利用相關回饋建立概念化的使用者興趣檔以協助使用者進行網頁查詢

周世傑
中央大學資訊管理學系

吳政穎
中央大學資訊管理學系

摘要

現今網路的搜尋引擎主要都是設計給大眾做資訊檢索，對於不同背景與需求的使用者，當提供一個相同的查詢句，所得到的搜尋結果會都是相同的大量網頁，這使得個人化搜尋的需求越來越高。使用者興趣檔描述了一個特定使用者的興趣，在資訊檢索的系統裡，通常用來幫助搜尋引擎提供個人化的搜尋結果，或者應用在推薦系統上。當今大部分的使用者興趣檔，是反映使用者長期的資訊需求，而非單次檢索的資訊需求。

本研究提出了一套方法，應用了概念擷取的技術，來幫助使用者建立一種短期的使用者興趣檔，希望藉由相關回饋來擷取出一個向量空間模型，用來代表使用者單次檢索的資訊需求，以幫助系統檢索出，與使用者需求相關的網頁給使用者瀏覽，以降低使用者的瀏覽時間，提高檢索的效率。

關鍵字：使用者興趣檔、資訊檢索、資訊需求、概念擷取、向量空間模型

感謝國科會(NSC96-2416-H-008-011-MY2)對本研究之經費補助。

Applying Relevance Feedback to Develop the Conceptual User Profile to Assist the User in Web Search

Shih-Chieh Chou
Department of Information Management, National Central University

Chen-Gying Wu
Department of Information Management, National Central University

Abstract

Nowadays, the search engine has been designed to retrieve information. With the same query, the users will get the same result which usually is a mass amount of web pages although the users' interests might be different. Therefore, enhancement on personalized search is always needed. The user profile which depicts the user's search interest has been applied to assist the user in the searching of personalized information in searching engine. Contemporary user profile mostly has been designed to reflect the user's long-term information interest.

This study has proposed a method to develop the user profile which could reflect the user' s short-term information interest. In the method, the information of relevance feedback has been utilized to extract the user's concepts of interest. These concepts of interest are then represented in a vector space model as the user profile for machine calculation. With the application of the vector space model of interest concepts, the web pages of the user's interest could be better retrieved to enhance retrieval efficiency.

Key words : User Profile, Information Retrieval, Information Need, Concept Extraction, Vector Space Model.

壹、導論

由於網際網路的成熟發展，使得資訊在網路上的流通越來越廣泛，網站上提供的資訊，逐漸成為人們獲取資訊的主要來源。在資訊檢索上，使用簡單的關鍵字來幫助使用者搜尋網路資訊，是廣泛而普遍的檢索方式。但是現今網路的搜尋引擎，主要提供給大眾使用，搜尋引擎並無法區分不同背景的使用者，也無法明確的得知使用者的資訊需求為何，所以對於一個相同的查詢句，所得到的搜尋結果都是相同的。研究指出，有超過80%的使用者，希望在資訊檢索時能有個人化的搜尋結果 (Sullivan 2004)。

因為網路上的資訊量越來越多，只在搜尋引擎上輸入幾個關鍵字，無法很明確的描述使用者的資訊需求，導致檢索到的資訊量通常過於龐大。對使用者來說，要在大量文章中找到真正相關的文章，往往相當耗時。使用者興趣檔(User Profile)是用來描述一個特定使用者的興趣，在資訊檢索的領域裡，通常用來幫助搜尋引擎提供個人化的搜尋結果，或者應用在推薦系統上。使用者興趣檔的建立方式可區分為直接式與間接式(Mulvenna et al. 2000; Sugimoto 1999)。不論是直接式的建立方式還是間接式的建立方式，使用者的瀏覽歷史(Browsing Histories)是最常拿來應用的興趣資訊，有許多的研究都使用瀏覽歷史來建立使用者興趣檔(Trajkova & Gauch 2004; Liu et al. 2002; Kim & Chan 2003; Spetetta & Gauch 2005)。

但是這些使用者興趣檔所反映的是使用者的長期資訊需求，並非單次檢索的資訊需求。舉例來說，一個具有電腦科學背景的人，在檢索java這個字詞時，不論他是否想找有關咖啡的文章，透過使用者興趣檔所檢索出來的結果，可能幾乎都是程式語言相關的文章。在這種情況下，使用者必須花時間去精確化查詢句，才能達到檢索目的。因此，本研究的目的是在提出一套方法，對於使用者回饋的相關文章和非相關文章，擷取出文章特徵，用以建立一個概念化的使用者興趣檔，以反映使用者的查詢需求。

