第一章

緒論

近年來，網際網路的快速成長，連帶著全球資訊網(WWW)也蓬勃的發展，有越來越多的使用者依藉著全球資訊網來搜尋資訊。也因此，搜尋引擎成為使用者在全球資訊網瀏覽時不可或缺的工具之一。包括 Google、Yahoo 等都是熱門的搜尋引擎。在使用搜尋引擎時，使用者多利用關鍵字來表達其資訊需求，以求助搜尋系統幫助使用者找尋其所需的資訊。

然而，使用者真的能利用搜尋引擎找到想搜尋的資訊嗎？對於一個剛剛接觸網路搜尋的**生手**(Novice User)而言，可以準確的抓到確切且適當的關鍵字給予搜尋引擎嗎？即使是網路搜尋的常客，如果遇到不熟悉或冷門的專業領域，是否也可以準確的查詢適當的關鍵字呢？舉個例子來說：假設有位剛接觸網路搜尋的生手，想找尋有關美國職業籃球運動球員 Jordan 的資訊。因此他下了一個關鍵字：Jordan。從搜尋引擎回傳“Jordan”關鍵字的網頁中，可能包含美國 NBA 職業球員 Michael Jordan 的網頁；也可能包含中東國家約旦的相關網頁；還可能包含線性代數中的 Jordan Form；或者也有可能是其他姓名中出現 Jordan 的個人網頁。又如一個不懂得資訊科學的使用者，他想搜尋有關於資料探勘(Data Mining)領域中的關聯法則(Association Rule)的相關資訊。不過，他可能不知道此種關鍵字該怎麼給予，因此可能會輸入“data mining”這樣的關鍵字，但是在資料探勘的領域中，有相當多的子領域，因此，回傳網頁包含的內容也是包羅萬象的。

對於上列生手所面臨的問題，傳統的資訊擷取(Information Retrieval)領域發展出**自**

動查詢修正(Automatic Query Refinement)的技術。最常見的自動查詢修正技術就是**相關回饋**(Relevance Feedback)，由使用者對查詢的結果給予回饋。系統根據使用者的回饋資訊，自動產生新的查詢。也就是說，搜尋引擎運用相關回饋技術時，在每一回合的查詢修正中，搜尋引擎會傳回使用者其所搜尋到的網頁，經由使用者確認之後，若回傳的網頁並不滿足使用者的需求，使用者必須去點選回傳網頁中較符合自己資訊需求的網頁。根據這些由使用者給的回饋資訊，搜尋系統再找出與前一次回饋結果中相同資訊需求的網頁。如此繼續下去，直至使用者的資訊需求符合為止。

上述回饋方式屬於**明確回饋**(Explicit Feedback)，系統必須由使用者提供明確的回饋資訊。但在明確回饋的系統中，使用者必須付出額外的時間。一般而言，使用者並不喜歡做這額外的工作。尤其，使用者的原始查詢所得的結果，往往並不是那麼的符合需求。使用者在得到符合的答案前，也往往必須歷經多回的相關回饋以修正查詢。特別是針對較複雜的或是較冷門的資訊需求。

有鑒於此，如何避免明確回饋的缺點？如何自動化地偵測到使用者的真正資訊需求？如何自動化的去做查詢修正，則成重要的課題之一。

為了瞭解使用者的查詢需求，除了使用者的原始查詢與對查詢結果的明確回饋外，還有很多隱含的資訊可以作為瞭解需求的線索。例如：使用者曾經查詢過的關鍵字、使用者曾經點選的相關網頁、使用者瀏覽的情形和使用者與系統互動的紀錄都可以提供隱含的回饋資訊。這種透過隱含的資訊提供相關回饋功能的方法稱之為**隱含回饋**(Implicit Feedback)。

隱含回饋所利用的資訊可分為兩大類。第一大類資訊稱之為**短期情境**(Short-term Context)，指的是在目前使用者的 **Query Session**(查詢期間**)**中，有助於瞭解使用者資訊

