

106 年公務人員普通考試試題

代號：42660

全一頁

類科：圖書資訊管理

科目：資訊系統與資訊檢索概要

考試時間：1 小時 30 分

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題上，於本試題上作答者，不予計分。

(三)本科目除專門名詞或數理公式外，應使用本國文字作答。

一、國家圖書館全國圖書書目資訊網(NBINET)提供眾多國內圖書館的書目資訊，國立臺灣大學圖書館建置的 MetaCat 亦可檢索國內眾多圖書館的書目資訊。請說明這二種資訊服務系統的異同。(25 分)

二、一般的網路搜尋引擎送回的每一筆檢索結果都會附加一段簡短的文字(Snippet)，這段文字可能包含使用者的檢索詞彙，也可能沒有包含使用者的檢索詞彙，這段文字的目的是讓使用者可藉以簡易判斷這筆檢索結果是否為其所需，並進而決定是否點選(Click)該筆資料。請說明搜尋引擎如何產生 Snippet。(25 分)

三、眾多的資訊系統需要使用者的輸入作為後續處理的依據。當使用者輸入資料時，有可能輸入錯誤的文字(例如：拼錯英文辭彙 covfefe)，或是輸入贅字(例如：個人電電腦)。請說明資訊系統如何偵測可能的錯誤，並能夠友善性修正錯誤。(25 分)

四、資訊檢索系統效能的評量指標(Metrics)有多種，例如眾所周知的 Recall(回現率、查全率、求全率)與 Precision(精確度、查準率、求準率)，以及查詢時間等等。請盡可能列出你所知的評量指標，並逐一說明其定義與適用的時機或情況。(25 分)

1. 國家圖書館全國圖書書目資訊網(NBINET)提供眾多國內圖書館的書目資訊，國立臺灣大學圖書館建置的 MetaCat 亦可檢索國內眾多圖書館的書目資訊。請說明這二種資訊服務系統的異同。(25 分)

Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
<ul style="list-style-type: none"> ● NBINET 的集中式資訊檢索 ● MetaCat 的分散式檢索 ● 異同 	<ul style="list-style-type: none"> ● 整合式查詢 ● 分散式查詢 	起(10%)：整合檢索技術 承(30%)：NBIBET & MetaCat 簡介 轉(50%)：比較異同 合(10%)：metacat + 的更版
參考書目 1. NBINET (http://nbinet.ncl.edu.tw)		
<p>自從網際網路的廣泛應用與各式數位圖書館計畫的發展，各種數位原生與重製的資訊資源大量產生，促使使用者可以跨越時空，隨時隨地取用網路化的電子資訊資源。圖書館館藏發展除了傳統的實體館藏資源外，也逐一擴展至網路電子館藏資源。陳亞寧在館藏發展未來發展趨勢文章中指出，不論是圖書館的紙本存儲與管理，亦或是 Dempsey 所提出之館藏網格與群體式館藏，其中，整合檢索技術則是圖書館跨館查詢的一種最佳方案。</p> <p>全國圖書書目資訊網（National Bibliographic Information Network，簡稱 NBINet）聯合目錄系統於 1998 年 4 月正式啟用，主要功能在提供國內各圖書館合作編目之用，並可供一般使用者查詢利用聯合目錄資料庫。主要發展目標為 1.建立完整的我國「國家書目」書目資料庫。2.建立全國中外文圖書資料聯合目錄資料庫。3.提供國內外新書資訊。4.提供學術研究及一般參考查詢利用。5.提供館際合作與互借。6.促進書目資訊著錄之標準化。7.提供各圖書館發展之全文影像及索引摘要服務。8.進行與國際間書目資料庫之聯繫與利用。MetaCat 為台灣大學圖書館自行開發的跨資料庫整合查詢系統，2017 年升級為 MetaCatPlus。</p>		
	NBINET	MetaCat
系統架構	集中式結構	分散式結構
收錄資料	國家圖書館及 86 所各類型合作單位陸續提供之書目及館藏資料、國際標準書號中心新書書目、漢學研究中心藏書目錄、民國 1-38 年參考書目、港澳地區參考書目。	整合了 73 所圖書館，北部 41 所，中部 10 所，南部 22 所，東部 1 所。以類型區分來說，大專院校圖書館 62 所，其他圖書館 12 所的圖書館館藏目錄。
功能	1. 提供合作館進行線上合作編目 2. 一般使用者可使用瀏覽器連線查詢並下載或轉錄多種格式資料 3. 提供符合 Z39.50 協定之跨系統查詢功能，連結其他具有相同功能的資料庫進行查詢	1. 整合國內多所重要圖書館之館藏查詢系統。 2. 可即時得知所需資料的借閱狀況，但查詢速度會受網路及各圖書館系統快慢之影響。 3. 將各圖書館之 WebPAC 系統按館藏類型及地區分類。 4. 可自行選擇多個圖書館整合查

