使用決策樹來抽取文件自動分類系統中之分類規則 Extracting Classification Rules in Automatic Document Classification Systems by Using Decision Trees

自從Maron於1961年提出首篇文件自動分類的論文以來，傳統的分類方法不外乎機率模式與向量模式。近年來的研究也加入了統計分析、專家系統、自然語言處理、和類神經網路等先進的技術，以提高分類的正確性。以上所提的諸方法中，其對文件自動分類而言，均可視為是黑箱作業，因其分類行為或分類規則無從得知。本研究利用機械學習技術中之Quinlan的C4.5決策樹(Decision trees)來抽取文件自動分類系統中之分類規則，期使文件自動分類系統之分類行為透明化，而人們可藉由所抽取之分類規則進一步來驗證文件自動分類之正確性。在本研究中，我們採用ACM Computing Reviews的分類法作為分類的依據。我們從該期刊共收錄了56個中類別，6424篇論文為實驗用資料。再以其中的論文題目和出處當作該文件的素描(Profile)。取其中十分之一作為測試資料，其餘為訓練資料。我們從訓練資料中，使用 Quinlan的決策樹共抽取出1162條分類規則。再利用此分類規則分別對訓練文件及測試文件做分類，實驗結果分別為：訓練資料召回率為67.7%，測試資料為 45.5%。若將上述規則再精簡成 29O條分類規則，則訓練資料召回率變為52.3%，而測試資料略降為 43.0%。