國立交通大學

資訊科學系

碩 士 論 文

中文自動作文修辭評分系統設計

Automated Chinese Essay Scoring System Rhetoric Aspect

研 究 生：張佑銘

指導教授：李嘉晃　教授

中 華 民 國 九 十 四 年 六 月

中文自動作文修辭評分系統設計

Automated Chinese Essay Scoring System Rhetoric Aspect

研 究 生：張佑銘　　Student：Yu-Ming Chang

指導教授：李嘉晃　　Advisor：Chia-Hoang Lee

國 立 交 通 大 學

資 訊 科 學 系

碩 士 論 文

A Thesis

Submitted to Department of Computer and Information Science

College of Electrical Engineering and Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

In

Computer and Information Science

June 2005

Hsinchu, Taiwan, Republic of China

中華民國九十四年六月

中文自動修辭評分系統設計

學生 ： 張佑銘　　指導教授：李嘉晃 博士

國立交通大學電機資訊學院　資訊科學研究所碩士班

中文摘要

本論文探討是否能夠以機器學習的方式來實行中文自動作文評分以協助作文的評分工作，此項評分限於修辭方面。主要基礎在於設計及擷取修辭技法來做為機器學習所需要的特徵，這些特徵包含詞數、形容詞數、成語數、譬喻手法、非口語化的喻詞、排比手法。由於不同老師批改同一篇作文也會有誤差出現，所以我們認為系統判斷出的分數及作文實際修辭分數的差距在一分(含)以內為可容許的誤差，因此以「避免產生過大誤差，但容許微小的差異」為主要概念建立改良式的 ID3 演算法做為機器學習的方式。訓練和測試作文都由老師先行評分，每篇作文的修辭分數由低到高為一到六分。利用改良式的 ID3 演算法產生決策樹，決策樹會將測試的作文分類到一分到六分其中的一個等級。以上述特徵及機器學習方式建立的修辭評分系統，其評量出的分數與真實分數的誤差比兩名沒有受過標準化作文評分訓練的老師間之誤差來得低。

Automated Chinese Essay Scoring System Rhetoric Aspect

Student ： Yu-Ming Chang　　Advisor：Prof. Chia-Hoang Lee

Department of Computer and Information Science
National Chiao Tung University

Abstract

This thesis explores the possibility of designing an Automated Chinese essay scoring system using machine learning method. Especially, we are interested in rhetorical aspect of the writing. The System focuses on designing and retrieving features related to rhetoric. Since human grader might assign different score to the same essay at different times, the system treats the error of one point as tolerable. Based on it, a modified ID3 algorithm is constructed.

誌謝

感謝指導教授李嘉晃老師兩年來在學業與生活上的指導及提攜，以及口試期間承蒙陳信希教授、李肇林教授、莊仁輝教授詳細審閱論文以及給予熱心的指正及寶貴的意見，使論文更臻完備，特此表達由衷感謝之意。

論文的完成歸功於許多人的協助，感謝張道行學長的耐心指導，以及實驗室同學林俊豪、高鳴遠、黃明超、蔡沛言提供的寶貴建議以及資料，在此表答感謝之意。

最後感謝我的父母，由於有你們的全力支持使我的論文得以順利完成。謹以此論文獻給所有關懷、協助我的人。

張佑銘 謹誌於交通大學

民國九十四年六月

目錄

圖目錄

表目錄

第一章　前言

1.1 研究動機

現今學生語文能力普遍低落，其部份原因在於網路及手機通訊的發達，傳達訊息變得迅速且即時，使得一般學生慣於以簡短且口語化的文字互相溝通，結構嚴謹、詞藻洗鍊的傳統書信因而逐漸被取代，這樣一來便減少了鍛鍊文筆的機會。其次由於升學考試取消作文測驗，學生因而不重視國語文能力的培養。台積電董事長張忠謀先生曾提出「中文優勢論」[10]，在大中國市場興起的時代，中文能力已成為台灣吸引外資的人才優勢。但是事實上隨著非華語系國家學習中文的人數不斷增加，台灣的中文優勢正逐步消失。為了改善這個情形，教育部在民國九十三年十月宣佈，國中基本學力測驗將在民國九十五年試辦加考作文，民國九十六年正式實施。

作文可以使學生明瞭本國語言之特質、加強組織及表達概念的能力、強化寫作技巧、亦可增進文藝欣賞及創作之能力。不斷練習寫作可以整理學生從教學中零星片斷所獲得的寫作方法和知識，學生自然能增強寫作能力[2]。

不同老師的背景知識及主觀認知有相當大的差異性，因此批改作文要能有一致性的評分標準並不容易。老師在批改作文時，如果能有自動化的評分系統做為輔助，便可提高閱卷效率及評分的一致性。在西方，英文的自動作文評分系統已有長久的發展，同時也有相當成熟的文法分析工具，但在中文作文的處理上，缺乏文法分析工具，也沒有高效能的評分系統。因此我們希望建立一個自動化的作文評分系統來協助作文的評閱工作。

一般而言，老師批改作文時，主要是以該文是否符合主題、句法是否通順、文章完整性、組織架構、修辭用字等來做為評分的標準。判斷文章是否符合主題通常需要分析

句子及段落對於文章主題在語意上的關聯性。判斷文章句法是否通順需要檢查句子是否符合文法。評估文章完整性及組織架構需要對作文各段落的主要概念之間做語意及連貫性分析。而評估修辭用字則需要觀察各種修辭手法及詞彙的使用技巧。

作文要寫得動人，必然要注重修辭，修辭學探討表答情思的技巧與規則，以期達到精確明瞭、優美生動的境界[1]。本論文針對修辭用字方面，以機器學習的方式來學習老師在修辭學上的評分準則，並且發展自動化的評閱系統來協助閱卷及訓練老師的工作。

1.2 研究假設

在這一節說明一般作文測驗的評分制度，以及本論文對修辭評分工作的兩個假設。

1.2.1 傳統的作文評分方式

一般作文的評量方式是採用級分制[3]，通常為六級分，每份試卷最少由兩名老師評分。老師之間由於背景知識及主觀認知有相當大的差異，常常會影響作文評分時的標準。因此閱卷老師通常要先經過培訓，學習批改及計分的方式，使評分時盡量能達到標準化及符合一致性，以免產生爭議。如果兩名老師評分差異太大，就會找第三人仲裁，評量出合理的分數。

1.2.2 模擬修辭評分方式

底下將簡略地描述本論文的兩個基本假設：

假設一：

作文評分採級分制，分為六個級分，修辭分數由低到高為一分到六分，雖然每位老師對於作文的評分標準不盡相同，但這個差異也不會非常大，因此我們認為相差一分為可容許的誤差，整個研究的演算法設計及效能評估都是以此為標準。