		詢。 5. 查詢結果依題名條列，並列出擁有此館藏之圖書館。
書目處理	將各合作館的書目資料進行選擇書目與匯集館藏的處理	會進行查詢資料簡目的書目整合。(去重複)
會員加入政策	1. 圖書館加入需有其義務與責任	無
合作編目政策	1. 有規範建檔格式、編目規則、分類法與主題/標題詞	無
是否提供館藏資訊	有各館索書號，但無館藏借閱現況與預約狀態	因為直接連結到各圖書館系統的館藏詳目，因此可查看該書的現況。
主要差異	1. 功能不同：NBINET 為國家書目中心，所收錄的書目資料為集中式處理，書目並經過品質控管。MetaCat 為跨圖書館自動化系統的整合查詢，主要功能是進行查詢。 2. 應用技術不同：整合 vs 分散 3. 書目品質：控管 vs 不控管 4. 檢索速度：MetaCat 為分散式檢索，會因圖書館自動化與網路速度而不穩定的現象。	
如前所述，MetaCat 2017 年升級為 MetaCat Plus，原因在於讀者希望能除了檢索書目之外，更能夠即時獲知館藏的獲得情形，以利之後利用館際互借平台申請館際互借。雖然分散式檢索有其缺點，但就使用者檢索資源的面相來說卻是舉足輕重。		

2. 一般的網路搜尋引擎送回的每一筆檢索結果都會附加一段簡短的文字(Snippet)，這段文字可能包含使用者的檢索詞彙，也可能沒有包含使用者的檢索詞彙，這段文字的目的是讓使用者可藉以簡易判斷這筆檢索結果是否為其所需，並進而決定是否點選(Click)該筆資料。請說明搜尋引擎如何產生 Snippet。(25 分)

Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
<ul style="list-style-type: none"> ● Snippet(或稱摘錄) ● 如何產生 	<ul style="list-style-type: none"> ● 複合式摘要 	<p>起(20%)：網路資源檢索的資訊爆炸 承(30%)：摘錄定義與功能，複合式摘要 轉(30%)：產生摘錄的方式與建立的依據 合(20%)：Google 已將摘要進化成複合式摘要</p>
<p>參考書目</p> <ul style="list-style-type: none"> ● 什麼是 Rich Snippets (複合式摘要)? https://seo.dns.com.tw/archives/4119 		
<p>資訊爆炸的時代，後現代重視的是片斷而非整體，同時亦經常是非線性模式思考，例如：以往要從文抄本中擷取一段文字，必須逐頁翻閱，現今藉由超文本（hypertext）技術即可輕鬆查得，因此，網路資源檢索中的自動摘要可協助使用者判別重要性的技術更顯重要。</p> <p>通常使用者產生檢索問題後檢索系統送回排序的檢索結果，每一筆檢索結果通常包含簡易的摘錄（Snippet）。摘錄是從文獻中選取句子構成摘錄，或是對文獻句子重組產生摘錄。摘錄產生方式主要為自動產生主題簡要文句、自動選擇重要句子、關鍵詞多寡、依句子位置、依據特數據行等。</p> <p>通常建立摘錄的方法如下</p> <ol style="list-style-type: none"> 1. 語法分析：利用辭典中的語言知識對文獻中的句子進行語法分析，獲得句子結構。 2. 語義分析：運用知識庫中的語義資料將句子結構轉換成具邏輯和意義為基礎的語義表示。 3. 資訊提取：根據知識庫中的知識進行推理，提取關鍵內容出來。 4. 摘錄生成：將資訊表中的內容轉換成完整連貫的文字輸出。 <p>而摘錄建立的依據</p> <ol style="list-style-type: none"> 1. 詞頻：文章相關主題的有效詞 2. 標題：作者提示文章內容的標題 3. 位置：論述題目主旨通常在段落首句 4. 句子結構：句法和句子之間存在某種關係，因此疑問句、感嘆句等不宜進入摘錄 <p>現在，Google 將摘錄的型式優化，成為 Rich Snippets (複合式摘要)。所謂的複合式摘要，就是在指自然搜尋結果列表中，Google 為了讓搜尋者能夠更清楚的瞭解列表到底是什麼網頁，而顯示比原先更多資訊的區塊。例如：查哈利波特時直接在下方的摘錄欄位顯示書評。Google 會在某些情況下決定何時顯示一般的 Snippets，何時顯示 Rich Snippets，而這關鍵就是網頁上需加入正確的標記，就能夠顯示。想要在自然搜尋列表中顯示 Rich Snippets (複合式摘要)，就必須正確的使用複合式摘要標記，例如微資料 (Microformats)、微格式 (Microdata) 以及 RDFa (RDF-attribute)等等。Rich Snippets (複合式摘要片段)就是指透過特別的標示而讓搜尋引擎抓出來的資料段，這些特別的標示就是如 Schema.org 在推動的</p>		

Microdata 之類的標準。而圖書館界之資料也可以用於網路搜尋引擎也可使用相同的標準，讓圖書館內的資源更能夠曝光於搜尋平台上。

3 眾多的資訊系統需要使用者的輸入作為後續處理的依據。當使用者輸入資料時，有可能輸入錯誤的文字(例如：拼錯英文辭彙 covfefe)，或是輸入贅字(例如：個人電電腦)。請說明資訊系統如何偵測可能的錯誤，並能夠友善性修正錯誤。(25 分)

Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
<ul style="list-style-type: none"> ● 輸入錯誤導致的錯誤修正 	<ul style="list-style-type: none"> ● 模糊查詢 ● Spelling suggestion/ do you mean 功能 ● 向量模式查詢 ● 機率查詢 	<p>起(20%)：以探索系統平台的 did you mean 功能談起</p> <p>承(30%)：模糊查詢的定義</p> <p>轉(40%)：系統如何偵測錯誤</p> <p>合(10%)：協助使用者友善修正錯誤</p>

參考書目

- 曾元顯、林瑜一著(1998)。模糊搜尋、相關詞提示與相關詞回饋在 OPAC 系統中的成效評估。中國圖書館學會會報，61，103-125。
- 圖書館學與資訊科學大辭典。<http://terms.naer.edu.tw/detail/1679031/>
- 卜小蝶(2006)，應用檢索紀錄於網路術語推薦之研究。

我們在使用圖書館所提供的探索資源系統平台時，會發現到在輸入檢索字串時，若系統查詢不到所輸入的檢索詞彙時，系統會利用拼字檢查或是您是否要找尋 (Did you mean) 功能指引讀者下正確的檢索用語。或是，在資訊系統中輸入李登飛，系統則在檢索結果中找出許多與李登飛相似的檢索結果，其中包括李登輝正確的標目以及筆數。這兩種不同的情境所使用的檢索技術不同，但目的都是相同的，都是指引讀者進行友善修正錯誤的方式。前者主要技術為互動式資訊檢索技術中的術語推薦服務，後者則是資訊檢索中的模糊檢索或是相似性回饋。

