106年公務人員高等考試三級考試試題　　代號：23470　　全一頁

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：㈠禁止使用電子計算器。
㈡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。
㈢本科目除專門名詞或數理公式外，應使用本國文字作答。

一、試說明大數據（Big Data，或稱巨量資料或是海量資料）的應用如何提升整合式圖書館管理系統（Integrated Library Management System，簡稱 ILMS）的使用者體驗。（25分）

二、一般而言，停用詞可分為二類：通用停用詞（Generic Stop Words），專用停用詞（Specific Stop Words，或稱領域停用詞，Domain Stop Words）。請分別說明這二類的停用詞，並說明如何建構這二種停用詞表（Stop-word List）。（25分）

三、相關回饋的二種主要形式：一為顯性相關回饋（Explicit Relevance Feedback）；一為隱性相關回饋（Implicit Relevance Feedback）。請申論隱性相關回饋的可能形式。（25分）

四、資訊檢索的發展已有數十年的歷史，有眾多學者與專家提出許多資訊檢索理論或模式。請任選二種，首先明確寫出其理論或模式名稱，接著說明其內容並申論優缺點。（25分）

申論題解答

1. 請說明大數據 (Big Data，或稱巨量資料或是海量資料) 的應用如何提升整合式圖書館管理系統 (Integrated Library Management System，簡稱 ILMS) 的使用者體驗。(25 分)

Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
● 大數據 ● 整合式圖書館管理系統 ● 使用者體驗 (UX)	● 大數據的 4V 特性 ● 鏈結資料(Linked Data) ● 混搭 (Mashup) ● 資訊視覺化 ● 資料探勘 ● 圖書館大數據的來源 ● 大數據於圖書館系統的應用	起(20%)：大數據興起的背景與圖書資訊系統的演變 承(30%)： 使用者體驗，ILMS 系統的功能與特性 轉(40%)：系統功能應用大數據改善使用者體驗 合(10%)：學術圖書館趨勢

參考書目

- 黃鴻珠，陳亞寧 (2010)。圖書資訊系統的演變與發展，教育資訊與圖書館，48 (4)，頁 403-428。
- 蔡天怡 (2015) ，巨量資料於圖書館的發展應用趨勢，104 年開放資料與巨量資料之應用與發展趨勢研習班研習手冊。

近來大數據以顯然成為顯學，分析此趨勢愈來愈熱門的原因為 1. 容易收集跟處理資料。隨著儲存資料的硬體設備愈加便宜、可以儲存的容量愈來愈大，再加上各種用於分析、處理資料的軟體相對普及，故透過資料分析，將分析結果應用於開發相關專案的機會也愈來愈多。2.拜網路便利所賜，容易有大量的網路使用者出現。3.政府推動開放資料，讓資料集（dataset）的取用更容易。4. 各種感測裝置（sensor）的普及，例如：錄音工具、手機感測軟體、相機錄影設備，讓大眾暴露在容易被感測、記錄相關資料的環境中。而圖書館，知識匯集地，在圖書館是一個成長的有機體下，如何隨著外在環境趨勢，化危機為轉機，就必須了解何謂大數據? 圖書館的核心系統—整合式圖書館管理系統提供了不少數據來源，對圖書館可如何透過數據分析改善使用者體驗的服務。

大數據指的是因應過去技術平台無法處理大量、快速產生、無結構性或需要即時回應的資料所衍生的新一代技術的集合。並不是數據量要大，要海量，才能有價值。

大數據的特性有 4V 特性，分別是大、快、雜、疑。分述如下：

第一是大量（Volume)：

伴隨電子商務的蓬勃、資料傳輸管道成本的降低，現今的資訊技術可以處理的資料愈來愈大量。如果處理的資料量規模達到數兆位元組（terabytes, TB）的資料量規模，即稱為大量。

第二是持續快速產生（Velocity)：

指每一秒鐘會即時地產生數十萬筆的紀錄檔，形成所謂的串流資料（Streams Data)，這種資料的特性是寫入的速度非常快速，會源源不絕不斷地寫入資料庫中。例如在 Twitter 或 Facebook 上每秒的發文、每秒在搜尋引擎中搜尋產生的紀錄檔。