假設二：

我們認為老師評斷修辭分數時，會觀察文中是否使用各種修辭技巧詞。在描述事情時，使用修辭技法，如譬喻、排比的學生應該有較優秀的修辭水準。本論文探討是否能夠以此為基礎並且以機器學習的方式來實行自動作文評分以協助作文的評分工作。本論文將著重於設計及擷取上述的各項修辭技法來做為機器學習所需要的特徵，並設計有效的機器學習方式。

1.3 研究目的

本研究嘗試建立一個系統，使用改良式 ID3 演算法學習訓練作文中的特徵後，生成決策樹產生規則，並以這些規則評斷測試作文的等級。老師在批改作文時，可以參考此系統所做出的分類來評分。或是以此系統作為輔助工具訓練老師批改作文的技巧。更進一步，在大型考試如基本學力測驗中，需要兩名以上閱卷老師以避免過於主觀造成不公時，加入本修辭評分系統的完整作文評分系統可以取代其中一名老師，這樣便能節省大量的人力以及時間。

1.4 論文架構

第一章為前言，內容著重於說明研究的動機、作文的重要性及所處理的問題。第二章為文獻探討，首先說明英文自動作文評分系統的發展，接下來解釋中文作文處理上的困難、與英文處理之間的差異性。第三章描述各項特徵及擷取方式。第四章提出一個以 ID3 演算法為基礎的改良式演算法。第五章為實驗實作及結果的呈現與分析。第六章為本研究的結論與未來發展。

第二章　文獻探討

在這一章我們將概略地描述英文作文評分系統(簡稱 AES)的發展，尤其著重於 e-rater 及 IEA 作文評分系統的介紹。最後將說明目前中文自動作文評分的困難之處。

AES 的發展可追溯到 60 年代[7]，其特徵的設計及擷取僅限於表面(surface)特徵，如平均字長、文章字數、標點符號數量、介系詞數量、罕用字數量等。Ellis Page 開發的 PEG(project essay grader)利用表面特徵及線性回歸的統計方法來預測老師的作文評分。這個方式並不被教育界所接受，原因在於學生容易掌握得高分的竅門，並且無法給予教學上的反饋[7]。80 年代早期的 WWB 系統開始著重教育性的反饋目標，包括拼字、措詞、以及可讀性等。雖然只是一些顯而易見的概念，卻是寫作品質分析自動化正確的一步[7]。90 年代隨著自然語言處理(NLP)及資訊擷取(IR)的發展，研究人員得以使用更新的工具及技巧，實行更有效的自動寫作品質分析，如評估句法變化(syntactic variety)、主題相關性(topic content)、概念結構(organization of ideas)等[7]。90 年代晚期的 e-rater、IEA 便是利用 NLP 及 IR 等新技術發展出的著名系統。

2.1 e-rater

1999 年 2 月，GMAT Analytical Writing Assessment 開始使用 e-rater[3]來協助文章的評分工作。原本人工評分的工作需要兩名閱卷者對同一篇文章評分，分數為六級分，如果兩位閱卷者認定的分數相差超過一分，則需要第三位閱卷者仲裁。而現在則以 e-rater 取代兩名閱卷者的其中一名，其中部份原因在於，e-rater 計算出的分數僅有 3%與實際閱卷者批改的分數相差超過一分，而這個數字與兩名受過訓練的閱卷者之間的誤差是相近的。

e-rater 包含了三個模組：句法(syntactic)、語段(discourse)、主題關聯分析

(topical analysis)。句法模組先將各個字標記詞性後，搜尋有意義的詞組並存入以動詞做次分類的結構樹中，藉此定義不定詞、完整、以及從屬子句。這些資訊提供文章中的各種句法給 e-rater。語段模組標記句子的主要概念，並且分割不同概念的句子集合。主題關聯分析模組則評估了字彙的發揮及與文章主題的相關性。藉著各模組提供的資訊，e-rater 決定文章的分數以及提供少許的反饋訊息。

2.2 IEA

Intelligent Essay Assessor(IEA)[4]是一個以Latent Semantic Analysis(LSA)[5]為基礎的軟體工具集合。LSA為一機器學習的方法，藉由SVD(singular value decomposition)方法在大量的文本中取得字與句在數學表達上的意義關係[6]。SVD方法不同於早期資訊擷取所使用的逐字比對(literally matching terms)方式，逐字比對的缺點在於，一個概念可能有許多種表達方式(synonymy)，會造成一篇文章中的關鍵字在相關文章內因使用不同的術語而無法被程式察覺。另外大多數的字有複數的意義(polysemy)，這會使得系統納入與關鍵字實際意義不同的資訊。而以SVD為基礎的演算法可以使用概念索引而非各別字的擷取。如此可以避免一些synonymy及polysemy所產生的問題。以LSA做為內容(content)模組，加上風格(style)、技巧(mechanics)兩個模組所組成的IEA系統可以達到與e-rater相同的精確率[7]。此外IEA亦可提供一些簡單的反饋訊息。

2.3 中文處理

任何一個語言處理的系統都必須先能分辨文本中的各個詞才能進行語言分析、自動翻譯、資訊擷取等進一步的處理。中文與英文最顯而易見的差異，在於中文的詞與詞之間並無空白加以區隔，因此中文自動分詞成為必要的工作。

由於新的辭彙會不斷增加，辭典無法列出所有的中文詞，參考辭彙不足造成中文自動分詞的處理相當困難。其中最重要的一項課題是需要擷取未知詞如專有名詞、外來語。另外一個重點在於，不同的系統往往只測試自行選定的測試文集，各個系統彼此間欠缺一個比較的基礎。為了彌補這樣的缺陷，[8]提供了一個分詞技術互相比較的環境。由於中文分詞不可能達到百分之百正確，因此在擷取文本中的資訊以進行各項處理時必然會受到負面的影響。

另外，AES 系統如 e-rater 等，使用文法分析工具來分析句型及修辭結構，但是中文與英文的文法有極大的差異，英文的文法分析工具無法在處理中文作文時使用，而且中文的文法分析工具亦相當缺乏，因此我們將研究目標集中在評估用字遣詞及使用修辭技法的能力。

第三章　中文自動作文修辭評分系統設計

— 特徵擷取

自動修辭評分系統的主要流程如圖 1，第一步先處理中文斷詞及標記詞性，然後擷取修辭特徵，接著使用改良式的 ID3 演算法生成決策樹，最後以決策樹判斷測試作文的分數。自動修辭評分系統設計的兩個重點為擷取修辭特徵以及設計機器學習方法。我們將在本章說明擷取修辭特徵的方式，在下一章描述機器學習的方法。

