



105年特種考試地方政府公務人員考試試題

代號：32340

全一頁

等 別：三等考試

類 科：圖書資訊管理（選試英文）

科 目：資訊系統與資訊檢索

考試時間：2小時

座號：_____

※注意：(一)禁止使用電子計算器。

(二)不必抄題，作答時請將試題題號及答案依照順序寫在試卷上，於本試題上作答者，不予計分。

- 一、除了傳統的 Precision 及 Recall 之外，試舉出兩個適用於評估系統排序 (Ranking) 效能的新評估指標。(25 分)
- 二、Gerard Salton 提出以 “Discrimination Value” 來決定一個索引詞的重要性，並以 $TF (Term Frequency) * IDF (Inverse Document Frequency)$ 來計算詞的權重。試解釋 “Discrimination Value” 的意義及 $TF * IDF$ 的計算方式。(25 分)
- 三、相較於資訊檢索系統，推薦系統更適用於何種資訊取用 (Information Access) 情境？請從資訊需求及文件的特性進行申論之。(25 分)
- 四、試解釋相關反饋的原理，並說明準相關反饋 (Pseudo Relevance Feedback) 及隱式相關反饋 (Implicit Relevance Feedback) 兩種取徑的不同。(25 分)



申論題解答

一、
答：

Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
● 資訊系統評估新評估指標	● 資訊檢索 ● AP ● R-Precision	起(10%)：資訊檢索 承(40%)：資訊檢索評估傳統方法 轉(40%)：資訊檢索評估新判斷指標 合(10%)：小結
<p>參考書目</p> <ul style="list-style-type: none"> ● 陳光華 (2004)。資訊檢索的績效評估。2004 現代資訊組織與檢索研討會。 ● 相關判斷與評量，圖書館學與資訊科學大辭典，http://terms.naer.edu.tw/detail/1679017/ ● 資訊檢索系統評估與測試，圖書館學與資訊科學大辭典， http://terms.naer.edu.tw/detail/1680704/ 		
<p>資訊檢索可定義為有選擇性、有系統性回收經過邏輯性貯存的資訊。早在 1960 年代，便有學者們試著去評估資訊檢索系統的執行能力，也陸續發展出許多不同的測試形式，例如研究測試(Research Tests)是一種沒有特定的應用目標，只是想擴展對資訊檢索系統相關認知的測試方式；有的是在某個特定的作業環境下進行的測試，目的在於設立一套新的系統，故而往往產生許多改進的建議；另外，則有些是有關操作系統方面的測試。進行各種評估測試的理由，主要是在找出所謂成功檢索的意義層面，找出任何工具或技術的轉變對系統執行能力的影響，以及做為資訊提供者的參考，有效的評估在設立一個資料庫或操作檢索上，會是一種很好的投資或修正的參考。</p> <p>資訊檢索系統的評估，範圍的設計上可大到包含整個檢索環境，也可只針對某一部分來進行，例如針對索引部分、資訊選粹服務(SDI)的功能或回溯檢索的部分來執行。而在做評估測試時，選擇的參與者或使用最好是一群有真正資訊需求的對象，因為這樣的使用者才能了解自己需要的資料是什麼，也才能有效地分析檢索的滿意度。至於評估的重點，傳統上是建立在回現率(Recall Ratio)和精確度(Precision)的分析研究。所謂回現率是指所檢索到的相關文獻占潛在可檢索到的文獻總數之比例，而精確度是指檢索到的相關文獻占此次所檢索到的文獻總數之比例，回現率愈高，則精確度愈低。怎樣才算是有關的文獻，則端視檢索者根據自己的資訊需求所下的判斷了。所以，有真正資訊需求的使用者才能真實反映出問題的狀況，而文獻相關性的判斷，則是使用者尋求資訊與解決問題的重點。當然這也是評估中最能引起討論的部分畢竟相關與否是主觀的判斷，而非客觀的比較。</p> <p>然而在進行文件檢索的評分時，常採用的 recall(查全率)、precision(查準率)與結合兩者的 F1-Measure 是一種適用於無排序之檢索結果(non-ranked retrieved list)的評量尺度。</p>		



