語意分析暨文件分類技術

本院覽號

05T-1010224

公告日期

智財權狀態

know-how

摘要

傳統分類方法通常利用字典比對、自然語言剖析、詞頻統計等方式取出關鍵字作為文件的特徵值，作為分類演算法的前處理。我們所發展的分類演算法則進一步納入時間因素與流行語等概念。 這項新的技術定期運用RSS機制，由特定來源網站匯入即時文章，再以自然語言擷取等傳統擷取文件特徵值的方式檢出語意關鍵詞，並統計分析其發生頻率與生命週期。同時，我們所發展的自我學習機制所訓練出的處理核心，已能有效處理時下部落格常用的口語用詞及熱門關鍵詞，例如：林來瘋、iphone4S等；亦能有效處理新聞媒體等網站使用的專業用詞。 我們利用pixnet公司收集的大量的部落格等網路文章測試，證實本演算法呈現令人滿意的分類準確率。

技術優勢

本文件分類演算法具有自我學習機制，可定期自網路上提供RSS之來源網站匯入即時文章並自我訓練，此訓練出的核心將能有效處理 時下部落格、口語用詞及熱門關鍵詞，例如:林來瘋、iphone4S。至於一般新聞等標準用字、用詞，也已經內含。所以，本演算法在分類準確率上，能有效處理含有口語用詞、熱門議題之文章。更適合，部落格等文章。 另一優點為效率，因本分類演算法在設計之時，已對效率做一處理，不同以往使用機器學習之分類演算法，需要較多時間於訓練及測試階段。本演算法已經過測試，達到每小時能處理超過10,000篇文章之效能。

應用範圍

可用於網路文件自動分類

創作人

何建明等

中央研究院
ACADEMIA SINICA