

[5] 李御璽、顏秀珍、杜季樺、謝旻棋，"結合使用者瀏覽路徑與購物行為之網頁使用探勘技術"，***2004 數位生活與網際網路科技研討會***， 2004。

[6] 郭煌政、林奕森，"以叢集分析技術探討病患就診屬性與看診時間之關係"，2002。

[7] 陳榮昌、蔡旺典，"以知識本體來輔助個人化排序"，***第十二屆資訊管理暨實務研討會***， 2006。

[8] 曾龍，***資料探礦概念與技術***，維科圖書有限公司，2003。

[9] Cooley, R., Mobasher, B. and Srivastava. J., "Data Preparation for Mining World Wide Web Browsing Patterns," ***Knowledge and Information System***, Vol. 1, No. 1, pp. 5-32, 1999.

[10] Chakrabarti, S., Dom, B. E., Kumar, S. R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D. and Kleinberg, J., "Mining the Web Link Structure," ***Computer***, Vol. 38, No. 8, pp. 60-67, 1999.

[11] Daugherty, P. J., Rogers, D. S. and Spencer, M. S., "Just-in-Time Functional Model: Empirical Test and Validation," ***International Journal of Physical Distribution and Logistics Management***, Vol. 24, No. 6, pp. 20-26, 1994.

[12] Deng, Ke. Y., W, L. Ng., and Dik-Lun, Lee., "Web Dynamics and their Ramifications for the Development of Web Search Engines," ***Computer Networks: The International Journal of Computer and Telecommunications Networking***, Vol. 50, No. 10, pp. 1430-1447, 2006.

[13] Manber, U., Pater, A. and Robison, J., "Experience with Personalization on Yahoo!," ***Communication of the ACM***, Vol. 43, No. 8, pp. 35-39, 2000.

[14] Mobasher, B., Cooley, R. and Srivastava, J., "Automatic Personalization Based on Web Usage Mining," ***Communication of the ACM***, pp. 142-151, 2000.

[15] Spiliopoulou, M., "Web Usage Mining for Web Site Evaluation," ***Communication of the ACM***, pp. 127-134, 2000.

[16] Zaiane, O. R., "Resource and Knowledge Discovery from the Internet and Multimedia Repository," ***Ph.D Dissertation***, Simon Fraser University, March, 1999.