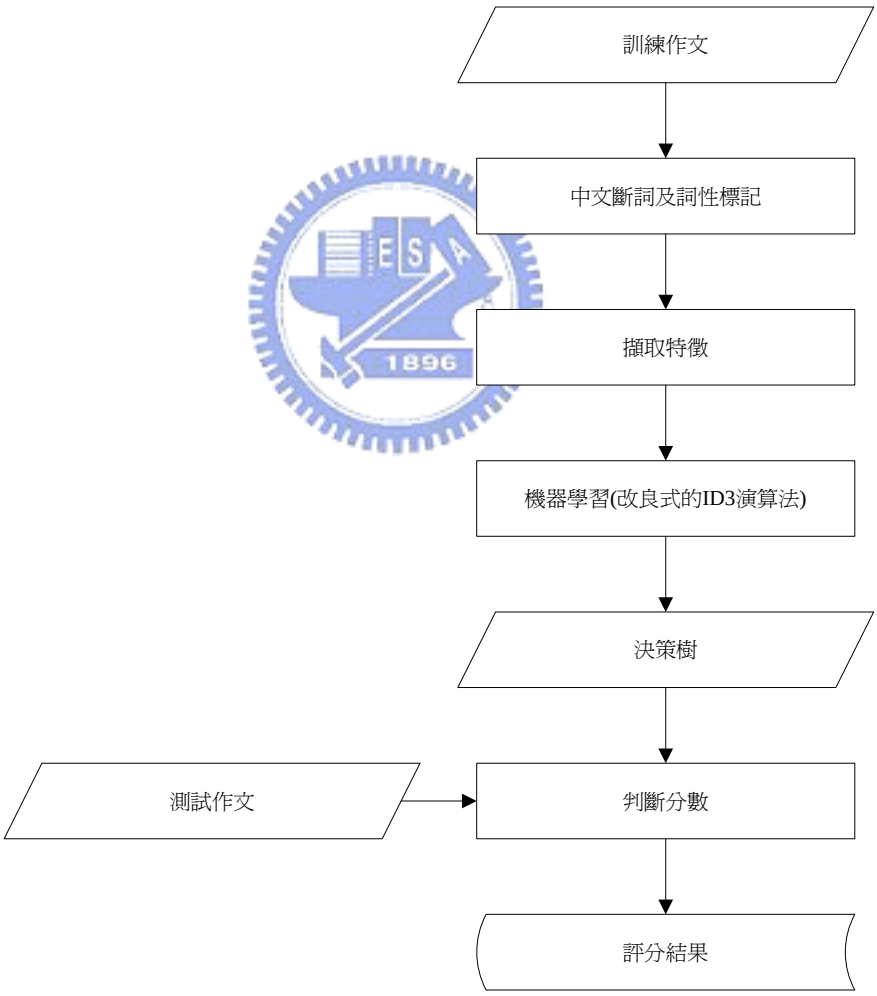

圖 1 中文自動修辭評分系統架構圖

作文中的特徵可分為間接特徵與直接特徵。間接特徵(或稱為表面特徵)的出現頻率

雖然與作文的分數有相關性，但是老師在批改作文時，並不是以間接特徵作為主要的評分依據，因此這些特徵比較缺乏教育上的反饋功能。一般老師閱卷時觀察的特徵我們稱之為直接特徵，這類特徵能夠更直接地評量文章的品質，因此本研究在評分系統中除了間接特徵外加入直接特徵，以期能更有效估計作文的修辭水準。

3.1 間接特徵

間接特徵包含詞數、形容詞數、及成語數。表達感情與思想需要透過形容詞的修飾。另外，使用成語代表寫作者受過較深入的國語文訓練，修辭能力會有一定的水準。兩者的數量又與文章長度有所關聯，因此我們加入了詞數這一項特徵。

圖 2、3、4 分別是詞數、形容詞數、成語數在所有作文樣本中的分布。很明顯隨著修辭分數的增加，平均詞數、平均形容詞數、平均成語數的數量也都會上升，也就是說這些特徵的統計數量與修辭能力有正相關性。

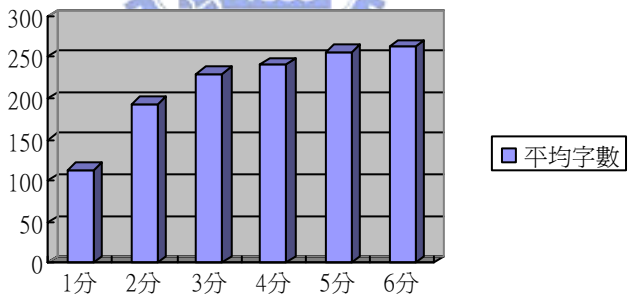

圖 2 各分數作文平均詞數統計圖。
1 分：114、2 分：194、3 分：230、
4 分：241、5 分：257、6 分：263。

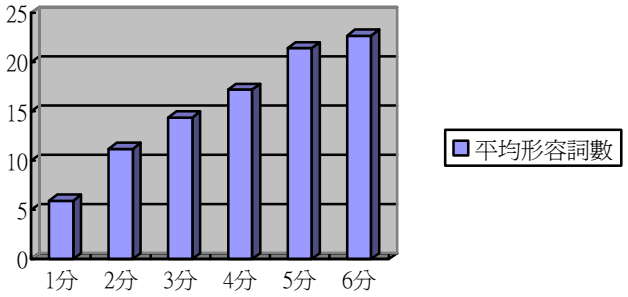

圖 3 各分數作文平均形容詞數統計圖。

1 分：5.95、2 分：11.19、3 分：14.4、

4 分：17.2、5 分：21.41、6 分：22.67。

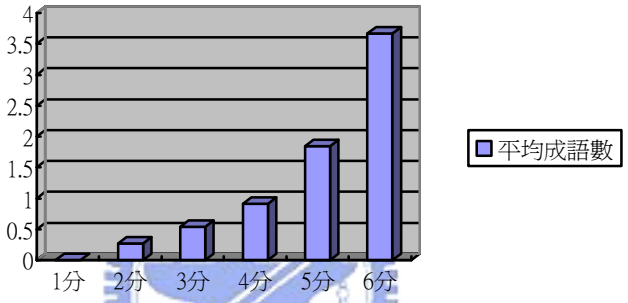

圖 4 各分數作文平均成語數統計圖。

1 分：0、2 分：0.27、3 分：0.54、

4 分：0.91、5 分：1.84、6 分：3.67。

3.2 直接特徵

本研究使用的直接特徵包含譬喻手法、非口語化的喻詞以及排比手法。

3.2.1 譬喻手法

譬喻就是用已知喻未知[1]，利用已知的知識描繪未知的情境。一個好的譬喻，往往能夠給人留下深刻的印象。例如：「下課十分鐘時，學校就像一個菜市場。」這樣的一個譬喻手法是借用菜市場很吵雜的已知知識來描述下課時學校的情境。