$$P = \text{Precision} = \frac{\# \text{ relevant retrieved}}{\# \text{ retrieved}}$$

$$R = \text{Recall} = \frac{\# \text{ relevant retrieved}}{\# \text{ relevant}}$$

$$F1 = \frac{2PR}{P + R}$$

然而，多數的資訊檢索系統的檢索結果都是排序的，這也符合使用者的期待，畢竟第 1 篇文件就是相關文件，與第 20 篇文件才是相關文件，對使用者而言，感覺是截然不同的。對於有排序檢索結果 (ranked retrieved list) 的評分尺度 (scoring metrics)，最常採用的是由 Buckley & Voorhee 提出的 average precision (AP) 與 R-Precision。

Average precision 是以下列方式計算而得，其意涵是平均每篇相關文件被檢索時的 Precision。

$$\text{Average Precision}_j = \text{AP}(j) = \frac{\sum_{i=1}^{r_j} \frac{i}{\# \text{Doc}_j(i)}}{r_j}$$

r_j 表示資訊檢索系統針對編號 j 的查詢問題，共檢索出的相關文件數。

$\# \text{Doc}_j(i)$ 表示資訊檢索系統針對編號 j 的查詢問題，在第 i 篇相關文件被檢索出時，總共被檢索出的文件數。

R-Precision 則是表示在檢索出第 R 篇文件時的 Precision， R 是查詢問題真正相關的文件數。以簡單的例子說明 P 、 R 、 AP 與 R-Precision。

某次查詢 $Q1$ ，資料集中 60 篇中應有 10 篇相關文章，但系統傳回的 15 篇文章中，只有 5 篇是相關文章，而排序從高到低的 15 篇文章中，+ 表示相關，- 表示不相關。

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
+	-	+	-	-	+	-	-	-	+	-	-	-	-	+

$$P = 5 / (5+10) = 0.33$$

$$R = 5 / (5+5) = 0.5$$

$$AP = (1/1+2/3+3/6+4/10+5/15) / 5 = 0.58$$

$$R = 4/10 = 0.4$$

在資訊爆炸的時代中，透過 Ranking 的排序方式優先顯示可提高使用者資訊檢索的效率。優化使用者經驗，除了要從資訊檢索系統的資訊架構著手，更要從資訊檢索排序效能加以優化。



二、
答：

Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
<ul style="list-style-type: none"> ● Discrimination Value ● TF ● IDF ● TF*IDF 		起(20%)：資訊檢索技術發展 承(50%)：詞頻，逆向文件頻率 轉(20%)：TF*IDF 與鑑別值之間的關係 合(10%)：優點

參考書目

- 逆向文件頻率，圖書館學與資訊科學大辭典，<http://terms.naer.edu.tw/detail/1679000/>
- 詞頻，圖書館學與資訊科學大辭典，<http://terms.naer.edu.tw/detail/1678993/>
- 黃雲龍 (2008)，中文全文文件群集索引理論研究與實證，圖書與資訊學刊，24，44-68。
- 楊德倫 (2014)，文字探勘之前處理與 TF-IDF 介紹，國立台灣大學計算機及資訊網路中心電子報，31。

資訊檢索技術的發展可分為以技術為導向的發展與以使用者為導向的發展模式，史派克瓊斯於 1972 年提出了逆向文件頻率 (Inverse Document Frequency, 簡稱 IDF)，並進行一連串實驗，發現使用 IDF 的檢索系統能夠產生比較有效的檢索結果。Salton 於 1973 年至 1975 年之間提出了數篇論文，進一步使用詞彙鑑別值 (Term Discrimination Value, 簡稱 TDV) 的觀念，加強資訊檢索系統的效用。

以下分別就詞頻、反向文件篇數與詞彙鑑別值 (或稱為索引詞區別值)

詞頻 (term frequency, 簡稱 TF) 是指文件中詞彙出現的頻率或是次數，是衡量一個詞彙重要性的一種指標。在各種資訊檢索模型中，如向量空間模型、機率模型、語言模型，都會用到詞頻的概念。