第三是多樣性（Variety)：

有兩種層面的意思，一種為資料領域的多樣性，當處理資料時，將完全不同領域的資料一起合併來做分析。例如：麵包店除了看原本麵包相關的報表，現在也可以把其他領域的資料共同納入考量，例如氣象、交通資料等等放進來一起分析，將異質性資料的結合做成本分析。另一種是資料格式上的多樣性，可以大致區分為結構性資料與非結構性資料，前者是能夠被量化、容易組織的資料，像是書目紀錄或稽核項，而後者則像是 Facebook 的發文、照片、通話紀錄或是影像……等較難處理。

第四為數據需經過真實性驗證 (Veracity) :
大數據分析中英分析並過濾資料有有偏差、偽造、異常的部分，防止這些「dirty data」損害到資料系統的完整跟正確性，進而影響決策。

圖書館大數據的主要來源可分為四種，分別是公部門或其他私部門的資料(ex: 其他圖書館的公開資料，政府公開資料等)；其次是圖書館各種裝置產生的資料 (ex: 監視器、門禁安全系統)；使用者自行產生的資料 (ex: 社群媒體的使用紀錄))；組織內部資料 (ex: 圖書館自動化資料、網站流量、參考服務使用統計等)。其中，整合式圖書館管理系統中流通模組、編目模組、採購模組、期刊模組與 webopac 模組中提供了讓圖書館可進行進一步數據分析的來源，其中，以流通交易檔為分析依據的資料挖掘則是明顯改善使用者體驗的一利器。

透過讀者借閱紀錄數據分析技術，圖書館可進行以下服務以提升使用者體驗

1. 應用借閱紀錄關聯規則分析，找出讀者個人特性與圖書之間的關聯性；利用讀者特性的相似性推薦圖書；將同質性的圖書推薦給適性的讀者；這是最多人使用的探勘技術。
2. 分類分析：藉由讀者不同的特性與借閱紀錄，判別讀者間的相似性與相異性，找出各類特性的讀者對圖書的興趣，並依此模式推薦新書給讀者。
3. 群集分析：找出圖書與圖書、讀者與讀者之間的關係，以探討使用者的群集特性，並找出其借閱行為的傾向。
4. 次序相關分析：依據讀者借閱的順序，來推薦給其他未借閱之讀者。

圖書館應用大數據的優點影響層面廣，並可協助館員問對的問題、行銷決策。方法包括整合不同來源，利用混搭的方式展現全方位資料，以提供加值服務與探索性服務。館員也可以透過數據分析進行服務評估。ACRL 2016 年學術圖書館趨勢即充分看到資料的價值，從提供服務到發展政策，到館員成為資料科學家，如何協助資料庚用。但，圖書館在應用時仍需要考量的挑戰有資料的取得成本、資料的透明度與公開程度、資料的隱私權與安全性、資料分析的基礎設施與技術設備到最後的資料解讀的專業知識與應用進行全方位的考量。

2. 一般而言，停用詞可分為二類：通用停用詞 (Generic Stop Words) ，專用停用詞 (Specific Stop Words，或稱領域停用詞，Domain Stop Words)。請分別說明這二類停用詞，並說明如何建構這二類停用詞表 (Stop-word List) （25分）

Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
● 停用詞 ● 專用停用詞 ● 領域停用詞 ● 如何建構	● 魯恩	起(20%)：自然語言檢索 承(30%)：停用詞，通用停用詞，領域停用詞 轉(40%)：建構方法 合(10%)：資訊檢索先驅- 魯恩

參考書目

- 化柏林 (2007)，知識抽取中的停用詞處理技術，現代圖書館情報技術， 8。
- 陳光華 (2006) ，知識的組織與擷取，圖書館學刊，12。
-

網際網路的出現使得資訊檢索研究更具挑戰性的環境，資訊檢索是由文件集合中檢索相關的資料，而資訊擷取是由文件中事先擷取預設所需的資訊，視為比資訊檢索更深一層的資訊服務，因資訊擷取不僅僅辨識重要的個體，還必須決定個體之間的關係。而今，透過資訊擷取自動化技術，系統會文件版面進行分析模組、分詞、語彙分析、語法分析、語義分析，而這其中，分詞系統擷取出來的詞彙，為了讓系統自動分詞之詞彙更具資訊檢索之價值，事先透過停用自詞表篩選過濾停用字，不僅僅能加快分詞速度，更能針對之後知識擷取所需要處理的語彙分析、語法分析與語義分析提高分析的速度與品質。