貳、概念的擷取

本研究主要在利用相關回饋的網頁，建立代表使用者查詢概念的使用者興趣檔，因此，如何將使用者的查詢概念從回饋網頁的內文擷取出來，即成為本研究重要的議題。在資訊檢索的領域中，常用單獨的字詞，有關聯的字詞或者一個段落來表達一個特定領域的概念。因此在概念的擷取上，常會以文章內文所傳達的語意來擷取概念。在Chang et.al. (2006)的研究裏，先以關鍵字群集、標題字詞頻率和句子位置這三種準則，找出哪些句子在文章中具有意義，再將這些有意義的句子轉換成向量模式，以萃取出文章的特徵向量，最後將特徵向量透過分群以形成概念。一般網頁上的連結文字，往往是網頁中較重要的文字，但由於連結文字常常由幾個少許的字詞所組成，如將Chang et.al.利用句子來擷取概念的方式，應用在網頁上，可能會因此忽略了網頁上某些重要字詞。

另外一種概念擷取的方法是去分析字詞之間的共現性(Co-occurrence)，當某幾個特定字詞總是出現在同一篇文章中，則這些字詞之間必定存在著某種關連，其背後也可能隱含著某種概念。在資料探勘(Data Mining)裡，常用關聯規則(Association rule)來找出字詞之間的關聯性。Rahman et. al. (2003)的研究，就是利用關連規則裏高頻項目集合[1](Frequent item set or Large item set)的觀念，找出關鍵字共同出現的關聯，而這些具有關聯的關鍵字集合，即可用來表示一種特定概念。因此，如果文章中同時出現具有關聯性的關鍵字，則可以認定此篇文章包含著某種概念。

Karoui et. al. (2006)的研究，則是利用網頁特徵[2]來判斷字詞的共現性。首先用一個上下文階層組織(Contextual Hierarchy)的標籤階層，來闡明整個HTML文件中，所有標籤之間可能存在的關係。藉由上下文標籤階層，即可定義字詞之間的連結，而找出字詞間的共現性，最後去分析HTML文件中標題標籤與關鍵標籤裡的字詞，找出與這些字詞存在著共現性的字詞後，將之轉成向量空間模式，進行分群來得到特定領域的概念。上述利用共現性來擷取概念的方式，需先蒐集大量的資料進行分析以建立樣式(Pattern)，並依此樣式找出字詞的共現性，因此，需耗費許多時間在建立樣式上。

除了利用文章語意和字詞共現性來擷取概念之外，Lecceuche (2000)則是利用對文件的鑑別力，來找出文件中重要的概念。其主要想法為，如果某一些字詞可以用來區分兩份文件間的差異，則這些字詞應視為一個重要的概念。Lecceuche使用part-of-speech tagger工具，將文件中無法有效區分文件的字詞過濾掉，接著以n元詞[3](n-gram)來識別文件中潛在的概念，在捨棄出現頻率較低的概念後，以一個基準文件當作比較對象，計算此概念在原文與基準文件中出現的頻率比率，如果在原文中的出現頻率越高則代表此概念是越重要的。但如何去選擇一個良好的基準文件來作為比對的對象，則成為這方法成敗的關鍵與否。

上述的文章概念擷取方法，皆是利用字詞的組合來表示概念，因此字詞權重的計算方式，就成為概念擷取的重要部分。Salton & Buckley (1988)所提出的TFIDF，是在向量空間模式(Vector Space Model)中，用來描述文章，最知名的字詞權重方法。TFIDF首先計算某個特定的字詞在某篇文章中出現的次數，如果此字詞在文章中出現的頻率越高，則認為這個字詞在這篇文章中越有意義；除此之外，如果某個字詞在整個文集中，出現的頻率越高，則此字詞在文章中的代表性越低，將有可能被認定為停用字(Stop Word)，換句話說，一個字詞的權重決定於上述兩種因素。

另一種決定字詞權重的方式，則多加考量了類別資訊(Category Information)。Soucy & Mineau (2005)所提出的信賴權重(Confidence Weight)，是基於統計上的信賴區間來決定字詞的權重。其主要想法為，如果在類別c_j中，文章包含字詞t_i的比例，跟在非類別c_j中，文章包含字詞t_i的比例，有明顯的差距時，則t_i會被認定是具有重要性的字詞，應該

[1] 若項目集出現頻率大於最小支持度，則稱此項目集為高頻項目集合。

[2] 網頁的特徵是指HTML文件包含許多不同的標籤屬性，例如：標題標籤、表格標籤、段落標籤…等不同的標籤。

[3] n元詞是在文件中，一連串連續出現的n個字詞。

得到高的權重值。上述的字詞權重方法主要是用於一般純文字文件，Fresno & Ribeiro (2004)則提出了一套混合式的字詞權重方法(Analytic Combination of Criteria, ACC)，用於決定網頁上的字詞相關度。這套方法考量了字詞的頻率，標題字詞的頻率，強調標籤[4]中的字詞頻率和字詞在文章中的位置頻率來算出字詞的權重值。

