以資料採礦技術建置查核缺失編碼網路資訊系統

吳丞曜[1] 何岳峰[2] 黃濬彥[3] 謝孟勳[4]

[1] 中興大學土木工程學系碩士生，uiyyeu@gmail.com
[2] 中興大學土木工程學系博士生，hoamon@gmail.com
[3] 中興大學土木工程學系博士生，johnisacoolboy@gmail.com
[4] 中興大學土木工程學系副教授，mchsie@gmail.com

摘要：公共工程案進行查核（督導）時，查核委員所撰寫的『品質缺失描述』，其編號工作目前仍採人工判讀的方式，此法非常曠日廢時，而且編碼的工作非常容易產生錯誤。故本研究之目的為期望針對查核委員的『品質缺失描述』，拆解成電腦能判讀的『特徵』，並於 765 條缺失編號中，搜尋最相關的缺失編碼。

本研究在特徵方面採用斷詞處理方式來擷取特徵，並以「Yahoo!搜尋『斷章取義』 API」所提供的服務來取代詞庫的建立，獲得『1000』個關鍵詞。

相關函數方面，本研究採 TF*IDF(term frequency–inverse document frequency)技術。為取得缺失內容與缺失編碼之『相似度』，本研究分析 Data-Mining 技術中，4 種較常見的距離公式：餘弦法、歐幾里德、曼哈頓以及內積法，作為計算相似度的依據。

最後的實驗結果顯示，以歐幾里德距離計算的成果有 82.72%的機率，可幫使用者大幅縮減工作量；同時，在工作量減輕的情況下，可預期使用者判讀編號的品質將提升，有益於往後進行的各種統計分析。本研究可於稽查、督導、查核等相關系統中，以公共工程委員會頒定之缺失編號作廠商缺失改善紀錄之統計，可有效辨別缺失紀錄的編號屬性，提高缺失統計正確率。

關鍵字：文件分類，工程查核，缺失編號