

106 學年度大四工工專題摘要

第 8 組	以文件特徵值為基礎之文章風格評斷模式
指導教授	侯建良 教授
參與學生	103034006 高漢佑 103034023 李松霖 103034026 呂冠昇
摘 要	
隨著資訊科技的普及，我們可以輕易地從網路上的文件獲取大量的資訊，然而在撰寫及散佈文章時，衍生出許多問題，如：文件抄襲、冒用他人身份發佈文章。在文件抄襲方面，由於電腦方便的文書處理能力，每個人都能輕易地複製貼上並修改他人的著作。而這樣的問題大從網路上之公開文件，小至學生的心得與報告，都有可能發生。而冒用他人身份的問題則是衍生自部分社群媒體帳號安全的不完善或個人資料的維護問題。為了解決以上問題，我們想出了一套分析文章風格的方法與系統，來比較兩兩文章之間的風格相似度，希望藉此協助使用者進行文章風格之判定。而本研究的目的除了協助我們找出可能來自同一作者的文章外，也希望使用者能透過一次輸入大量文章進行比較，並由系統產生分析結果來讓使用者能在較短的時間內找出文章風格相似的文章。本研究除了分析現有的網路文章，還能提供使用者自行加入任意文章進行分析。 本專題於第一階段從文章中取出特徵向量，包含文章用詞與文章句長等向量。在詞的部分，我們利用 tf-idf 演算法來計算出每個詞在每篇文章中的權重，並建立詞頻矩陣。第二階段則是將不同文章間的特徵向量分別作餘弦分析，計算兩兩文章在不同特徵下的相似值，並建構數學模型來計算其綜合評比之相似值。第三階段則是將這些關係值利用 K-means 分群法分成三個群分別代表相似度高、中、低，並將結果呈現給使用者。使用者能查看任意一篇文章與其他文章之間的文章風格相似度。	