參、概念化使用者興趣檔

本研究的構想，是對使用者在搜尋引擎上所檢索回來的搜尋結果，利用使用者的相關回饋，建立概念化的使用者興趣檔。其方法為，當搜尋引擎傳回資訊需求者所檢索的網頁，我們先針對每篇網頁進行分析，擷取網頁中具有代表性的字詞，形成特徵向量來表示此篇網頁，然後在資訊需求者進行相關回饋的動作後，我們再對相關和非相關網頁，擷取出此次使用者資訊檢索的查詢概念，並用此概念來建置使用者興趣檔，圖1顯示了該使用者興趣檔的建置流程。在完成使用者興趣檔建置後，接著可使用餘弦函式去計算使用者興趣檔和網頁的相似度，以協助網頁查詢。後續將先對圖1中各階段工作加以說明，之後再說明相似度的處理。

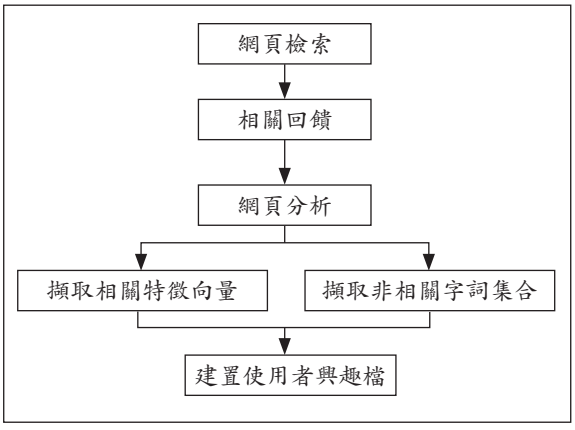

圖1：使用者興趣檔建置流程

一、 網頁檢索

本階段工作在將使用者的查詢字串發送給外部搜尋引擎，以獲取搜尋結果，然後把搜尋結果回傳給使用者，其間採用 Google 作為搜尋引擎。採用Google 的原因，是因本研究著重在網頁文字的處理，而 Google 是用連結分析，查詢的結果可以遠超過其索引資料庫的範圍。

[4] Fresno and Ribeiro將html標籤<*b*> </*b*>, <*u*> </*u*>, <*em*> </*em*>, <*i*> </*i*>, <*strong*> </*strong*>皆視為強調標籤。

二、相關回饋

本階段工作主要在提供使用者介面，在使用者瀏覽過搜尋結果後，能夠方便的從搜尋結果中勾選相關與非相關網頁，並將此些資訊回傳系統，以完成相關回饋，此些相關回饋提供的資訊，將成為後續分析使用者資訊需求的基礎。

三、網頁分析

在Fresno & Ribeiro (2004)的研究裏，對於網頁字詞相關性的決定，考量了字詞頻率、標題中的字詞頻率、強調標籤中的字詞頻率和字詞在文章中的位置頻率這四個準則，植基於這套機制，我們發展了本研究的網頁特徵向量擷取模式。

首先我們將每篇檢索回來的網頁內容使用CLEF[5]的停用字列表(StopList)去除停用字(Stopwords Removal)，再用Poter演算法將剩下的字詞做字根化(Stemming)的動作，如此即可得到代表每篇網頁的字詞集合，即網頁$webpage_i=\{term_1,term_2,term_3, ...term_i\}$。

得到網頁的字詞集合後，我們計算在字詞集合裡每個字詞的字詞頻率。字詞頻率計算方式如式3-1所示，其中$n_f(i)$指字詞i在網頁內文中出現的次數，N_{tot}指網頁內文的字詞總數。接著分析此篇網頁裡標題標籤中和強調標籤中的出現的字詞，之後針對分析到的字詞，分別去計算這兩種標籤的字詞頻率。式3-2定義了標題字詞的頻率，$n_t(i)$表示字詞i出現在標題標籤內的次數，N_{tit}則表示在標題標籤內的字詞總數。式3-3定義了強調字詞的頻率，$n_e(i)$表示字詞i出現在強調標籤裡的次數，N_{emph}則表示在整篇文章中，所有強調標籤內的字詞總數。

因為現今網頁的內容越來越豐富，也越來越複雜，一個網頁的內容可能不再以文字為主，取而代之的可能是圖片或影音檔案，而這類型的網頁內文並沒有明顯的段落，可能只有幾句敘述的文字，所以我們捨棄Fresno & Ribeiro所提的字詞位置頻率這個準則。最後字詞$term_i$權重值的計算方式如式3-4所示，$f(i)$代表字詞頻率值，$t(i)$表示標題的字詞頻率值，$e(i)$表示強調字詞頻率值。在係數的設定上，我們係依據Fresno & Ribeiro的係數設定，並把原本位置頻率的係數值轉加到字詞頻率的係數上，各係數設定為：$C_f=0.60$, $C_t=0.15$, $C_e=0.25$。計算完權重值後的字詞集合，可以轉成一個字詞向量，$webpage_i=[term_1,term_2,term_3, ...term_i]$，此字詞向量即為代表網頁的特徵向量。