一般而言，一個詞彙在某一篇文章中出現的次數越高，即詞頻越高，則其在該篇文件中的代表性越重要。例外的情況，有虛詞、連接詞、代名詞等功能詞 (function words)，這些詞彙，經常有高詞頻，卻不帶有任何內容意義，在文件的詞彙處理過程，常被特意地停用、過濾掉，因而被稱為停用詞 (stop words)。

此外，詞彙的重要性，也會考慮到這個詞彙出現在所有文件中的篇數，亦即其文件篇數 (document frequency, 簡稱 DF)。若其出現在越多篇文件中，即 DF 越高，則表示該詞彙可能為常用字，故而重要性越低，相對於其反向文件篇數 (inverse document frequency, 簡稱 IDF) 也低。因此，詞彙的重要性，經常以詞頻 (TF) 以及反向文件篇數 (IDF) 這兩個概念，一起考慮，TF-IDF 是一種用於資訊檢索與文字探勘的常用加權技術，為一種統計方法，用來評估單詞對於文件的集合或詞庫中一份文件的重要程度

1. TF (Term Frequency)：

假設 j 是「某一特定文件」，i 是該文件中所使用單詞或單字的「其中一種」，n(i,j) 就是 i 在 j 當中的「出現次數」，那麼 tf(i,j) 的算法就是 $n(i,j) / (n(1,j)+n(2,j)+n(3,j)+...+n(i,j))$ 。例如第一篇文件中，被我們篩選出兩個重要名詞，分別為「健康」、「富有」，「健康」在該篇文件中出現 70 次，「富有」出現 30 次，那「健康」的 $tf = 70 / (70+30) = 70/100 = 0.7$ ，而「富有」的 $tf = 30 / (70+30) = 30/100$



= 0.3；在第二篇文件裡，同樣篩選出兩個名詞，分別為「健康」、「富有」，「健康」在該篇文件中出現 40 次，「富有」出現 60 次，那「健康」的 $tf = 40 / (40+60) = 40/100 = 0.4$ ，「富有」的 $tf = 60 / (40+60) = 60/100 = 0.6$ ， tf 值愈高，其單詞愈重要。

2. IDF (Inverse Document Frequency)：

換個角度來看，假設 D 是「所有的文件總數」， i 是網頁中所使用的單詞， $t(i)$ 是該單詞在所有文件總數中出現的「文件數」，那麼 $idf(i)$ 的算法就是 $\log (D/t(i)) = \log D - \log t(i)$ 。例如有 100 個網頁，「健康」出現在 10 個網頁當中，而「富有」出現在 100 個網頁當中，那麼「健康」的 $idf = \log (100/10) = \log 100 - \log 10 = 2 - 1 = 1$ ，而「富有」的 $idf = \log (100/100) = \log 100 - \log 100 = 2 - 2 = 0$ 。

所以，「健康」出現的機會小，與出現機會很大的「富有」比較起來，便顯得非常重要。

最後，將 $tf(i,j) * idf(i)$ (例如： $i =$ 「健康」一詞) 來進行計算，以某一特定文件內的高單詞頻率，乘上該單詞在文件總數中的低文件頻率，便可以產生 TF-IDF 權重值，且 TF-IDF 傾向於過濾掉常見的單詞，保留重要的單詞，如此一來，「富有」便不重要了。

傳統上，資訊檢索的研究通常使用詞頻 (Term Frequency, 簡稱 TF) 作為選擇索引詞彙的標準，認為排除所謂的功能詞彙 (Function Word) 之後，文件中出現越多次的詞彙越能夠代表該文件的特性。然而，若是相同的詞彙在許多文件都出現，則其代表性會比較不可靠，因為其鑑別性 (Discriminativity) 比較低。根據 Salton 實驗，詞彙的鑑別性、詞彙頻率與文件頻率之間有一些關係存在，隨著詞彙頻率與文件頻率的增加，詞彙鑑別性會從 0 到正值，然後逆轉到負值。詞彙的鑑別值與詞彙頻率之間的現象，提供深入於語言學構詞分析的思考方向，而文件頻率的增加也不全然代表內容訊息的增加，因此 Salton 建議，將頻率低的詞彙以詞組 (Phrase) 方式，改善系統的檢出率。然後將頻率高的詞彙以索引典的方式，改善系統的精確率。

