109 年公務人員普等考試四級考試試題
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

1. 試分別說明內容為基礎之多媒體檢索(content-based multimedia retrieval) 如何利用影像 (image)，聲音(audio)及視訊(video)的內容特徵，進行多媒體檢索? 相較於利用文字 (Text) 描述多媒體資訊，並以文字間接進行多媒體檢索的方式，直接以內容特徵進行多媒體檢索的優缺點為何? (25 分)

Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
● 以內容為基礎之多媒體檢索 ● 以文字為基礎之多媒體檢索 ● 優缺點	● 人工智慧	起(20%)：多媒體與資訊檢索的定義 承 20(%)： 三種多媒體的內容特徵為何 轉(40%)： 相較於文字的多媒體檢索的比較 合(10%)： 帶出人工智慧的機器學習

參考書目

● 曾元顯 (1996)，多媒體資訊檢索技術之探討。21 世紀資訊科學與技術的展望國際研討會。

多媒體泛指多種媒體型態的融合，包括文字(text)、圖像(image)、影像(video)、語音(speech)、音樂(music)、動畫(animation)等等。資訊檢索是有關於資訊館藏的儲存、組織與查詢的資訊服務。由於影像、聲音、視訊等媒體比傳統文字媒體更能有效傳達資訊，以及目前資訊技術已具備處理多媒體資料的能力，隨著多媒體資料的快速累積，繼傳統文字資料檢索之後，多媒體資訊檢索將成為資訊服務的重要課題。

以內容為基礎的多媒體資料檢索 (Content-based Multimedia Retrieval) 就是以多媒體可備電腦擷取的視聽感知之實體特徵進行索引與檢索，就影像而言，內容特徵可以是顏色、紋理、形狀、空間與佈局。就聲音而言，內容特徵可以是音符頻率、節奏、語調等。就視訊而言，內容特徵可以是動作、圖像與聲音。以圖像檢索為例，圖像檢索按描述圖像內容方式的不同可以分為兩類，一類是基於文本的圖像檢索(TBIR, Text Based Image Retrieval)，另一類是基於內容的圖像檢索(CBIR, Content Based Image Retrieval)。以下分別針對這兩種檢索的方法進行比較。

比較	TBIR	CBIR
時間	較早出現，約 1970 年代	1992 年
定義	利用文本標註的方式對圖像中的內容進行描述，從而為每幅圖像形成描述這幅圖像內容的關鍵字，比如圖像中的物體、場景等	利用電腦對圖像進行分析，建立圖像特徵向量描述並存入圖像特徵庫。
方法	人工標註方式或圖像識別技術進行半自動標注	當使用者輸入一張查詢圖像時，用相同的特徵提取方法提取查詢圖像的特徵得到查詢向量，然後在某種相似性度量準則下計算查詢向量到特徵庫中各個特徵的相似性大小，最後按相似性大小進行排序並順序輸出對應的圖片。
使用者檢索	根據自己的興趣提供查詢關鍵字	● 使用者以圖找圖 (Query by example) ● 直接檢索 (Direct query): 例如要找顏色 ● 相關回饋: 選擇相關的例子進行檢索
優點	易於實現，且在標注時有人工介入，所以其查準率也相對較高	● 將圖像內容的表達和相似性度量交給電腦進行自動的處理，克服了採用文本進行圖像檢索所面臨的缺陷， ● 發揮了計算機長於計算的優勢，大大提高了檢索的效

		率
缺點	● 需要人工介入標註過程，使得它只適用於小規模的圖像資料 ● 對於需要精確的查詢，使用者有時很難用簡短的關鍵字來描述出自己真正想要獲取的圖像 ● 人工標註過程不可避免的會受到標註者的認知水準、言語使用以及主觀判斷等的影響，因此會造成文字描述圖片的差異。	● 為特徵標註與語義之間存在著難以填補的語義鴻溝(semantic gap) ，並且這種語義鴻溝是不可消除的。
應用層面	● 傳統的圖像資訊檢索系統	● 基於內容的圖像檢索技術在電子商務、版權保護、醫療診斷、街景地圖等工業領域具有廣闊的應用前景。