停用詞，在不同任務導向的檢索系統，有其不同的意義。在基於詞的資訊檢索系統中，是指出現頻率高、不具檢索意義的詞。例如：的、是、太、of、the 等。若是該系統為自動分類系統，停用詞則為沒有意義的虛詞與中性詞。在自動問答系統中，停用詞可能是針對問題的不同有不同的動態變化詞。停用詞可分為通用停用詞與專用停用詞，通用停用詞為不限系統內的文件領域，系統在進行分詞時一律過濾，像是人類語言中包含的功能詞，這些功能詞極其普遍，與其他詞相比，功能詞沒有什麼實際含義，比如'the'、'is'、'at'、'which'、'on'等。但領域相關的停用詞進一步將代表一般概念而非學科領域中較具專指意義之詞彙加入停用字表。例如，科技研究報告資料庫中會將[本研究]進行過濾。

停用詞表的來源有人工建構與基於統計的自動學習兩種方式。
通常建構停用詞表的方法有以下方式

1. 詞頻：將系統文件中的高詞頻的詞彙找出，刪除可用於檢索之詞彙後則為停用詞。
2. 文件頻率：如果一個詞在文檔中出現的頻率太高的詞彙。
3. 語法剔除：將動詞的現在進行式、未來式、過去式與動詞等磁性的詞視為停用詞。Ex: running, runner, runs, ran 。
4. 包含數字以及特殊符號的詞

漢斯· 彼得· 盧恩為資訊檢索的先驅，創造了以概念為檢索的理論基礎。在魯恩之後，停用詞的觀念就逐漸出現在資訊檢索文獻中。

3 相關回饋的二種主要形式：一為顯性回饋 (Explicit-Relevance Feedback)；一為隱性相關回饋 (Implicit Relevance Feedback)。請申論隱性相關回饋的可能形式。(25 分)

Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
● 相關回饋 ● 隱性相關回饋 ● 顯性相關回饋 ● 可能形式	● 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。

4. 資訊檢索的發展已有數十年的歷史，有眾多學者與專家提出許多資訊檢索理論或模式。請任選兩種，首先明確寫出其理論或模式名稱，接著說明其內容並申論優缺點。(25 分)

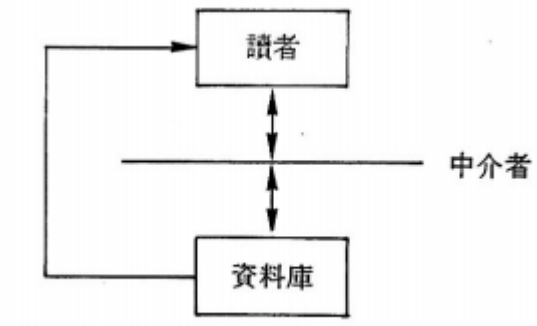
Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
● 資訊檢索理論 ● 內容及優缺點	● Ellis ● Kuhlthau	起(20%)： 資訊檢索系統的要素與資訊檢索理論 承(20%)：資訊檢索研究理論的範疇 轉(40%)： 挑出兩點理論進行內容說明與申論優缺點 合(20%)：資訊系統理論為實驗性科學

參考書目

- 圖書館學與資訊科學大辭典 http://terms.naer.edu.tw/detail/1679001/
- 吳美美 (1994)，試論資訊檢索理論。當代圖書館事業論集。台北市，正中書局。
- 王秀卿 (2001)，網路使用與資訊尋求行為之文獻探討。大學圖書館 6(1) ，頁 144-162。
- 吳美美 (2001)，中文資訊檢索系統使用研究。臺北市，臺灣學生。

資訊檢索理論在探討從各種型態的文件中，分析、轉換、提取、過濾出有用的訊息，並加以排序、組織、呈現，再以主動、被動或互動等方式，來滿足使用者資訊需求的各種議題。傳統資訊檢索的理論，主要探討圖書館館藏中書目資訊的儲存與提取。在電腦網路普及、數位文件風行之後，舉凡網頁資料查找、學術文件搜尋、法律前案檢索、新聞事件歸類、垃圾郵件過濾、文件自動摘要、關聯資訊擷取、生物資訊探勘、主題趨勢辨識、商業智慧分析、自動詢答系統等，都成為資訊檢索理論探討的課題。