譬喻手法包含明喻、隱喻、略喻、借喻等[1] 。因為略喻、借喻需要語意的分析，所以本研究只使用明喻及隱喻手法。明喻及隱喻主要是由本體、喻體、及喻詞所構成。舉例來說：「下課十分鐘時，學校就像一個菜市場。」這句話中「學校」是本體、「像」是喻詞、「菜市場」是喻體。本論文採用的明喻及隱喻手法基本形式有兩種，一種是本體、喻詞、喻體皆在同一句；另一種是喻詞及喻體在同一句，本體在前面一句的句子組合，如：「原本寧靜的校園，霎時變成了熱鬧的菜市場。」底下是從作文中找出明喻及隱喻手法的步驟及實例說明。參照圖 5，首先要找出所有可能的喻詞，再搜尋符合譬喻手法形式的句子。這個演算法需要使用辭典，我們使用的辭典為「中文詞庫」，其中記載了 78219 個詞以及各個詞的詞性。該演算法是半自動化的，有些步驟需要語意及文法分析，因此我們以人工處理。

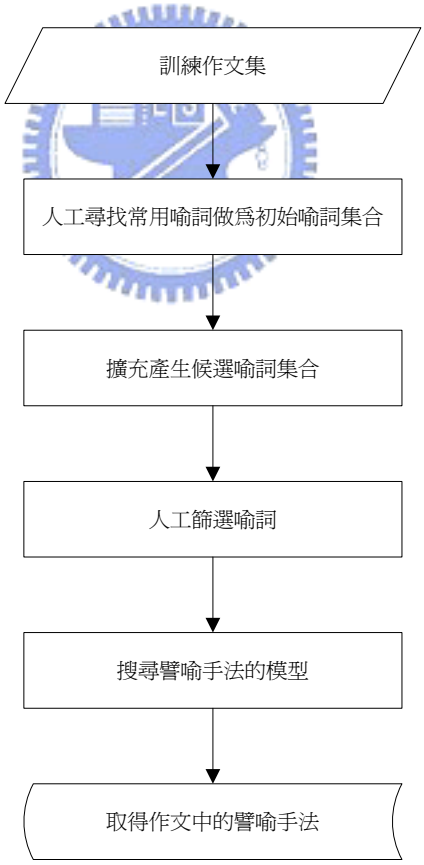

圖 5 擷取譬喻手法流程圖

I. 尋找常用喻詞。

1. 以人工在訓練作文中找出一個或數個包含本體、喻體、喻詞的譬喻手法，如「學校就像一個菜市場。」，取出其喻體—菜市場。

2. 到訓練作文中搜尋出現 I.1 所取出之喻體的句子，將句子中符合語意的喻詞以人工挑出。例如從「由安靜的教室一下子轉變成熱鬧的菜市場。」中挑出「轉變成」這一個喻詞。這些人工挑出的詞稱為「初始喻詞集合」。

將資料庫中的所有作文都做為訓練資料時，這個步驟由人工所產生的初始喻詞集合有 15 個。基本上，首先由人工挑出菜市場這個詞之後輸入程式，而後由程式在資料庫所有作文中找出包含菜市場的句子，再從這 45 個句子以人工選出合乎語意的譬喻手法並且取出該譬喻手法所使用的喻詞做為初始喻詞集合，該集合內計有 15 個喻詞。

II. 擴充產生候選喻詞的集合。

1. 在辭典中查出「初始喻詞集合」中各個喻詞的詞性，例如：“好像”的詞性為 VG(分類述詞)、“如同”的詞性包含 VG 及 P(介詞)。將這些詞性稱為「候選詞性集合」。

2. 在辭典中找出所有詞性屬於「候選詞性集合」並且有出現在訓練作文中且其後在同一句內有名詞的詞，稱為「候選喻詞集合」。例如“變成”這個詞出現在「使學校變成一間超大間的菜市場。」中，且其後有一個名詞“菜市場”，便把“變成”納入候選喻詞集合之中。其後需要名詞的原因在於明喻及隱喻手法必然會有喻體存在，而喻體往往在喻詞的後方，且兩詞會在同一句之內。這個集合的數量相當龐大，其中包含了許多不能做為喻詞的詞。

將資料庫中的所有作文都做為訓練資料時，這個步驟所產生的候選詞性集合包含 13 個詞性，候選喻詞集合包含 766 個詞。

III. 篩選喻詞。

以人工在「候選喻詞集合」中篩選語意上適合做為喻詞的詞，稱為「喻詞

集合」。

將資料庫中的所有作文都做為訓練資料時，這個步驟所產生的喻詞集合有34個。

IV.搜尋譬喻手法

這裡使用兩種句型來擷取譬喻手法：

句型一、對於每一個屬於「喻詞集合」的詞，到作文中找出該詞前後皆有名詞的句子，認定這是一個使用譬喻手法的句子。如：

整個　校園(NC)　看起來　「像」　菜市場(NC)　。

喻詞“像”之前有名詞“校園”，之後有名詞“菜市場”，符合我們定義的條件，因此將之視為譬喻手法。

句型二、這個方法分兩個步驟：

a.找出喻詞之後有名詞及形容詞，喻詞之前沒有名詞的單一句子。

b.在a.中找出的句子的前一句如果有出現名詞，則認定這是一個譬喻手法。如：

原本　寧靜　的　校園(NC)　，

霎時　「變成」　了　熱鬧(VH)　的　菜市場(NC)　。

第二句中的喻詞“變成”之前沒有名詞，之後有名詞“菜市場”及形容詞“熱鬧”，且第一句中包含了名詞“校園”，符合我們定義的條件，因此將之視為譬喻手法。

利用上述演算法在作文中找出譬喻手法，做為機器學習時使用的特徵。圖6為所有作文樣本中各分數的平均譬喻數量，愈高分的作文愈有可能出現譬喻手法。

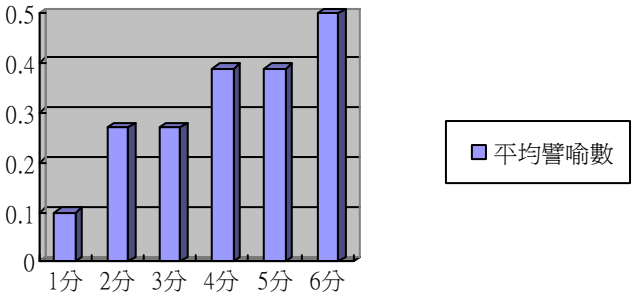

圖 6 各分數作文平均譬喻數統計圖。
1 分：0.1、2 分：0.27、3 分：0.27、
4 分：0.39、5 分：0.39、6 分：0.5。

3.2.2 非口語化的喻詞

我們觀察資料庫中的作文後發現，譬喻手法有著口語化及非口語化的現象，例如：「學校變得「跟」菜市場一樣。」這個用法較口語化，而：「菜市場「般」的學校」則相反，這兩者的差別在於所使用的喻詞不同。因此我們假設一般學生會使用口語化的喻詞，而修辭程度高的學生則會使用非口語化的喻詞。計算每一個喻詞在作文中的分布狀況，在高分作文出現比例遠高於在低分作文出現比例的喻詞，便歸類為非口語化的喻詞。篩選的步驟如下：