英文中有言：A picture is worth a thousand words.（一圖勝千言），對描述多媒體資料來說，反過來說也對：A word is worth a thousand pictures （一個關鍵字彙讓人聯想到上千張圖片）。除非讀者展示出他想要的圖片、樣式，否則光從語言文字的描述，一位館員實在很難想像讀者真正需要的圖片或影像。目前人工智慧中有許多圖像辨識的應用就是利用內容檢索的技術，由機器可自動辨認出有意義的內容特徵並不斷的學習，也是目前人工智慧第三波浪潮興起的原因之一。

2. Word2Vec 是一種 Word Embedding 的技術，請說明 Word2Vec 的技術內涵為何？相較於 TF * IDF 所決定的詞向量，採用 Word2Vec 所決定詞向量於向量空間模型 (Vector Space Model) 所設計的資訊檢索系統，其優缺點分別為何？(25 分)

Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
● 詞嵌入 ● Word2Vec 技術內涵 ● 與 TF*IDF 的優缺點	● 自然語言處理 ● 人工智慧 ● 詞袋模型 ● 機器學習	起(20%)：自然語言處理中的詞袋模型 承(20%)： 詞袋模型的缺點 轉(30%)： Word Embedding 與 Word2Vec 的技術內涵 合(30%)： 優缺點綜合論述

參考書目

- 林頌堅 (2017)，以開放資料的教師學術專長彙整表為基礎之學科標準分類分析。教育資料與圖書館學，54(1)。
- 林昆賢、蔡俊明(2019)。基於深度學習的自然語言處理中預訓練 Word2Vec 模型的研究。國教新知，66(1)，15-31。
- 圖書館與資訊科學大辭典 https://terms.naer.edu.tw/detail/1678997/?index=1

自然語言處理，是針對人類語言文字進行各種自動化處理的技術，其目標是要讓電腦認識、分析、理解、合成人類語言，進行各式運算，希望最終能以自然語言為媒介，讓電腦跟人類順暢的溝通，以完成各項指定的任務。目前自然語言的資訊檢索的處理方法大多基於詞袋模型，比對文字資料上出現詞語種類與次數的相似性。而系統內的詞袋中的詞，則是藉由詞頻 (TF) 與逆向文件頻率 (IDF) 組成之詞，模型中將詞語轉換為一組向量，向量上的每個元素的數值代表詞語在文字上的權重。文字資料的比對便轉成為向量間的相似性比對。

然而，依 TF*IDF 所決定的詞向量之詞袋模型有幾個嚴重的缺點，簡述如下：

(一) 詞語的出現都是獨立的，缺乏詞語在文本上下文的脈絡。

(二) 無考慮多義詞及同義詞，每個詞語都對應向量上的一個特定元素，缺乏語法和語意。

(三) 詞語種類多，導致向量的維度 (Dimension) 相當大，但向量上的數值相當稀疏。

(四) 自動斷詞時無考慮到未知詞。

人工智慧的第三波翻轉的關鍵技術是深度學習，利用深度學習進行語意分析、自然語言處理、自動分類等相關應用時，事先的準備訓練詞嵌入(word embedding) 是必要的。詞嵌入的技術就是將在將中文字的最小處理單位「詞或詞組」轉換為電腦能處理的向量型態，作為文字輸入銜接深度學習其他程序的重要步驟。Word2Vec 為學者 2013 年提出之詞向量的訓練模型，該模型使用二種不同的非監督式學習的神經網路模型訓練方式：(一) CBOW、(二) Skip-gram，訓練結果可獲得帶有豐富訊息的詞向量，供後續深度學習詞嵌入使用。以下簡述這兩種不同的訓練模型：

(一) CBOW（Continuous bag of words）

CBOW 模型是輸入去除中心詞的其他上下文（context），訓練模型去預測該中心詞出現機率的方法。

(二) Skip-gram

Skip-gram 模型與 CBOW 模型同樣是以預測詞為訓練的目的，但二者用來訓練的輸入詞及輸出詞恰好相反，Skip-gram 模型輸入一個詞，以該詞為中心去預測其上下文。

優點是 Word2Vec 的方法具有隱含上下文脈絡的關係，轉換出來的特徵向量在空間向量上的計算較能夠保留每個詞語的語法或語意訊息。另一個優點則是 Word2Vec 所產生的特徵向量維度通常遠小於詞袋模型的特徵向量維度，在計算速度上明顯的會比 TF*IDF 所決定詞向量於向量空間模型 (Vector Space Model) 所設計的資訊檢索系統效能較佳。