Salton 提出文件的向量空間模型所建構起來的群集索引的優點就是減少傳統反轉索引方式的龐大索引空間，同時提升資訊檢索系統的效率與效能。

三、
答：

Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
<ul style="list-style-type: none"> ● 推薦系統 ● 從資訊需求及文件的特性 	<ul style="list-style-type: none"> ● 互動式資訊檢索技術 ● 資訊過濾 	起(10%)：互動資訊系統的產生背景 承(40%)：互動型資訊檢索技術的種類 轉(40%)：從資訊需求及文件特性申論 合(10%)：建議
參考書目 <ul style="list-style-type: none"> ● 卜小蝶，2006。「應用檢索記錄於網路術語推薦之研究」。Workshop of Recent Advances in Library and Information Science。美國資訊科學與技術學會台北分會。 		



隨著網路的普及與資源的快速成長，有愈來愈多使用者依賴搜尋引擎來查詢資訊。但由於網路資源數量龐大，往往輸入任一檢索詞彙，都會得到數以百萬計的檢索結果，造成使用者不少資訊負擔。再者，當使用者因某種資訊需求而產生資訊尋求行為時，其當下既有之知識狀態其實並不足以明確表達其需求，常常無法輸入適切的檢索詞彙。

為了克服上述資訊超載、認知負擔、及資訊不足等問題，已有不少互動式資訊檢索技術（Interactive Information Retrieval Techniques）的發展。這類技術主要是藉由使用者與系統間的互動來提昇檢索效益。相關技術包括相似性回饋（Relevance Feedback）、詞彙擴展（Query Expansion）、及術語推薦（Term Suggestion）等。相似性回饋的作法為系統將前一階段所檢索出的文件，提供使用者選取後，系統再由這些文件中抽取重要特徵進行檢索，以尋得更多相關文件。其中抽取比對的特徵如果是文件本身，則稱為相似性回饋功能；若為文件中的相關詞彙，則稱為術語推薦（或是相關詞提示）。而詞彙擴展其實與術語推薦的作法相似，只是前者為系統自動將相關詞彙進行擴展，後者則由使用者自行選取。過去有關相似性回饋技術應用，以文件回饋較為常見，但使用者必須判斷那些文件相關，一來造成認知負擔，二來由於回饋的文件包含資訊相當多元，回饋結果不一定符合檢索需求。換言之，術語推薦較不需額外資訊，同時對使用者而言，也較易判斷，因此是互動式資訊檢索中比較可行的作法。

從資訊取用時的資訊需求而言，使用者會從資訊物件本身的主題、個人推薦產生的個人興趣檔、瀏覽次數、下載次數、點擊次數作為推薦系統推薦時的判斷依據，除了系統紀錄檔產生的紀錄（ex: 瀏覽次數）以及透過資訊過濾系統更強調建立使用者對資訊需求與偏好的描述資料（profile），透過此描述資料，過濾系統才能自動地協助使用者篩選不相關與不值得注意的資訊。資訊過濾的方式基本上分為三種：基於內容特徵的過濾方式，基於個人認知與偏好的過濾方式，以及基於社會性、多人協同（social/collaborative filtering）的過濾方式。

若資訊文件內容為文字型資料居多，則需要透過關鍵詞自動抽取的方法進行。基本上關鍵詞的抽取方法約分三類：一是統計分析法，這也是資訊檢索領域中最常使用的方法，其主要是分析文件中詞彙的統計特性，如分析詞彙出現的頻率及位置，若符合預先設定的門檻值，即被視為關鍵詞。這類方法由於未使用任何控制字彙，因此較不受語文的限制，且所抽取的詞彙其新穎性較高，也可抽取到控制字彙未收錄的詞彙。但其缺點是所抽取詞彙常有錯誤組合或無意義的情形。第二種方法是詞庫比對法，即利用事先已建立的控制字彙表或詞庫，與文件來進行比對。這類方法雖然簡便，但詞庫無法收錄所有詞彙，同時也需人力維護，此外未知詞（Unknown Term）一般也不會被詞庫所收錄。第三種方法則是文法剖析法，即利用自然語言處理技術對文件進行文法剖析，所獲得的詞彙雖較有意義，但仍需詞庫的配合，此外，在文句的文法並不完整的情況下，這類方法的成效其實不太明顯。上述關鍵詞自動抽取技術基本上多以文件作為抽取來源，而所抽取出的詞彙，也多作為系統的索引詞彙依據。然而在實際的資訊檢索環境中，使用者所