吳美美認為資訊檢索系統的組成三要素，分別是讀者、中介者、資料庫。圖示如下。



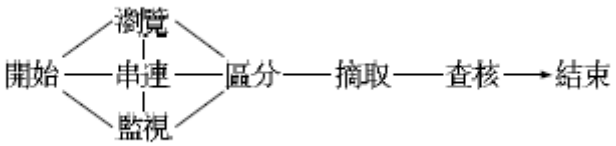
從上述模式所包含的三要素來看，資訊檢索研究的範疇至少包括五個要向，除了三個要素的研究- 讀者研究、中介者研究、資料庫研究外，由於檢索系統有互動的本質，以及檢索系統研究的目的在在設計，或是改進設計，應再加上互動研究和設計。

資訊檢索系統模式中，使用者被界定為有某種資訊需求的人。使用者研究有三個課題，資訊需求 (information need) 是使用者研究的第一個議題，資訊尋求行為 (information seeking behavior) 是第二個議題，資訊問題表示法是第三個議題。資訊需求是資訊尋求行為的前導，人有資訊的需要才會有接著而來資訊尋求行為。資訊尋求行為是人有了資訊需求，要向外求。此時內在的認知活動，是疑問的產生，而顯現在外的行動就是資訊尋求行為。資訊需求就是使用者的資訊問題，也有人認為是使用者的檢索問題。

資訊需求相關理論要論述的有 T.D. Wilson 於 1981, 1996, 1999 年分別提出的資訊行為模式，為文本理論之基礎。Ellis

及 Kuhlthau 再研究對象資訊行為過程中所經歷的資訊行為階段與特色。因此本題列舉 1. Ellis 的資訊尋求行為策略模組與 Kuhlthau 資訊尋求行為階段模式進行論述。

Ellis 認為資訊尋求行為是由八個特色所組成，分別是 1. 開始 (starting): 使用者開始尋求資訊時使用的方法，如詢問其他使用者。2. 串連 (Chaining): 查看文獻中的附註和摘要，並進一步串聯已知款目。3. 瀏覽 (Browsing)：以半導向，辦架構的方式找尋資訊。4. 監視 (monitering): 保持最新穎資訊的檢索。5. 區分 (differentiating)：區分資訊來源，過濾所得資訊。6. 摘取 (extracting)：在資訊來源中選擇相關的資訊。 7. 查核 (verifying): 核對資訊的正確性。 8. 結束 (ending)：結束檢索。Ellis 認為，任何人的資訊尋求模式之特色間的互動，和個人所處時間點之資訊尋求環境有絕對關係。



此理論的優點在於 Eills 明確提出完整的資訊搜尋模式，並提出使用者的情境理論。

Kuhlthau 的資訊搜尋過程模式可謂資訊行為研究在認知取向的典範，也是自 1990 年代後最常被引用的相關模式之一。人們在建構資訊需求的過程中，通常是經歷情感、思想及行動上一系列明顯變化的階段，ISP 從使用者資訊尋求過程中的全面性經驗，提出一個自開始（initiation）、選擇（selection）、 探索（exploration）、形成（formulation）、收集（collection）至呈現（presentation）等六個階段的模式，從使用者的認知觀點分析其資訊搜尋的過程。ISP 認為人們的資訊搜尋過程是一種建構的過程，過程中的每個階段皆整合了個人的情意（A-ective）、認知（Cognitive）和行動（Physical）之全部經驗。

此理論的優點為 Kuhlthau 應用個人建構理論以描述個人如何建構它們遇到的資訊，該理論之最基本的假設為使用者因為資訊需求不確定性所引起的疑惑和挫折，隨著檢索過程的推展，使用者收集到愈來愈多相關資訊，使這些負面的感覺轉變成有信心，很滿意和有方向感的正面感覺。Kuhlthau 認為使用者之檢索問題會逐漸被修正，而資訊尋求的過程是不斷修正的過程，因此在分析檢索行為時，多採用連續檢索過程之分析。

由於這類技術、系統與理論的發展，在滿足使用者需求或特定的應用目的，沒有標準答案可供遵循，常須實驗加以比較驗證，因此實驗所需之文件測試集的蒐集或製作，在資訊檢索領域的研究當中，扮演非常重要的角色，而正確可靠的成效評估方法，也是不可忽視的議題。可以說，資訊檢索是一種實驗性科學。而資訊檢索理論涵蓋的議題，在圖書資訊學的領域中，除了跟圖書館學、使用者行為、資訊計量等課題有長久的關係外，也和電腦科學中的資料庫、自然語言處理、機器學習、文件自動處理等研究議題，有相當高的相互影響性。