1. 將訓練作文分為高低分兩群，高分群為 4 到 6 分，低分群為 1 到 3 分。

2. 取出喻詞集合(A)中的各個喻詞 a_i(i∈{1 to size(A)})，高分作文中包含喻詞 a_i 的作文數量記為 high(a_i)，低分作文中包含喻詞 a_i 的作文數量記為 low(a_i)。

3. 計算 high(a_i)在出現喻詞 a_i 的所有作文中所佔的比例：

$$Proportion(high(a_i)) = high(a_i)/(high(a_i)+low(a_i)) \quad (1)$$

4. 如果 Proportion(high(a_i))大於 0.6，也就是喻詞 a_i 主要出現在高分的作文之中，那麼便認定喻詞 a_i 是一個非口語化的喻詞，使用這個喻詞的譬喻手法會是一個非口語化的譬喻手法。

由以上的步驟會篩選出一個非口語化的喻詞集合。我們觀察到以此方式篩選出使用這些喻詞的句子大多是相當優秀的譬喻手法。圖 7 為所有作文樣本中各分數的平均非口語化喻詞數量，高分作文中非口語化喻詞的數量遠超過低分作文。

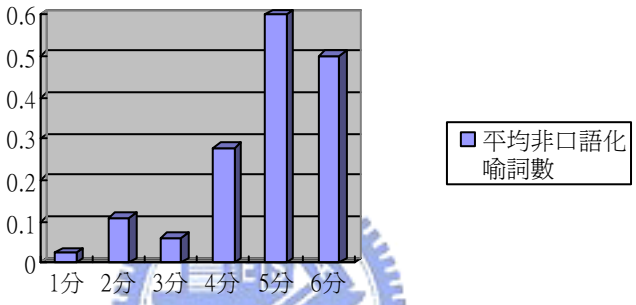

圖 7 各分數作文平均非口語化喻詞數統計圖。
1 分：0、025 分：0.11、3 分：0.06、
4 分：0.28、5 分：0.6、6 分：0.5。

3.2.3 排比手法

對於同一範圍、同一性質的意念，用兩個以上結構相似的句法來表達的一種修辭技巧，叫做排比。如：「打球的打球，散步的散步。」是一個排比手法。排比和對偶頗為相似，但兩者之間也有區別：對偶必須字數相等、兩兩相對，排比則不拘；對偶須避免用意或相同字，排比卻往往意思相同、字也相同[11]。上述概念為建立排比句型的基本原則。

在本論文中使用兩個明顯的排比手法句型，用以找出類似的排比句子。在此說明兩種排比句型：

句型一、相鄰的兩個句子詞數相同，相對應詞的詞性也相同：

假設 S_1 與 S_2 是鄰接的兩個句子，而且

$$\text{Size}(S_1) = \text{Size}(S_2),$$

$$\& \ \forall i \in \{1 \text{ to } \text{Size}(S_1)\}, \ \text{POSTag}(S_{1\,i}) = \text{POSTag}(S_{2\,i}). \quad (2)$$

則 S_1 與 S_2 便是一個排比手法，其中 Size(S) 是指句子 S 的詞數，$\text{POSTag}(S_i)$ 是指 S 中第 i 個詞的詞性。例如：

「打球的打球，散步的散步。」前後兩句可各分為

打球(VA) 的(DE) 打球(VA) ，

散步(VA) 的(DE) 散步(VA) 。

兩句的詞數相同，相對應詞的詞性也相同，於是將這兩句視為一個排比句法。

句型二、選定一個句子，在之後 6 個句子之內有相同詞數、相對應詞性亦相同的句子，即：

$$\forall i \in \{1 \text{ to Sentence number of the essay}\},$$

$$\forall j \in \{2 \text{ to } 6\}, \ \forall k \in \{1 \text{ to } \text{Size}(S_i)\},$$

$$\text{Size}(S_i) = \text{Size}(S_{i+j}) \ \& \ \text{POSTag}(S_{i\,k}) = \text{POSTag}(S_{i+j\,k}). \quad (3)$$

i 的範圍從作文的第一句開始直到最後一句，若能找出符合上述條件的 i 與 j，則視句子 S_i 到句子 S_{i+2j-1} 為一個排比手法。例如：

「到操場走走，可以看到有人悠閒的慢跑；到合作社走走，可以看到大家都快樂的買著自己需要的東西。」以分號為區隔，前後兩部份的第一句為

到(P)　操場(NC)　走走(VA)　，

以及

到(P)　合作社(NC)　走走(VA)　，

兩個句子的詞數相同，相對應詞的詞性也相同，於是將這個句型組合視為一個排比句法。

利用這兩種句型在作文中找出排比手法，做為機器學習時使用的特徵。圖 8 為所有作文中各分數的平均排比手法數量。高分的作文出現排比手法的機會較高。

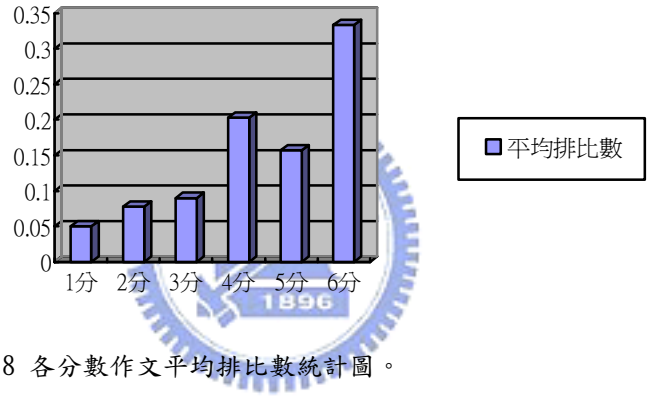

圖 8 各分數作文平均排比數統計圖。

1 分：0.005、2 分：0.078、3 分：0.09、

4 分：0.203、5 分：0.157、6 分：0.333。

第四章　中文自動作文修辭評分系統設計

— 演算法設計

本系統要模擬的對象是批改作文的老師，而老師不可能有無限的時間評閱作文，在有限的時間內要完成閱卷工作，通常第一印象佔有相當重要的地位。第一印象近似於顯著的特徵，而 ID3 演算法建立決策樹時，即是拿最顯著的特徵來做第一次分類，也就是整個樹的根結點，因此我們使用 ID3 演算法作為基本架構，並配合資料的分布及允許一分誤差的概念修改此一演算法，使得整個系統能更符合我們的需求。其他機器學習方法如類神經網路、貝氏學習、concept learning 等並不在本論文的討論範圍內。

