壹、緒論

早期的中文古典詩詞選編之編纂體例大多以作家詩詞選編或以歷代詩詞選編。但時代在發展、社會在進步，人們的閱讀能力、欣賞習慣及品鑒水平也在變化（張占國、王鐵柱，2006）。後來便出現了根據詩詞本身內容所表達的思想、情感來進行分類的體例。

目前的詩詞選編都是由編者以人工方式進行分類，當面對大量的中文古典詩詞需要進行分類時，人工分類的速度將非常緩慢，且多疏漏；所幸以英文爲主的文件分類技術已發展多年，本研究希望能利用這些技術來應用於中文古典詩詞的分類。

自動化文件分類系統，是近代學術研究中的一個熱門題目，相關應用的研究也是層出不窮，但少見應用於中文古典詩詞分類的研究。本研究希望能建立一套中文古典詩詞自動分類系統，來檢驗各種自動分類方法、參數配置等的分類績效。

貳、文獻探討

機器並不像人類，本身能自我理解文件的內容及大綱，僅能透過存在於文件集中數據來分析並建立分類規則。目前最常使用的方法爲向量空間模式（Vector Space Model）（古倫維，2000；杜海倫，1999；汪若文，2004；林頌華，1999；高志強，2004；莊慧美，2000；許志全，2009；陳昭安，2002；曾元顯，2002；黃冠中，2007；楊林波、王士同，2009）。向量空間模式一般定義爲將文件中之字、詞或片語視爲特徵，再將特徵資訊轉換成向量，建立出特徵向量空間，再於特徵向量空間中進行向量間相似度之計算。使用向量空間模式來建構分類規則，其處理方法大致可分爲四個步驟，分別爲前處理（preprocessing）、特徵萃取（feature extraction）、相似度計算（similarity calculation）及分類方法（classification method）。

一、前處理

爲了使自動文件分類的過程能更有效率、分類結果更正確，通常會在特徵萃取之前對文件內容做一些處理，以去除雜訊（noise），令萃取出來的特徵能更爲精準地代表該文件，此一步驟通稱之爲「前處理」。以下是常見的前處理方法：

（一）刪除停用字（removing stop-word）

在英文或中文文件中常有很多主詞、冠詞及介詞，還有常出現的慣用字及標點符號，這一類字雖有其存在的意義，但出現在文件中的頻率常過高，簡單來說，即是所有文章都共有的字，這些字稱之爲「停用字」（stop-word），而這些文字的集合稱爲「停用字列表」（stop-word list）。由於本研究是以中文古典詩詞爲分類主體，字數本就不多，因此不再做停用字處理，僅將標點符號刪除。

（二）中文斷詞（word segmentation）

英文的詞均以空白區隔，因此並無斷詞問題；然而中文並無詞與詞之間的間隔，因此需先做斷詞的處理，才能取出正確的特徵。目前中文斷詞器的製作大致以中央研究院資訊科學研究所發展的斷詞規則爲依據（江振宇，2004），然中文古典詩詞的字數較少，大部分規則不適用，故本研究只採用其中的第一個規則即長詞優先。

中文斷詞的方法主要是將文章中出現的字跟詞庫做比對，本研究採用YAHOO! 奇摩字典，因其爲電子化詞庫且可在網路上免費取得。一般中文文章中，四字詞以上的比例已相當少（余清祥，1998），詩詞中則更少，以五言絕句爲例，一句只有五個字，爲了豐富詩意，作者不太可能使用四字詞。因此本研究只斷出三字詞以下的特徵。

二、特徵萃取

特徵萃取是從前處理後的文件內容中，以詞爲特徵主體，透過各式計算建構出屬於該文件的特徵向量。應用在文件分類上的特徵粹取方式常見有下述幾種：

（一）詞頻（Term Frequency, TF）

定義爲文件 d 中出現詞 t 的次數。表示爲：

$$TF(d, t) \qquad \text{（公式 1）}$$

（二）逆文件頻（Inverse Document Frequency, IDF）

定義爲訓練集中，含有詞 t 的全部文件數之倒數，上述文件數愈少，特徵愈強烈，爲了公平比較，通常以下列公式計算：