需求的立即情境資訊。短期情境包括資訊需求的種類、先前的查詢、點選過的網頁等等。而 Query Session 指的則是使用者在搜尋引擎上，針對同一搜尋主題，由開始查詢一直到結束此次主題的搜尋，稱為一次的 Query Session。短期情境這類的資訊是與目前使用者的資訊需求最有直接關聯的。第二大類隱含回饋所利用的資訊則是**長期情境**(Long-term Context)，代表全部使用者的所有 Query Session 中，使用者與搜尋系統之間所有的互動歷史，包括查詢歷史、**點選連結**(Clickthrough)歷史等等。而此類資訊則是可應用到所有 Query Session 中。

以前述的舉例而言，如果搜尋引擎生手心目中想查詢的是線性代數中 Jordan Form 的相關資訊，那麼我們可以從過去的使用者搜尋紀錄中，比對找出過去與使用者的隱含回饋資訊相似的紀錄。例如，對於查詢的結果，他們都沒點選與 NBA 職籃或約旦相關的網頁，但他們都有點選與數學相關的網頁。如果由過去的紀錄中發現過去這些使用者最後都點選了有關鍵字“Jordan Form”的網頁。我們即可以將這樣的經驗法則套用在目前這位生手身上。因此，我們由這個例子中可以看出以往使用者的經驗對於查詢修正有相當大的幫助。

由於本研究想依藉著以往具有相似資訊需求的**經驗使用者**(Experienced Search User)之查詢行為來幫助目前使用者，因此，我們想根據以往研究的一些數據統計，來明瞭搜尋引擎生手與搜尋引擎經驗使用者的查詢過程各是如何的不同。[11]這一篇研究主要是針對經驗使用者與搜尋引擎生手，利用他們在網路搜尋上的一些數據統計分析，藉此來比較其查詢行為。其中提到，根據統計，搜尋引擎生手每次給予搜尋系統的查詢平均為 1.66 個關鍵字；而經驗使用者則是平均為 3.64 個關鍵字，而且越有經驗的使用者的一次查詢則是包含了越多關鍵字。另外研究也指出，有經驗的使用者會經常地使用進階的搜尋功能，例如使用 AND、OR、NOT 等運算子，使用片語搜尋等等。研究同樣也指出，搜尋引擎生手會很頻繁地變換所查詢的關鍵字，而且經常會變換查詢中部分的關鍵字，

表 1.1： 近幾年來針對使用者查詢資料的數據統計分析[28]。

Variables	1997	1999	2001
Mean terms per query	2.4	2.4	2.6
Terms per query			
1 term	26.3%	29.8%	26.9%
2 terms	31.5%	33.8%	30.5%
3+ terms	43.1%	36.4%	42.6%
Mean queries per user	2.5	1.9	2.3
Mean pages viewed per query	1.7	1.6	1.7
Pages viewed per query			
1 page	28.6%	42.7%	50.5%
2 pages	19.5%	21.2%	20.3%
3+ pages	51.9%	36.1%	29.2%
Users modifying queries	52.0%	39.6%	44.6%
Session size			
1 query	48.4%	20.8%	30.8%
2 queries	60.4%	19.8%	19.8%
3+ queries	55.4%	19.3%	25.3%
Boolean queries	5.0%	5.0%	10.0%
Terms not repeated in the data set	57.1%	61.6%	61.7%
Use of 100 most frequently occurring query terms	17.9%	19.3%	22.0%

且這些變換都是無助於萃取有用資訊的。在一次的搜尋過程中，搜尋引擎生手點選的網頁數目通常都很少，並且事後證明，這些由搜尋引擎生手點選的網頁大都不是相關的網頁。

[28]也是另外一篇針對使用者在網路上搜尋行為的專題研究，表 1.1 為近幾年來，針對一般使用者查詢資料的一些數據統計分析。

我們主要可以看到表 1.1 中，2001 年時，使用者每次的查詢包含的關鍵字個數平均為 2.6 個，而每一個使用者在一次查詢過程中的平均查詢次數為 2.3 次，每一次查詢所瀏覽的網頁平均為 1.7 個。

因此，本研究希望藉著隱含回饋的方式，由查詢日誌(Query Logs)中，利用過去具有相同資訊需求的使用者經驗，以幫助搜尋引擎生手有效地搜尋網頁，以達到自動查詢修正的目的。在下一章，我們將先回顧查詢修正隱含回饋的相關研究。第三章將介紹我們所提出的方法。接著，在第四章將敘述本系統的實作以及實驗評估方式與結果。最後一章則是結論與未來研究方向。