4.1 ID3 演算法

ID3 演算法[9]利用訓練資料建立決策樹，從樹的根節點到任一葉節點路徑上的內部節點用來比對例子(instance)中的屬性(attribute)，將例子加以分類，每條路徑亦可視為一個 if-then 的規則。ID3 演算法利用各屬性的 entropy 來挑選最有鑑別力的屬性，entropy 的定義為：

$$Entropy(S) \equiv \sum_{i=1}^{c} -p_i \log_2 p_i \quad (4)$$

S 為一個概念的集合，c 代表目標(target)屬性的類別數，p_i 為 S 集合中目標屬性的 i 分類所佔的比例。以下以一個簡單的例子說明整個演算法：

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

表 1 Play Tennis 範例

表 1 中有 14 個例子，記錄在哪一些天候條件下會去打網球，有 Outlook、Temperature、Humidity、Wind、Play Tennis 等 5 個屬性，其中 Play Tennis 是目標屬性。假設 S 是 Table 1 中 14 個例子的集合，c 為 Yes 和 No 兩個分類，p_i 是 i 分類的比例，其中 i∈{Yes, No}。則 S 的 Entropy 為：

$$Entropy([9+,\ 5-]) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$=0.940$$

挑選最有鑑別力的屬性時使用一個量化的值，這個值稱為 Information gain，屬性 A 在集合 S 之中的 information gain 記為 $Gain(S,\ A)$，其定義如下：

$$Gain(S,\ A) \equiv Entropy(S) - \sum_{v\in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (5)$$

屬性 A 可能出現的值記為 $Values(A)$，S_v 代表 S 的一個子集，這個子集的範圍是 S 集合中屬性 A 的值為 v 的所有例子。以 Humidity 與 Wind 這兩個屬性為例，如圖 9，Humidity 可能的值為 High 與 Normal，Humidity 為 High 的例子有 7 個，其中有 3 個會去打網球，4 個不會，其 Entropy 便為

$$-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}=0.985$$

接著算出 Humidity 為 Normal 時的 Entropy 後，代入 Information Gain 公式可求得 Gain(S, Humidity)為 0.151，以同樣算法可以得到 Gain(S, Wind)為 0.048，兩相比較，Humidity 會有較高的分類能力。

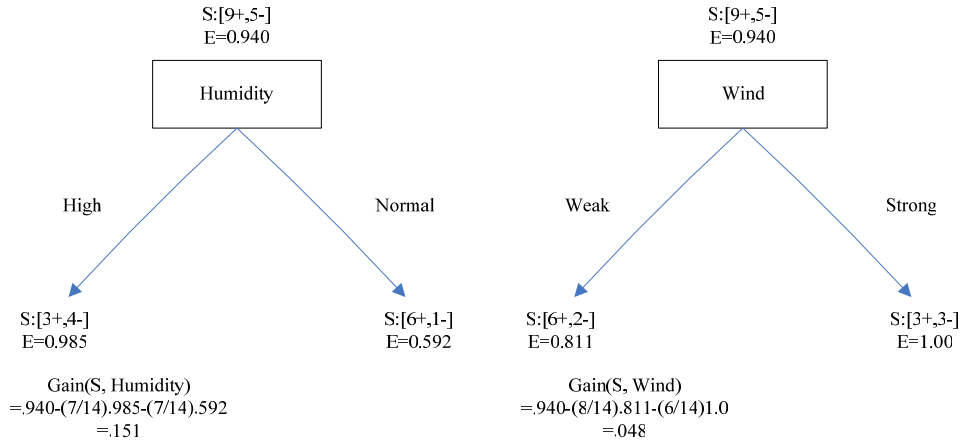

圖 9 決策樹範例圖

算出所有屬性的 Information Gain 之後，取該值最高的屬性當作決策樹的根節點，在表 1 的例子中為 Outlook，接下來再將所有例子分成 Outlook 為 Sunny、Outlook 為 Overcast 以及 Outlook 為 Rain 的三個子集，每個子集再對尚未始用的屬性計算 Information Gain，以此類推直到子集中的分類全都為同一類或是用完所有屬性為止。這樣的一個決策樹建立完成後，可以在未知分類的例子出現時，根據這個例子的屬性來判斷它應該屬於哪一個分類，在上面的例子中便能判斷在一個新的天候狀態下是不是會去打網球。

如果一個決策樹的節點太多，會產生 overfitting 的現象。以建立決策樹時所使用的訓資料做為測試資料來判斷效能時，由於決策樹與訓練資料有高相關性，因此會有相當高的精確率。但以訓練資料以外的資料做為測試資料時，隨著節點的增加，精確率會隨之降低。為了減低 overfitting 的產生，使用 prune 的來減少節點是必要的。執行 prune 的時候需要把部份訓練資料獨立出來做為測試決策樹準確率的確認(validation)資

料，而僅使用剩餘的訓練資料來建立決策樹。在某個節點做 prune 的動作是在不影響以確認資料測試時之精確率的前提下移除以這個節點為根的子樹，把這個節點視為葉節點，並將訓練資料中與從決策樹根部到此節點所表示的 if-then 規則相同的所有資料裡，數量最多的類別做為葉節點的分類值。

4.2 適合作文評分的分類方式

我們在實驗過程中發現，符合一個 if-then 規則的所有作文中各個分數的數量分布有時會相當均勻。如圖 10，符合詞數數量級為 0、形容詞數數量級為 2 的規則之訓練作文有 5 篇為 1 分、4 篇為 2 分、3 篇為 3 分、1 篇為 4 分。而 ID3 分類一個 if-then 規則是使用訓練資料中符合此一規則最大數量的類別，上面例子中的分布雖然部份候選類別頻率相近，但是由於 1 分作文的數量最多，所以 ID3 會選擇 1 分做為這個規則的分類，但這樣做在作文評分時是不恰當的，因為這麼一來可能會造成許多 3 分及 4 分作文被評為 1 分。以 e-rater 為例，初步評分時的兩名老師如果對一篇作文的評分差異超過 1 分(如 1 分與 3 分)，便需要第三名老師的加入以評估出最後的分數。因此我們希望能盡量使系統預測的分數及實際分數相差不超過一分，在此使用一個類似求變異數的函式來計算最佳的分類：

$$\text{class(S)} = \underset{\text{i}\in\{1\text{to }6\}}{\arg\min} \sum_{j\in\{1,2,3,4,5,6\},\, j\neq i} (i-j)^2 * freq(j) \qquad (6)$$

S 是符合決策樹中某一規則的訓練作文集合，i 的範圍為 1 分到 6 分，j 代表分數 i 之外的其他分數，freq 是訓練資料中分數為 j 的作文數量，$(i-j)^2$ 的意義是 j 離 i 越遠懲罰值越高。使得 $\sum (i-j)^2 * freq(j)$ 的值最低的 i 便做為這個規則的分類分數。在訓練作文中有 5 篇為 1 分、4 篇為 2 分、3 篇為 3 分、1 篇為 4 分的情況下，利用上述方式計算之後求得的最佳分類為 2 分，這時僅有 1 篇作文的預測分數與實際分數相差 2 分以上，而非使用原始 ID3 演算法分類方式時的 4 篇。這個分類方法對作文評分系統的效能有實質上的幫助。

