

99 學年度大四工工專題摘要

第十二組	建構論文摘要關鍵內容分析與自動擷取模式
指導教授	侯建良 教授
參與學生	9634011 沈君霓 9634027 陳昶昇 9634050 劉書豪
摘 要	
近年來由於網際網路技術之盛行，資訊技術之普遍運用，各企業與機構之電子化文件不斷快速累積，如何自廣泛資訊中取得使用者真正需要之資訊，同時節省大量資訊過濾與篩選時間，實為現今資訊服務與知識管理之重要課題。以往資訊需求者利用搜尋引擎以關鍵字為條件搜尋文件，而搜尋引擎取得相關文件後乃擷取文件中含搜尋條件的前後文作為文件段落，供資訊需求者參考以判斷文件傳達之資訊是否符合其需求，進而作為是否繼續進行深度閱讀的基礎。然而，此種以「搜尋條件」為基礎的文件段落呈現方式時常無法清楚讓資訊需求者了解該文件之真正核心內容。因此，本研究針對表達文件關鍵意義的關鍵字於文件中的位置進行探索，期能揭露文件關鍵字於文件中分佈位置的特性，並發展一套「文件關鍵內容分佈特性擷取」方法論，利用此方法論將此些關鍵字分佈特性建立一「關鍵內容擷取對照表」；之後，本研究再以「關鍵內容擷取對照表」為基礎發展一套「決定文件關鍵內容」方法論，並將此方法論之結果應用於文件搜尋，使搜尋引擎能以關鍵內容作為文件提示段落，並提供更具代表性與合理性之參考資訊予資訊需求者。 為達上述之目的，本研究首先乃廣泛地收集文件，以統計方法與分群方法探討文件關鍵字於文件中之分佈位置的特性，並依此發展一套「文件關鍵內容分佈特性擷取」方法論。此方法論之詳細作法乃先依關鍵字位於文件中的百分比位置進行分群並計算各分群之質心位置及全距；之後，再以不同濃縮率為基礎計算各分群之不同信賴區間並推斷所有分群之信賴區間所對應之信賴水準，將分析結果以「關鍵內容擷取對照表」予以維護。之後，本研究再以此「關鍵內容擷取對照表」為基礎，發展一套「決定文件關鍵內容」之方法論。而此方法論之詳細作法乃先取得目標文件之內容與對應之資訊，再利用「關鍵內容擷取對照表」內的資料進行比對，以取得目標文件對應之關鍵範圍落點；最後，利用關鍵範圍落點擷取文件之關鍵範圍內容，期能在無需預建詞庫的前提下，取得文件中可能包含關鍵字之簡短關鍵內容，並推斷其對應之信賴水準。 本研究除發展「文件關鍵內容分佈特性擷取」及「決定文件關鍵內容」方法論外，尚以此兩方法論為基礎，利用文件關鍵資訊呈現之概念建置一套「文件搜尋系統」，此系統乃利用關鍵內容作為文件提示段落，以更具代表性地呈現文件搜尋結果。本研究並以全國碩博士論文為案例進行系統績效驗證，以確認方法論之可行性。整體而論，本研究所提出之「文件關鍵內容分佈特性擷取」及「決定文件關鍵內容方法論」可有效擷取文件之真正核心內容提供予資訊需求者，以協助資訊需求者判斷文件傳達之關鍵資訊是否符合其需求，進而更有效率地決定其真正需要之文件，以進行深度瀏覽。	