$$f(i)=n_f(i)/N_{tot} \qquad \text{式3-1}$$

$$t(i)=n_t(i)/N_{tit} \qquad \text{式3-2}$$

$$e(i)=n_e(i)/N_{emph} \qquad \text{式3-3}$$

$$weight_i = C_f f(i) + C_t t(i) + C_e e(i) \qquad \text{式3-4}$$

[5] 歐盟成立的「跨語言資訊檢索論壇」(Cross Language Evaluation Forum, 簡稱CLEF)，為一個主要的資訊檢索評比單位。

四、擷取相關特徵向量

在此階段，我們從使用者相關回饋的網頁中，隨機選擇幾篇相關網頁，並把這幾篇網頁的內文結合起來，視為同一篇網頁內文，以前述的方法進行網頁的特徵向量擷取，得到一個相關特徵向量，$relevant=[term_1,term_2,term_3, ...term_i]$，此相關特徵向量即代表此次查詢的正向興趣資訊。

五、擷取非相關字詞集合

在此階段，我們同樣從使用者相關回饋的網頁，隨機選擇幾篇非相關網頁，並把網頁內文結合起來，進行去除停用字與字根化的動作，接著我們蒐集剩餘的字詞形成一個非相關的字詞集合，$nonrelevant=\{term_1,term_2,term_3, ...term_j\}$，此非相關的字詞集合即代表此次查詢的負向興趣資訊。

六、建置使用者興趣檔

此階段在建置使用者興趣檔。關於使用者興趣檔的建立，本研究應用了字詞在相關與非相關文章中出現的特性。假設某個字詞同時出現在相關與非相關的文章時，我們認為，這個字詞對於此次資訊檢索的重要性應不高，其相關程度也應該不高；而當某個字詞只在相關文章中出現，我們認為，這個字詞對於此次資訊檢索的重要性不低，其相關程度應該提高。我們將這種字詞出現特性稱為字詞敏感度(Sentivity)，並將字詞敏感度應用在使用者興趣檔的建立上。

Zacharis and Panayiotopoulos (2001)曾利用字詞的相關程度來調整字詞頻率，我們依此概念發展出另一公式來調整字詞頻率值，如式3-5所示，Tf表示字詞頻率，$Sensitivity$表示字詞的敏感度。對於字詞的敏感度，我們依序比對相關向量裡的字詞$term_i$，如果字詞$term_i$沒有出現在非相關字詞集合裡，表示此字詞$term_i$的敏感度較高，反之，則表示敏感度較低。我們依據前置測試，將$Sensitivity$分別設定為1.2與1.0，如表1所示。我們用此方式結合使用者的正向興趣資訊與負向興趣資訊，並依此產出使用者的查詢概念。調整完字詞頻率值後的相關特徵向量，即為此次資訊檢索的概念化使用者興趣檔。使用者興趣檔的字詞權重值的算法如式3-6所示，$f_{new}(i)$表示調整完後的字詞頻率值，$t(i)$表示標題的字詞頻率值，$e(i)$表示強調字詞頻率值，$C_f=0.60$, $C_t=0.15$, $C_e=0.25$。

$$Tf = Tf \times Sensitivity \qquad \text{式3-5}$$

表1：字詞敏感度係數

字詞出現特性	Sensitivity
沒有出現在非相關字詞集合	1.2
有出現在非相關字詞集合	1.0

$$weight_i = C_f f_{new}(i) + C_t t(i) + C_e e(i) \qquad \text{式3-6}$$

在完成使用者興趣檔的建置後，我們接著使用餘弦函式去計算使用者興趣檔和網頁的相似度，用以協助查詢使用者的需求網頁。餘弦函式計算相似度的向量空間模型是由Salton & Lesk (1968)所提出的，其主要的概念是將文章先轉換成為多維度空間中的一個向量，並將文章中的重要字詞當作空間中的維度，向量中的量值則代表字詞的權重值。當文章轉換完向量後，接著使用餘弦函式(cosine)來計算兩向量的相似度，其值介於0到1之間，兩向量的夾角成為90度時，餘弦值等於0，代表兩篇文章完全沒有任何相似之處，反之如果兩向量的夾角為0度時，餘弦值等於1，則表示著兩篇文章在此多維度空間中是高度相似的。

我們將每篇網頁在網頁分析階段中，所得到的網頁特徵向量，$webpage_i$=[$term_1$,$term_2$,$term_3$, ...$term_i$]，一一與概念化使用者興趣檔進行餘弦函式的相似度計算，相似度函式如式3-7所示。其中，$\sum_{i=1}^{n} t_{1i}^{2}$為文章$d_1$關鍵字詞的權重值平方和；$\sum_{i=1}^{n} t_{2i}^{2}$為文章$d_2$關鍵字詞的權重值平方和；$\sum_{i=1}^{n} (t_{1i} \times t_{2i})$為關鍵字在文章$d_1$的權重值與在文章$d_2$的權重值相乘和。

$$sim(d_1,d_2) = \frac{\sum_{i=1}^{n} (t_{1i} \times t_{2i})}{\sqrt{\sum_{i=1}^{n} t_{1i}^{2}} \times \sqrt{\sum_{i=1}^{n} t_{2i}^{2}}} \qquad \text{式3-7}$$