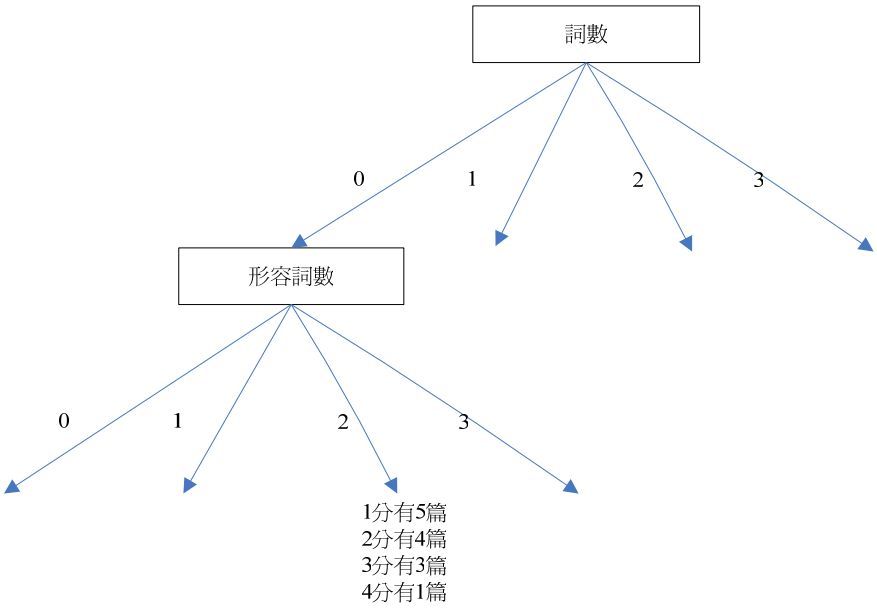

圖 10 決策樹範例圖

4.3 調整規則分類

在完成決策樹的建立之後，我們利用所有的訓練(training)和確認(validation)作文來調整各規則的分類。使用這個方法是因為我們取得的作文數量僅有 693 篇，相對來說訓練作文及確認作文的數量也就受到限制，因此藉由結合兩者的數量以期減少隨機取樣時作文特徵在訓練、確認、測試作文中分布不均勻而影響系統效能的情況。

在訓練和確認作文中篩選出符合 if-then 規則 $rule_p$(p 的範圍為“1”到“決策樹所形成的規則總數”)的作文集合 S，先取出該規則的分類，記為 c。計算作文集合 S 中，實際分數與規則分類分數 c 相同的比例，記為 same(c)；接著計算作文集合 S 中，實際分數與規則分類分數 c 相差不超過一分的比例，記為 close(c)。如果 same(c)<0.3 或 close(c)<0.8，則將此一規則標記為“困惑規則”。其中選擇 0.3 及 0.8 做為邊界的原

因在於，作文資料庫中所記錄的任意兩名老師對相同作文評分分數經過統計之後，完全相同的大約佔30%，而相差一分以內的大約佔80%。因此，使得same(c)低於0.3或close(c)低於0.8的規則，其分類我們認為可能是一個可信度不高的分類。

接下來挑出所有被標記為“困惑規則”的規則，對於每個困惑規則，取出訓練和確認作文中所有符合該規則的作文，利用公式(6)：

$$\text{class(S)} = \arg\min_{i\in\{1\text{ to }6\}} \sum_{j\in\{1,2,3,4,5,6\},\, j\neq i} (i-j)^2 * freq(j)$$

求出這個作文集合的最佳分類分數，並和該困惑規則的分類分數作比較，如果最佳分類分數大於規則的分類分數，便將這個規則的分類分數加1分，反之若是最佳分類分數小於規則的分類分數，則將這個規則的分類分數減1分。

如圖11，形容詞數數量級為3，成語數數量級為1，非口語化譬喻手法數量級為0的規則其分類為3分，符合這個規則的訓練作文有2篇為3分、2篇為4分，確認作文有6篇為3分、4篇為4分、4篇為5分。總共有8篇為3分、6篇為4分、4篇為5分。經過計算後發現實際分數與規則分類分數差距不超過一分的比例為77.8%，小於80%，符合“困惑規則”的條件。而這些作文的最佳分類分數為4，大於此困惑規則的原始分類分數—3，因此將此規則的分類分數加1分。

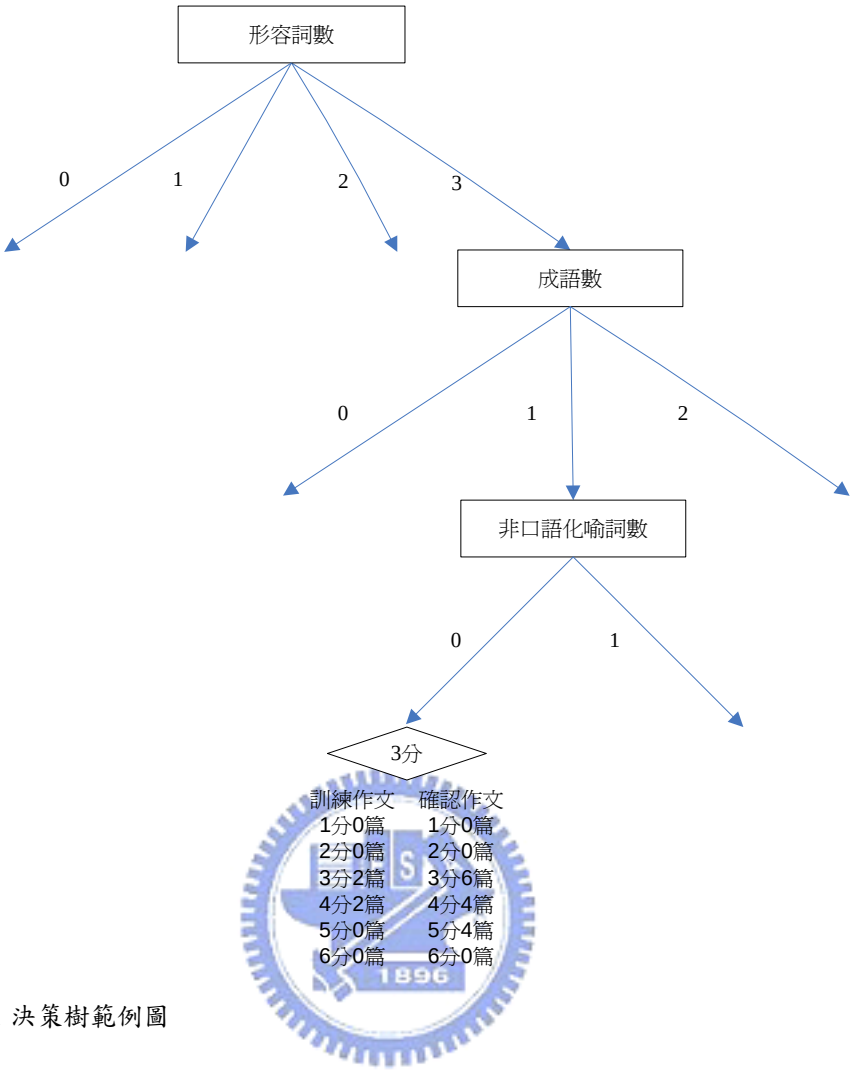
形容詞數
0
1
2
3
成語數
0
1
2
非口語化喻詞數
0
1
3分
訓練作文
1分0篇
2分0篇
3分2篇
4分2篇
5分0篇
6分0篇
確認作文
1分0篇
2分0篇
3分6篇
4分4篇
5分4篇
6分0篇

