國立交通大學

資訊科學與工程研究所

碩 士 論 文

基於貝氏機器學習法之中文自動作文評分系統

A Bayesian Based Chinese Essay Scoring System

研 究 生：林信宏

指導教授：李嘉晃　教授

中 華 民 國　九 十 五　年 六 月

基於貝氏機器學習法之中文自動評分系統

A Bayesian Based Chinese Essay Scoring System

研 究 生：林信宏　　Student：Shin-Hung Lin

指導教授：李嘉晃　　Advisor：Chia-Hoang Li

國 立 交 通 大 學

資 訊 科 學 與 工 程 研 究 所

碩 士 論 文

A Thesis
Submitted to Institute of Computer Science and Engineering
College of Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in

Computer Science

June 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年六月

基於貝氏機器學習法之中文自動作文評分系統

學生 ： 林信宏　　指導教授：李嘉晃 博士

國立交通大學電機資訊學院　資訊科學與工程研究所

中文摘要

在本論文中，我們探討文章直接與間接特徵對於寫作評分之間的關係，並以此作為基礎，建立一套以貝氏機器學習法為主的中文作文自動評閱系統。在本研究中我們認為，文章的間接特徵(外在特徵)雖然無法提供足夠的語義資訊，卻往往深切的影響評分老師對於文章好壞評斷的第一印象；如文章字數、分段數、標點符號的正確使用與否等其他多項外在因素，皆為評分老師在尚未深入細讀文章內容時用以作為評分標準的圭臬。但一篇文章的好壞，不僅僅只是以外觀的特徵來決定，且須更進一步探討文章的各段內容。據此想法，本系統對於文章的評分流程共分為三個階段：1.Holistic Scoring-整體評鑑 2.Paragraphic Scoring-分段評鑑 3.Integration-評鑑整合。而根據實驗結果，本系統評閱的正確率可達 95%~97%，是作為閱卷老師評分時的良好工具之一。

A Bayesian Based Chinese Essay Scoring System

Student : Shin-Hung Lin Advisor：Prof. Chia-Hoang Lee

Institute of Computer Science and Engineering

College of Computer Science

National Chiao Tung University

Abstract

This paper proposes an efficient method based on Bayesian Theorem to score Chinese essay according to the direct and indirect features of an essay. It includes words, nouns, themes, oral writing, average number of words of a section and concepts of an essay. In this study, we determined the holistic and paragraphic score of a testing data firstly. Subsequently calculate the grade of the testing data by integrating the relation of holistic and paragraphic score. Experimental results show that our approach compares favorably with some other Automatic Chinese Essay Scoring (ACES) systems.

目錄

圖表

表格

第一章、緒論

1.1 研究動機

世界上各個國家都重視自己的語言文化傳承，在教育學習的階段，尤其重視語文寫作表達能力的訓練，其中，作文教學是語文表達能力訓練中最重要的一環，作文不但可培養學生的語文能力，更能訓練其思考、邏輯與綜合思辨能力。但伴隨著時代的變遷與科技的進步，傳遞訊息的管道已不再侷限於傳統的書信往來，透過網際網路的通訊溝通方式，文字傳遞變得迅速且即時，也因而導致大部份學生慣於以簡短且口語化的文字作為溝通方式。

寫作能力是一種綜合訓練，既是語言文字的訓練，同時也是思維能力的訓練。藉由寫作過程，可以訓練一個學生的思考、理解、推理、及創作等能力；同時亦可檢視該學生是否已理解本國語言文字並能加以靈活應用。在日常生活中，舉凡書信、公文、簡報、履歷表、及婚喪喜慶等各類應酬文書等，皆屬作文的範疇，絕大部份的人或多或少都會接觸到作文，由此更突顯出作文教學的必要性。

一般的作文批改模式，往往須耗費大量的人力、物力以及長時間的作業過程，但除此冗長的過程之外，如何維持批閱作文的公正性則是另一重大考量。在升學考試的作文測驗中，為求公平與一致性，一篇文章往往需要經過多位閱卷老師所批改，再由各閱卷老師的批閱結果中做更進一步考量；但不同的老師有著不同的想法與看法，若非經過標準化的訓練，很難做到閱卷公平的一致性。因此，本研究嘗試建立一套自動化系統，藉由觀察文章特性，結合自然語言與貝氏機率學習法，建立一系統模型，用以提高批改作文時的可靠性，並改善閱卷時的工作效率。