在計算完所有搜尋網頁與概念化使用者興趣檔之間的相似度後，我們可將每篇搜尋網頁附上各自的相似度回傳給使用者，提供使用者參考，使之可依照相似度判定此篇網頁為相關網頁或者非相關網頁，希冀藉此模式以降低使用者整體瀏覽的時間。

肆、實驗分析

本研究主要在探討，如何對於使用者回饋的相關和非相關文章，擷取出文章特徵，用以建立一個概念化的使用者興趣檔。因此，本研究的實驗設計在驗證，依本系統方法所建立的概念化使用者興趣檔，是否正確反映使用者的查詢需求。

一、實驗設計

本章進行之系統測試，共分兩個實驗，其中實驗一之主要目的，除了評估使用者回饋的負向資訊能否有效的提升使用者興趣檔的正確性之外，同時也是為了驗證本研究所提出的字詞出現特性之應用是否成立，而實驗二的主要目的，則在評估本研究提出的方法所建置的概念化使用者興趣檔，是否能有效的代表使用者的查詢概念。參與本研究之實驗測試者共有5人，教育程度皆在研究所以上，具英文閱讀能力，且使用全球資訊網搜尋資料之經驗至少為5年。受測者查尋主題為自由選定，本系統並不限定使用者的查詢主題。

在實驗一的部分，使用者透過Google搜尋引擎進行資訊檢索，並從回傳的搜尋結果中選取15篇相關網頁與15篇非相關網頁回饋給予本系統，接著本系統從回饋的網頁中隨

機選取3篇相關網頁文章與3篇非相關網頁文章，來當作使用者的正向與負向興趣資訊，並利用本研究提出的方法，從中擷取出相關特徵向量與非相關字詞集合，來建立概念化使用者興趣檔；同時僅利用正向興趣資訊，另外再建立的一個不包含負向興趣資訊的使用者興趣檔。然後將此兩個使用者興趣檔，分別與相關網頁和非相關網頁進行相似度比對，來評估包含負向興趣資訊的使用者興趣檔，是否能有效的提升檢索的正確性。

在實驗二，我們將實驗一建立的概念化使用者興趣檔，與利用TFIDF建立的使用者興趣檔作比較。我們先利用TFIDF建立用一個特徵向量來代表TFIDF使用者興趣檔，再將兩個使用者興趣檔，同樣進行相關網頁與非相關網頁的相似度比對，來評估何種方式建立的使用者興趣檔，較能代表使用者的查詢概念。

二、實驗分析

在實驗一，使用者興趣檔與相關網頁進行相似度計算的部分，我們係將使用者回饋的15篇相關網頁扣除掉建立特徵向量的3篇網頁後，拿剩餘的12篇相關網頁文章當作相似度計算的對象。圖2為平均相關相似度，由圖得知，使用者回饋的負向興趣資訊，對於使用者興趣檔與相關網頁間的相似度計算，並無明顯的影響力。我們同樣拿使用者回饋的15篇非相關網頁，扣除掉建立非相關字詞集合的3篇網頁後，以剩餘的12篇非相關網頁文章當作相似度計算的對象，圖3為平均非相關相似度，由圖得知，包含負向興趣資訊的概念化使用者興趣檔，與非相關網頁間所計算出的相似度較低，這表示使用者回饋的負向興趣資訊對於使用者興趣檔與非相關網頁間的相似度計算具有一定的影響力。

表2為包含負向興趣資訊的概念化使用者興趣檔，與不包含負向興趣資訊的使用者興趣檔，之相似度比較。我們將每位使用者在這兩種不同興趣檔所算出的平均相似度拿來做對照，其中包含使用者負向興趣資訊的概念化使用者興趣檔所算出的相關相似度平均提高了0.00137，而在非相關相似度平均降低了0.03124。在建立概念化使用者興趣檔時，我們省略了非相關字詞集合裡的字詞，增強了只在正向興趣資訊裡出現的字詞，因此包含負向興趣資訊的使用者興趣檔在與非相關網頁進行相似度比對，所算出的相似度，自然比不包含負向興趣資訊的使用者興趣檔還低，所以雖然負向興趣資訊對於檢索出相關網頁的效能不大，但確實能有效的過濾掉非相關網頁，提升檢索時的正確性，同時也證明了本研究所提出的字詞出現特性之應用是成立的。