Word2Vec 固然解決了新詞或斷詞系統不完善引起的缺失，但也產生了一些過度匹配的問題。另外，因為詞與向量是一對一的關係，多義詞的問題在此方式中也無法獲得解決。透過預先訓練之詞向量且賦予文字上下文間的關聯資訊之特徵值，可應用於不同的自然語言深度學習任務中，並具有提升模型學習成效之泛用性，

3. 現今網路上充斥著許多形形色色的資源，怎樣選擇，評估這些資源已成為重要課題。試分別說明何謂資訊的正確性 (Accuracy)，權威性 (Authority) ，客觀性 (Objectivity)，涵蓋性 (Coverage) 及時效性 (Currency)？如何從這五個面向來判斷檢索資訊的可信度? (25 分)

Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
● 網路資訊可信度 ● 正確性 ● 權威控 ● 客觀性 ● 涵蓋性 ● 時效性	● 網路素養 ● IFLA 如何識別假新聞	破(20%)： 網路素養破題 論(20%)： 可信度定義 結(60%)： 五面向說明與判斷網路資訊

參考書目

● 陳世娟、邵婉卿（2014）。臺灣民眾網路素養之調查研究。大學圖書館，18(1)

21 世紀是以網際網路和數位資訊為主流的世紀，使用網路已經成為現代人生 活中不可或缺的一部份，因此培養良好的網路素養十分重要。網路素養，是指具備在網路環境中取得優勢的必要條件，是有能力藉由網路獲得 資訊、服務及資源，即能將網路資訊和其他資訊相結合，增加網路資訊的價值，能使用網路資訊的服務，用來分析並解決工作或個人相關事物，以提高生活品質。

如果要能有效處理和利用網路上取得的資訊，就要進行判斷，因此網路資訊品質的評估是網路素養中極重要的一環。自從網路資訊品質開始受到關注，複雜的網路資訊以及網路內容品質的可信度也開始成為熱門的研究議題。Rieh 與 Danielson 定義「可信度」是指可以相信的程度，資訊來源是可以被信任的、值得相信的、或可依賴的，那麼這個資訊就是可信靠的， 雖然對象和來源可以用來提供線索，使用者最終將因為他們個人的經驗和知識做出不同的知識品質評估。

文獻中共通的網路資訊品質評估指標分別是資訊的正確性 (Accuracy)，權威性 (Authority) ，客觀性 (Objectivity)，涵蓋性 (Coverage) 及時效性 (Currency)。以下分別說明如何從五個面向判斷檢索資訊的可信度：

(一) 正確性：正確性是指資訊的事實性、詳細程度、精確與完整性。可透過網站的內容豐富且正確、網站標題與內容一致

(二) 權威性：是否有註明著者與資料來源。網站內的資訊會註明資料來源、註明聯繫網頁設計者的資訊及網站基本資料、可了解發布網站的目的與動機。亦可在網路上思尋著者的名字，間接證實著者的存在與是否可信。

(三) 客觀性：網站內資訊發布的目的客觀且具有價值。需要特別注意網路資訊是否太過戲劇化或具有煽動性，不合理或錯誤的描述。

(四) 涵蓋性：網站資訊能夠滿足使用者需求、能讓使用者方便搜尋。

(五) 時效性：網站所提供的資訊內容是很新穎的、網站的資訊很新且更新迅速。網站有標註發布時間或修改時間。

以上至少有四個評估標準 (除了權威性) 可以從資訊內容本身就可以找到。另外，除了這五項之外，IFLA 也提供了八種如何識別假訊息的方式讓圖書館員或讀者透過這八種方式判斷資訊的可信度。這八種方式分別是 1 考慮資訊來源。

2. 詳細閱讀。3. 查核作者。4. 資料來源。5. 檢查發布日期。6. 這是個笑話嗎? 7. 摒除偏見。8. 向專家請教。其實這八種方法跟五個面向息息相關。在現今資訊爆炸的時代，如何透過素養的能力判斷資訊的可信度，這些評估指標提供了讀者或館員一種查核方法，同時也可以將這些評估放法融入於中心學資訊素養的教育課程設計的參考。

4. 試說明搜尋引擎提供動態查詢語句建議 (Dynamic Query Suggestions) 的目的為何？要達成有助於使用者的動態查詢語句建議，在技術上需要考量哪些面向？提供這樣的服務對於搜尋引擎效能的主要挑戰為何？ (25 分)