使用的檢索詞彙經常與系統所使用的索引詞彙不一致。詞彙抽取後為了讓使用者能有效瀏覽所推薦之相關詞彙，系統需要進行概念的分群。概念的分群可用樹狀架構或是放射狀知識圖譜的方式展現，現在有許多資訊視覺化的技術可以補強這方面的呈現效果，有利於使用者利用界面上的功能選取系統推薦詞彙。

術語推薦能藉由提示與使用者查詢相關的詞彙，幫助其釐清問題及需求，以獲得更完整精確的檢索結果。



四、
答：

Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
<ul style="list-style-type: none"> ● 相關回饋原理 ● 準相關回饋 ● 隱式相關回饋 	<ul style="list-style-type: none"> ● Google Map 與 Google 查詢的相關回饋 	起(30%)：直接破題闡述原理 承(30%)：準相關回饋 轉(30%)：隱式相關回饋 合(10%)：小結

參考書目

- 相關回饋，圖書館學與資訊科學大辭典 <http://terms.naer.edu.tw/detail/1679018/>

相關回饋意指以初次檢索結果為基礎，透過使用者或是系統自動回饋額外的訊息，以利二次檢索。相關回饋的目的是為了進行二次檢索，由原始之查詢問句透過相關回饋產生修正之查詢問句，這個過程被稱為「查詢問句擴展」(query expansion)。因此，相關回饋通常僅是查詢問句的擴展的一種作法。

利用相關回饋在原始查詢問句中追加額外詞彙的作法，是查詢問句擴展常見的技術，且具有相當的效益。然而，對於大部分使用者而言，要提供相關回饋所需之額外詞彙並不容易，或是不願意花費額外的時間，勾選初次檢索結果中的相關文件，在這種情況下，經常採用準相關回饋(pseudo relevance feedback)。準相關回饋並非實際要求使用者回饋有用的資訊，而是利用初次的檢索結果，不經使用者判斷即假定所有文件(或是前 20 篇)皆為相關，再將這些假定的相關文件經由相關回饋的程序建構新的查詢問句，從而利用其做進一步的檢索。此方法有一明顯的缺點，若假定之相關文件清單中，實際上不相關的文件占大部分，那麼加入原始查詢問句的擴展詞彙與原檢索主題並不相關，則擴展後查詢問句的檢索品質會變差。

準相關回饋屬於明確回饋，系統必須由使用者提供明確的回饋資訊並付出額外的時間。因此，系統如何自動化偵測到使用者的真正資訊需求，例如使用者曾查詢過的關鍵字或是點選過的相關網頁，這種透過隱含的資訊提供相關回饋功能的方法稱之為隱式相關回饋。

隱式相關回饋可分為兩大類，第一大類為短期情境用，指的是在目前使用者使用的查詢期間中，有助於了解使用者資訊需求的立即情境資訊。第二大類所利用的資訊則是長期情境，代表全部使用者所有的查詢期間中，使用者與搜尋系統之間所有的互動歷史，包括查詢歷史或點選連結的歷史。

以 Google Map 為例，準相關回饋與隱式相關回饋的資料取徑方式就有所不同，Google Map 的準相關回饋會利用簡單的詢問要求使用者輸入問題的回答，而 Google Map 隱式相關回饋的方式則是透過記錄定位位置與行事曆結合，判斷與提供相關資訊給使用者參考。

雖然相關回饋能顯示驚人的效果，但是相關回饋的效益隨原始查詢問句、排序的公式及相關詞彙的數量、初次檢索結果品質而改變。許多研究指出加入太多擴展詞彙之後所導致的失敗；隨著文件資料庫的不同或是文件清單排序方式的不同，也會有不同的結果；對於利用相關回饋資訊進行的自動查詢問句擴展，新加入詞彙的數目亦是決定檢索效益的重要因素。