互動式資訊檢索功能可分為三類技術，相似性回饋、術語推薦及詞彙擴展。相似性回饋做法為系統將前一階段所檢索出的文件，提供使用者選取後，系統在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，抽筆比對的特徵若是文件本身，則稱為相似系回饋功能，若為文件中的相關詞彙，則稱為術語推薦(或是相關詞提示)。而詞彙擴展與術語推薦作法相似，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，但使用者必須判斷哪些文件相關，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。術語推薦則不需額外資訊，同時對使用者來說也較容易判斷。一般而言，術語推薦與查詢問句擴展的技術不外乎是 1. 以查詢問句為基礎。2. 以語料庫為基礎。3. 以語言分析特性為基礎。

模糊檢索比較像是互動式查詢功能中的相似性回饋，模糊檢索 (fuzzy search) 是一種資訊檢索演算法，顧名思義是針對資料庫中的文件內容進行近似字串比對 (approximate string matching)，從而搜尋出與使用者需求相近的資料。傳統上，以精確比對 (exactly matching) 為導向的搜尋演算法允許使用者透過查詢詞彙找出資料庫中完全符合的文件，但無法找出字詞相近的資料。而模糊檢索則具備類似自然語言或近似字串查詢的功能，藉由演算法計算詞彙與詞彙之間的相似度，能找出字串相近、字義相近或字音相近的資料，可使資訊檢

索系統具備容錯、同音、相關詞提示、相關詞回饋、相關程度排序等功能，進而協助使用者擴展查詢範圍。

系統如何偵測錯誤

資訊檢索系統的任務是比對檢索條件與資料庫文件，以找出符合檢索條件的相關文件。傳統作法大多採用布林邏輯進行比對，將文件庫中的待檢文件與使用者輸入的檢索詞進行布林運算，結果只有 1 或 0（亦即符合或不符合），符合者檢索出來，不符者略過不取，所以又稱為精確比對。其優點是簡單、快速，非常適合用於需求明確的檢索。但是，如果使用者無法明確的表達檢索需求，或是輸入的檢索詞出現錯字、冗字，或是查詢條件比較複雜以致無法清楚描述，精確檢索就顯得有所不足。由於模糊檢索可允許使用者輸入任意查詢字串，即使有同音、同義、錯字、冗字也能檢索出相關資料，因此又稱為近似字串檢索或是容錯檢索。

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。同時由文件本身及使用者檢索記錄兩種來源抽取相關詞彙，兼顧文件特性及使用者導向需求，以提高相關詞彙抽取的精確率及回收率。

目前，許多搜尋引擎皆支援模糊檢索功能，使用上方便許多。以 Google 搜尋引擎為例，假如使用者在輸入時拼錯字，模糊檢索程式將藉由計算詞彙的相關性而主動向使用者提示正確的字詞。例如：在 Google 輸入「磨湖搜尋」將得到「目前顯示的是以下字詞的搜尋結果：模糊搜尋」。由於模糊檢索日益普及，許多圖書館檢索系統也都同時支援精確檢索與模糊檢索功能，通常當使用者選擇精確檢索找不到資料或資料過少時、或是不確定檢索詞是否正確時、或想藉由單一檢索詞擴展查詢範圍時，便可使用模糊檢索來找出相關性資料。

4 資訊檢索系統效能的評量指標(Metrics)有多種，例如眾所周知的 Recall(回現率、查全率、求全率)與 Precision(精確度、查準率、求準率)，以及查詢時間等等。請盡可能列出你所知的評量指標，並逐一說明其定義與適用的時機或情況。(25 分)

Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
<ul style="list-style-type: none"> ● 資訊系統評估新評估指標 	<ul style="list-style-type: none"> ● 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

參考書目

- 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/>

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。

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者最好是找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。怎樣才算是有相關的文獻，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。

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。

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

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

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

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。