以資料庫集合運算發展的字串比對系統

應鳴雄,林健宇,鄧光宏
資訊管理學系
資訊學院
mhying@chu.edu.tw

摘要

近年來學術界的論文、作業抄襲事件頻傳，文件比對技術的研究成為重要的議題。過去的文件比對研究，主要是運用統計、向量、矩陣及移動位置等方式進行，但只要抄襲者在字串間加入贅詞或將句子進行部分修飾後，抄襲比對系統通常無法正確的比對出抄襲的段落，因而助長了學生抄襲的投機心態。本研究利用中文斷詞(Chinese Word Segmentation)及資料庫集合運算(Database Set Operation)為基礎，建構一個字串比對系統，以解決贅詞過多及詞彙順序問題。本研究利用中文斷詞方式將字串斷成許多詞彙，並使用資料庫中已建立好的集合運算式進行交集和合併運算來正確比對出字串間相同之處。使用資料庫集合運算式來做比對，會比過去將大量文字放入程式記憶體中還來得有效率。本研究在雛型系統效能驗證方面，與過去字串比對方式進行比較，並以效率及準確率做為績效指標，而結果顯示本研究雛型系統準確率方面能有更好的績效。

關鍵字：中文斷詞、資料庫集合運算、比對系統、字串比對