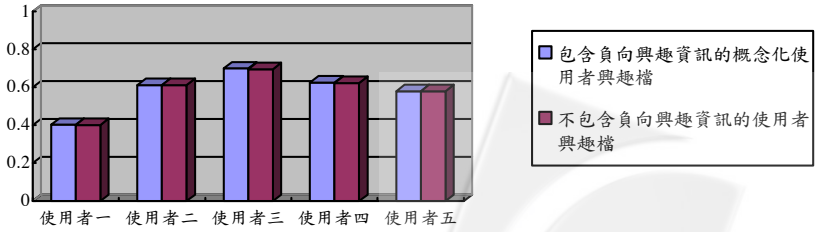

圖2：平均相關相似度

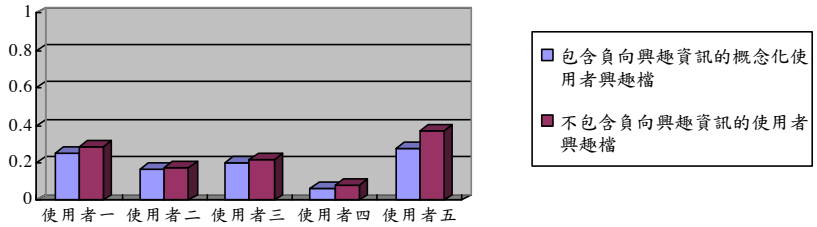

圖3：平均非相關相似度

表2：概念化使用者興趣檔與不包含負向興趣資訊使用者興趣檔相似度比較表

興趣檔\使用者	使用者一	使用者二	使用者三	使用者四	使用者五	平均
相關相似度						
含負向資訊	0.40276	0.61150	0.69990	0.62554	0.57936	0.00137
不含負向資訊	0.40176	0.61189	0.69584	0.62272	0.57997	
相差值	0.00100	-0.00039	0.00406	0.00282	-0.00061	
非相關相似度						
含負向資訊	0.25584	0.16942	0.20178	0.06544	0.28165	-0.03124
不含負向資訊	0.28667	0.17647	0.21777	0.07755	0.37184	
相差值	-0.03083	-0.00705	-0.01599	-0.01211	-0.09019	

在實驗二，使用者興趣檔與相關網頁計算的部分，我們同樣拿實驗一12篇的相關網頁文章當作相似度計算的對象。由於使用者回饋的網頁文章隱含其查詢概念，因此我們針對每位使用者的每筆相關相似度進行加總的動作，並用此數值來量化使用者的查詢概念，如果相關相似度的加總值越高，則代表此興趣檔越接近使用者的查詢概念。圖4為相關相似度加總圖，由圖可知，以概念化使用者興趣檔所算出的相似度加總值明顯的高於TFIDF使用者興趣檔的相似度加總值，即表示概念化使用者興趣檔在進行相關網頁比較的效能，是優於TFIDF使用者興趣檔。我們同樣拿實驗一12篇的相關網頁文章當作相似度計算的對象，並進行非相關相似度加總的計算。同理，我們用非相關相似度加總值來量化使用者的查詢概念，如果非相關相似度加總值越低則代表越接近使用者的查詢概念。圖5為非相關相似度加總圖，可看出以概念化使用者興趣檔所算出的相似度加總值略高於TFIDF使用者興趣檔的相似度加總值，這也表示概念化使用者興趣檔在進行非相關網頁比對的效能，略遜於TFIDF使用者興趣檔。

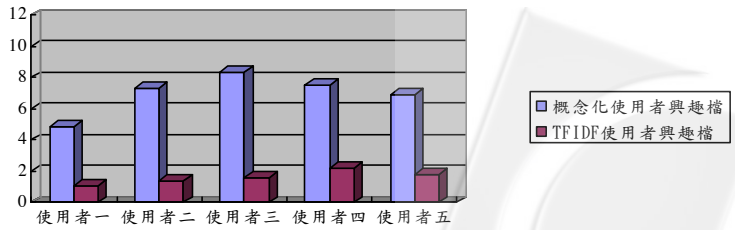

圖4：相關相似度加總

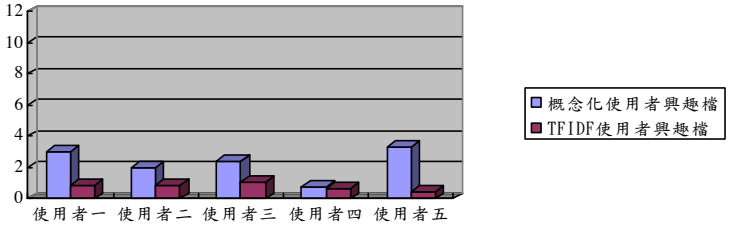

圖5：非相關相似度加總

雖然概念化使用者興趣檔在與使用者回饋的非相關網頁比對時，相似度略高於TFIDF使用者興趣檔所比對的相似度，但其非相關相似度提升的幅度小於相關相似度的提升幅度。表3為概念化使用者興趣檔與TFIDF使用者興趣檔相似度的比較表，我們將每位使用者在兩種不同興趣檔所算出的平均相似度拿來做比較，其中相關相似度平均提高了0.44831，而非相關相似度平均只提高了0.12839。表4則為兩種使用者興趣檔的相關相似度與非相關相似度的差距比較表，我們將表4的實驗數據轉成圖6的相似度差距圖，由圖可看出，概念化的使用者興趣檔在相關相似度與非相關相似度間的差距值較大，這表示，相較於TFIDF使用者興趣檔，當以概念化使用者興趣檔進行網頁文章相似度比對，更可以有效的區別出相關網頁與非相關網頁，亦即其鑑別力較高。