Step 1：拆解題幹	Step 2：概念延伸	Step 3：重組配分
● 動態查詢語句建議定義與目的 ● 技術面向 ● 主要挑戰	● 人工編制詞表 ● 統計分析之共現索引詞表 ● 應用使用者檢索紀錄	破(%)： 定義與目的 論(%)： 三個技術面向方法與優缺點 結(%)： 主要挑戰

參考書目

- 圖書館學與資訊科學大辭典 https://terms.naer.edu.tw/detail/1678988/?index=15

資訊檢索五大面向之一為透過查詢界面的功能強化資訊檢索的功能，而目前在搜尋引擎提供動態查詢語句建議 (Dynamic Query Suggestions) ，在圖書館學與資訊科學大辭典中又可稱之為自動語彙推薦服務（automatic term suggestion）、關聯詞提示或相關查詢詞建議，屬於互動式資訊檢索（interactive information retrieval）技術的之一。主要是藉由使用者與檢索系統之間的互動機制，提示使用者與原查詢字詞相關的關聯詞詞組建議，以輔助使用者明確地定義及表達其資訊查找的需求，使其查詢的建構能符合檢索系統的索引語言和檢索規範，進而獲得完整而正確的檢索結果。目的不僅僅可以在使用者進行檢索時建議與其原檢索主題或需求相關的同義詞及相關詞，亦能協助使用者在查詢建構時更加明確地定義和表達其資訊檢索的需求。隨著資訊時代的來臨，未來資訊超載、認知負擔，以及資訊不足的問題勢必將日益嚴重，類似的互動式資訊檢索技術發展已是必然的趨勢。

自動語彙推薦服務的實現，通常利用下列的技術來達成：

(一) 人工編製推薦詞組：人工編製推薦詞組由專家建立控制詞彙，例如：詞庫或索引典。此方法的優點是精準度高，且通常定義詞彙的同義或從屬(上／下位)關係，因而能完整地呈現出主題概念的架構；但是，其缺點在於人工製作的成本高、耗時，沒有辦法快速進行推薦詞組的維護和更新，且不易於跨領域的環境下使用。

(二) 統計分析之共現索引詞表：統計分析之共現索引詞表則利用關鍵字詞在文件資料庫中共同出現（co-occurrence）的頻率統計自動建構詞彙關聯，或是運用共現技術（collocation）萃取相關的組合詞組，例如：複合字詞（compound）或片語（phrase）。其好處是共現索引詞表的建構快速且成本低；然而，壞處是準確度略低，且詞彙關聯僅是主題概念相關，並非主題層級的語意關聯。

(三) 應用使用者檢索紀錄：應用使用者檢索紀錄則是蒐集所有使用者曾經檢索過的關鍵字詞，透過統計分析抽取相關詞彙，以達到建立關聯詞詞組的目的；這種方法最大的優點是以真實的使用者資訊需求為導向，因而推薦的關聯詞詞組比較符合使用者對於資訊檢索需求的認知，但是因其關聯詞詞組衍生自使用者所輸入的檢索需求，使得其建議的相關詞與原查詢關鍵字詞間關聯性較差，不易協助使用者釐清其真正的資訊檢索需求。

在搜尋引擎中，動態查詢語句建議的應用可分為「即時提示」及「靜態推薦」兩大類。即時提示乃是在使用者輸入檢索關鍵字詞的同時，透過即時查找的技術與後端的系統溝通取得相關的關聯詞詞組，並且藉由下拉式選單的方式在使用者端即時呈現。其好處是當使用者尚未完整輸入檢索關鍵字詞時，系統便可以預測使用者的檢索需求，使用者只需從建議詞組中選擇適當的關鍵字詞即可進行檢索。而靜態推薦則是在檢索結果中動態萃取與使用者查詢關鍵字詞有關的關聯詞詞組，但僅在檢索結果頁面的上方或下方呈現，建議使用者可以透過這些相關詞來進一步縮小範圍查詢。提供這樣的服務對搜尋引擎的主要挑戰為為了要即時顯示查詢語句，搜尋引擎的四大子系統中之一網頁索引（indexing）就必須特別注意斷詞的合理性以及在網頁檢索（querying）時如何判斷哪些語句跟使用者的查詢語句是相關的予以建議。另外，如何在檢索框即時顯示的效能也是挑戰之一。