一個處理概念漂移的垃圾郵件分類演算法

垃圾郵件氾濫的問題一直沒有得到徹底的解決，各種垃圾郵件防治機制紛紛興起，其中以機器學習為主的垃圾郵件分類過濾最為盛行，而這些方法，主要都是基於所有的資料在固定不變的環境下之假設，但是在實際環境中，郵件內容會隨著概念的漂移而不斷變動，使得分類器在模型建立之初，都有不錯的分類效果，但隨著時間的演進與概念的漂移，郵件的分類正確率會逐漸下滑，因此必須有一個學習的機制，針對資料集中新進與舊有郵件做相關的學習與調整。另一個郵件分類的問題是資料的偏斜，由於垃圾郵件的氾濫，垃圾郵件個數通常明顯的比正常郵件來的多，在分類的過程中，雖然大類別都有著較高的召回率，但是小類別的召回率卻相對不佳，本研究提出一個演算法IFWB，以貝氏分類為基礎，採用IGICF擷取關鍵字，結合一個漸進遺忘機制與分類成本架構，來解決郵件分類中概念漂移與資料偏斜的問題，並且透過實驗來驗證本研究所提出的郵件分類方法。