表3：概念化使用者興趣檔與TFIDF使用者興趣檔相似度比較表

興趣檔\使用者	使用者一	使用者二	使用者三	使用者四	使用者五	平均
相關相似度						
概念化	0.40276	0.61150	0.69990	0.62554	0.57936	0.44831
TFIDF	0.08870	0.11600	0.13656	0.18718	0.14911	
相差值	0.31406	0.49550	0.56334	0.43836	0.43025	
非相關相似度						
概念化	0.25584	0.16942	0.20178	0.06544	0.28165	0.12839
TFIDF	0.07154	0.07450	0.08728	0.05747	0.04135	
相差值	0.18430	0.09492	0.11450	0.00797	0.24030	

表4：相關相似度與非相關相似度差距值比較表

網頁\使用者	使用者一	使用者二	使用者三	使用者四	使用者五	平均
概念化使用者興趣檔						
相關網頁	0.40276	0.61150	0.69990	0.62554	0.57936	0.38953
非相關網頁	0.25584	0.16942	0.20178	0.06544	0.28165	
相差值	0.14962	0.44208	0.49812	0.56010	0.29771	
TFIDF使用者興趣檔						
相關網頁	0.08870	0.11600	0.13656	0.18718	0.14911	0.06908
非相關網頁	0.07154	0.07450	0.08728	0.05747	0.04135	
相差值	0.01716	0.04150	0.04928	0.12971	0.10776	

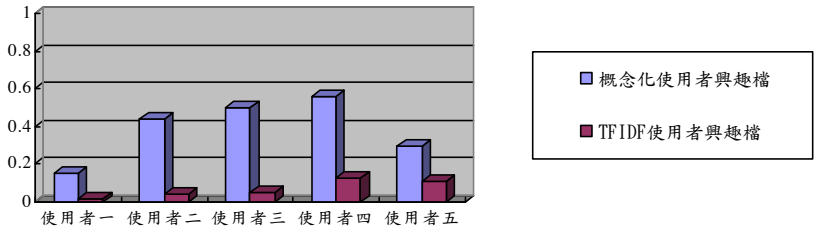

圖6：相似度差距圖

對於前述各項實驗所驗證的事項，目的，方法與結果，我們將之整理於表5。這些實驗分析與結果顯示，本研究所驗證的方法與系統，其基於使用者回饋的網頁文章，所建置的概念化使用者興趣檔，確有可行之處。其中包含了正向資訊的應用，正負向資訊交互利用，以及字詞出現特性的應用；另外，也證明了本系統所建置的概念化使用者興趣檔，確比TFIDF使用者興趣檔，更能表示出使用者的查詢概念。

表5：實驗彙整

驗證事項	驗證目的	驗證方法	驗證結果
使用者負向興趣資訊的重要性。	評估使用者回饋的負向資訊能否有效的提升使用者興趣檔的正確性，同時也驗證本研究所提出的字詞出現特性之應用是否成立。	建立包含負向興趣資訊與不包含負向興趣資訊的使用者興趣檔，將此兩種使用者興趣檔，分別與相關網頁和非相關網頁進行相似度比對，並觀察比對的相似度。	雖然負向興趣資訊對於檢索出相關網頁的效能不大，但確實能有效的區別非相關網頁，提升檢索時的正確性，同時也證明了本研究所提出的字詞出現特性之應用是成立的。
概念化使用者興趣檔的效能。	評估本研究提出的方法所建置的概念化使用者興趣檔，是否能有效的代表使用者的查詢概念。	建立TFIDF使用者興趣檔與概念化使用者興趣檔，將兩者分別與相關網頁和非相關網頁做相似度比對，並觀察比對的相似度。	相較於TFIDF使用者興趣檔，概念化使用者興趣檔在進行網頁文章相似度比對時，更可以有效的區別出相關網頁與非相關網頁。

伍、結論

本研究提出了一個個人化的網頁查詢方法。我們利用使用者相關回饋的網頁，來擷取出相關特徵向量與非相關字詞集合，並依此建立出代表使用者單次查詢概念的概念化使用者興趣檔。本研究的主要貢獻，即在於概念化使用者興趣檔的建立，我們除了利用包含使用者的正向興趣資訊的相關特徵向量之外，同時利用了非相關字詞集合，並搭配本研究提出的字詞出現特性來調整興趣檔字詞的權重。經由實驗的測試結果，驗證了在利用使用者興趣檔做資訊檢索時，使用者回饋的負向興趣資訊，的確能有效的提升檢索時的正確性，同時也證明了本研究所提出的字詞出現特性之應用是成立的。