圖 11 決策樹範例圖

第五章　實驗實作

我們使用的作文為國中二年級學生作文，這些作文輸入成電子檔時保留所有的錯字以及標點符號，不加以修改，以維持學生作文的原貌。修辭分數從 1 分到 6 分，1 分為最低，6 分為最高。每篇由二到三名老師評分，取平均分數。實驗資料包括 693 篇作文。修辭分數 1 分到 6 分的文章篇數分別為 40 篇、166 篇、234 篇、177 篇、70 篇以及 6 篇。每次實驗時隨機取出 394 篇訓練作文，其中 197 篇用來建立決策樹，197 篇用來作為 prune 時的檢驗資料，299 篇測試作文則用來評估整個系統的效能。

5.1 實驗流程

本實驗是藉由 ID3 演算法建立決策樹，因此要將連續資料予以離散化。先將所有訓練作文中 2 個連續性特徵—詞數及形容詞數—由小到大排序得到兩個數列，以這兩個數列各自的第一四分位數、中位數、第三四分位數做為邊界，將詞數及形容詞數各分成 4 群。成語數分為沒出現、出現 1 次、出現 2 次以上共 3 群；譬喻手法、非口語化的喻詞及排比三項特徵分為沒出現、出現 1 次以上共 2 群。

訓練作文先經過「中央研究院資訊科學研究所詞庫小組中文斷詞系統 1.0 版」進行斷詞與詞性標記後，開始擷取各項特徵。特徵擷取完畢之後經由改良的 ID3 演算法產生決策樹，測試用的作文同樣經過斷詞與詞性標記後，取出特徵，比對決策樹由根部到所有葉節點的路徑後可以得到該文的分數，所有測試作文經比對規則得到分數後，再與實際評分比較，以評估系統的效能。

5.2 實驗結果與討論

三次隨機抽取訓練及測試作文，並分別記錄兩種機器學習方式的效能，包括原始的 ID3 演算法(ID3)、以及本修辭評分系統使用的演算法—“Modified ID3”。我們使用兩

個數值來評估系統的效能：由於本論文是建立在“作文評分時相差一分為可容許的誤差”之概念上，因此計算測試作文中系統預測分數與實際分數相差一分以內的比例，記為 Adjacent value。另外一個數值名為 Exact value，代表系統預測分數與實際分數完全相同的比例。在三次的實驗中如，使用 modified ID3 方式的機器學習方法其效能皆高於原始的 ID3 演算法，平均 Adjacent value 高出 4.7%。其結果記錄在表 2 及表 3 中。表 2 為三次實驗中，以原始 ID3 演算法以及本系統所使用的 Modified ID3 演算法對測試作文評分後所得之 Adjacent value。表 3 為三次實驗中，以原始 ID3 演算法以及本系統所使用的 modified ID3 演算法對測試作文評分後所得之 Exact value。

	ID3	Modified ID3
實驗一	0.85	0.93
實驗二	0.857	0.89
實驗三	0.884	0.913
平均	0.864	0.911

表 2 Adjacent value 統計表

	ID3	Modified ID3
實驗一	0.345	0.375
實驗二	0.395	0.408
實驗三	0.362	0.385
平均	0.367	0.389

表 3 Exact value 統計表

	Exact	Adjacent
Two teachers	0.297	0.749
Modified ID3	0.389	0.911

表 4 系統效能評估表

在作文資料庫中，每篇作文由二到三名老師批改，每一位老師批改的作文數目為 50 到 100 不等，此外這些老師都沒有接受標準化批改作文方式的訓練。我們計算每一篇作文的任兩名閱卷老師所評分數的差距，如表 4，兩名老師之間的平均 exact value 為 0.297，adjacent value 為 0.749，而本研究之自動評分系統與作文平均分數之間的 adjacent value 平均值為 0.911，較兩名老師間的 adjacent value 高出 16.2%。因此老師在批改作文時，本系統可以提出一個具有可信度的參考分數。

第六章 結論與展望

在本論文中，我們使用六種修辭特徵及一個改良式的 ID3 演算法做為機器學習的方式來預測作文分數。這個改良式的演算法在決定一個 if-then 規則的分類時，為了符合作文評分的特性，於是以「避免產生過大誤差，但容許微小的差異」為原則定義了一個適合作文評分的分類方式，這個方式也確實帶來了效能上的提升。在系統與老師，以及兩名未受過訓練老師間的評分差異性評估中，系統與老師間的差異要比兩名不同老師間的差異來得低。因此本系統可以嘗試做為批改作文時的協助工具，提出作文分數以供參考。

往後我們希望能建立完善的修辭技法分析工句，以擷取作文中其他重要的修辭技法。另外，設計一個線上評分介面，輸入作文後顯示該作文的評分，並提供使用者一些修辭方面的建議。

參考文獻

[1]黃麗貞。《實用修辭學「增訂本」》台北，文津出版社，2004。

[2]陳品卿。《中學作文教學指導》國立臺灣師範大學中等教育輔導委員會，1989。

[3] Jill Burstein. The E-rater Scoring Engine: Automated Essay Scoring With Natural Language Processing. *Automated Essay Scoring: A Cross-Disciplinary Perspective* (2003), pp. 113-121

[4] Thomas K Landauer, Darrell Laham, Peter W. Foltz. Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor. *Automated Essay Scoring: A Cross-Disciplinary Perspective* (2003), pp. 87-112

[5] Thomas K. Landauer, Darrell Laham, Bob Rehder, and M. E. Schreiner. How Well Can Passage Meaning be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans. *In M. G. Shafto & P. Langley (Eds.), Proceedings of the 19th annual meeting of the Cognitive Science Society* (1997), pp. 412-417.

[6] Berry, M. W., Dumais, S. T. and O'Brien, G. W.. Using linear algebra for intelligent information retrieval. *SIAM: Review* (1995), 37(4), 573-595.

[7] Hearst, M.. The debate on automated essay grading. *IEEE Intelligent Systems* (2003), 15(5), 22-37, IEEE CS Press.

[8] Richard Sproat, Thomas Emerson. The First International Chinese Word Segmentation Bakeoff. Which was held as part of the *Second Meeting of SIGHAN*, July 11-12, 2003 in Sapporo, Japan.

[9] Tom Mitchell, McGraw Hill. *Machine Learning* (1997).

[10]92.10.22 經濟日報

http://mag.udn.com/mag/life/storypage.jsp?f_ART_ID=6732

[11] http://163.26.9.12/noise/hcjh-ca/1-12.htm