為了驗證概念化使用者興趣檔是否能有效的表示使用者單次的查詢概念，我們進行了概念化使用者興趣檔與使用者回饋的網頁文章間的相似度評估，並以TFIDF當做比較對象。實驗的測試結果說明了，相較於TFIDF使用者興趣檔，概念化使用者興趣檔在進行網頁文章相似度比對時，是更能有效的區別出相關網頁與非相關網頁，這也驗證了概念化使用者興趣檔可以有效的表示出使用者的查詢概念。受制於時間與物力，本研究提出的方法，僅限於以文字為主之查詢，未來，可以擴展相關研究如下：

1. 處理不同形式的資訊型態：本研究僅利用網頁上的文字部分來分析網頁內容與建置使用者興趣檔，未來可考量處理其他更複雜的資訊型態，例如網頁上的圖片、音樂、影片…等多媒體資訊，如此，即可以以更多的資訊來分析網頁內容，尋求建立出更精確的使用者興趣檔。
2. 擴充使用者興趣檔的應用面：本研究僅利用使用者興趣檔來協助使用者進行網頁查詢，如能擴展使用者興趣檔的應用面，例如將使用者興趣檔應用在查詢延伸擴展(Query Expansion)上，或者利用在網頁的分群上，將能增加使用者興趣檔的效用，提升資訊檢索的效能。

參考文獻

1. Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval*, Addison-Wesley, New York, 1999.
2. Chang, Y., Kim, M. and Raghavan, V. V.“Construction of Query Concepts Based on Feature Clustering of Documents,”*Information Retrieval* (9:3), 2006, pp. 231-248
3. Fresno, V. and Ribeiro, A.“An Analystical Approach to Concept Extraction in HTML Environments,”*Journal of Intelligent Information Systems* (22:3), 2004, pp. 215-235
4. Griffith, A., Luckhurst, H.C. and Willet, P.“Using Inter-Document Similarity Information in Document Retrieval Systems,”*Journal of the American Society for Information Science* (37:1), 1986, pp. 3-11
5. Karoui, L., Fufaure, M.A. and Bennacer, N.“A New Extraction Concept based on Contextual Clustering,”*Proceedings of the International Conference on Computational Intelligence for Modeling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce*, Washington, D.C., USA, 2006, p.91
6. Kim, H. R. and Chan, P. K.“Learning Implicit User Interest Hierarchy for Context in Personalization,”*Proceedings of the 8th international conference on Intelligent user interfaces*, Miami, Florida, USA, 2003, pp. 101-108
7. Lecceuche, R.“Finding Comparatively Important Concepts between Texts,”*Proceedings of the 15th IEEE international conference on Automated software engineering*, Washington, D.C., USA, 2000, p.55

8. Liu, F., Yu, C. and Meng, W. "Personalized Web Search by Mapping User Queries to Categories," *Proceedings of the eleventh international conference on Information and knowledge management*, McLean, Virginia, USA , 2002, pp. 558-565
9. Mulvenna, M. D., Anand, S. S. and Buchner, A.G., "Personalization on the Net using Web Mining," *Comm. Of the ACM* (43:8), 2000, pp. 123-125
10. Rahman, M., Chowdhury, F. A., Sohel, P. N. and Kamruzzaman, S. M. "Text Classification Using the Concept of Association Rule of Data Mining," *Proceeding of the International Conference on Information Technology*, Maribor, Slovenia, 2003, pp. 23-26
11. Salton, G. and Buckley, C. "Term-Weighting Approaches in Automatic Text Retrieval," *Information Processing and Management* (24:5), 1988, pp. 513-523
12. Salton, G. and Lesk, M. E. "Computer Evaluation of Indexing and Text Processing," *Journal of the ACM* (15:1), 1968, pp. 8-36
13. Soucy, P. and Mineau, G. W. "Beyond TFIDF Weighting for Text Categorization in the Vector Space Model," *International Joint Conference on Artificial Intelligence*, Edinburgh, Scotland, 2005, pp. 1130-1135
14. Speretta, M. and Gauch, S. "Personalized Search Based on User Search Histories," *Proceedings of the 2005 IEEE/WIC/ACM international Conference on Web intelligence*, Washington, D.C., 2005, pp. 622-628
15. Sugimoto, M. "User Modeling and Adaptive Interaction in Information Gathering System," *Journal of Japanese Society for Artificial Intelligence* (14:1), 1999, pp. 25-32
16. Sullivan, D. "The Older You Are, The More You Want Personalized Search," 2004, http://searchenginewatch.com/searchday/article.php/3385131
17. Trajkova, J. and Gauch, S. "Improving Ontology-Based User Profiles," *Proceedings of RIAO*, University of Avignon, Vaucluse, France, 2004, pp. 380-389
18. Zacharis, Z. N. and Panayiotopoulos, T. "Web Search Using a Genetic Algorithm," *IEEE Internet Computing* (5:2), 2001, pp. 18-26