1.2 研究假設

在觀察文章特性的過程中，我們認為文章的外在特徵，雖然無法直接作為評分的標準，但對於閱卷老師的評分結果確具有一定的影響力。例如，字數較多的文章雖不見得必定屬於高分範疇，但相較之下，字數過少則絕不能成為高分文章；而從文章的結構方面來看，一篇好的文章，內容必定包含「起、承、轉、合」四個段落，相較於低分文章分段過多或過少，則有著明顯的區隔。據此觀念，我們擷取文章的多項直接與間接特徵，作為機器學習法所需的評分依據，並設計更有效的學習方式以提升系統準確度。

1.3 研究目的與構想

藉由觀察文章的特性，本研究欲透過文章的多項直接與間接特徵，包括文章內使用的概念數、字數、名詞數、句號數等多項特徵，建立一結合 Bayesian Theorem 的系統模型，且該系統能自訓練資料（Training Data，即評閱過的作文）中自動建立評分規則，並利用此自動產生的規則對於其他作文進行評分的工作，同時亦可將此系統與其他不同的 AES 系統做結合，以改善系統效能及評閱時的正確性與可靠性。

1.4 論文架構

第一章為前言，內容主要為說明研究的動機目的、作文的重要性及系統設計時所採用的構想。第二章為相關研究，內容包括英文自動作文評分系統的發展，及貝氏機器學習法的理論基礎。第三章將介紹貝氏機器學習法中，各項類別屬性的擷取方式。第四章則詳細介紹系統的核心架構及說明。第五章為實驗實作及研究結果的呈現與分析，並透過流程圖瞭解整個實驗過程。第六章為描述本篇論文的研究總結，以及未來尚需繼續研究的相關工作。

第二章、相關研究與構想

在本章節中，將詳細的介紹本論文提及的相關研究與構想，首先概略的介紹英文的作文評分系統，隨之介紹設計系統的理念與想法。本章的章節順序安排如下，首先在 2.1 節簡單的介紹英文作文評分系統 e-rater 的發展原理。接著在 2.2 節詳細描述貝氏機器學習法（Bayesian Learning Theorem）所使用的概念與其機率模型。最後，在 2.3 節概述中文斷詞處理的重要性。

2.1 e-rater

GMAT(Graduate Management Admission Test)是美國商業學校入學測驗考試，這項測驗中除了與商業相關的筆試，同時也包含作文評量項目，其中作文的分數是以六級分制為主，最低分為一分，最高分為六分。在早期的閱卷過程中，一篇文章的評比必須經由兩位閱卷老師來進行評鑑工作，當兩位老師的評分結果相差超過一分時，則必須經由第三位閱卷老師來進行裁定任務；換言之，一篇測驗文章的分數依據，至少須有兩位閱卷老師達成一定程度上的共識。到了 1999 年 2 月，參加 GMAT 測驗的人數日漸增多，相對之下，所需作文評閱的工作量也大幅提升，於是引進了 e-rater 系統來代替初期所須兩位閱卷老師的其中之一，其主要原因為經由 e-rater 系統所評定的分數結果中，其 92%以上的結果皆與實際閱卷老師所批改的分數相差位於一分之內。事實上，這個統計結果與真實情況中兩名受過訓練的閱卷老師相比，兩者之間的誤差比率是非常相近的。

而在 e-rater 的系統在進行評閱的過程中，總共包含了三個模組：結構(structure)、組織(organization)、內容(content)。其中，結構模組主要的工作為分析句法的多樣性，搜尋有意義的詞組，進而判別不定詞、成語、慣用語、以及完整或從屬子句等分析任務。接著，由組織模組負責分析句法中的主要概念，包括句子與句子之間的轉折詞或連接詞，以及句中所使用的修辭結構等。之後，交由內容模組進行評估文章中所使用的字彙是否能反應文章內容與主題的相

關性。最後，對於各個模組所提供的資訊進行整合，e-rater 方可評定測驗文章的分數並提供少許的反饋訊息。

2.2 Bayesian Theorem

貝氏機器學習法是一個基於機率理論的分類方法，此方法假設所有的輸入屬性(attribute)間彼此獨立，而且各項屬性皆具有同等的重要性，雖然這樣的假設有失公允，但在實際運用上仍可提供可接受的結果。這個方法在一般機器學習領域中，相當廣泛而普遍的被應用於各類問題中。

一般而言，貝氏機器學習法的作法為，在特徵選取後，由已知資訊中計算出該特徵與該類別之間的條件機率關係，對資料提出事前機率分配（prior probability），再依所搜集到的資訊，來修正事前機率分配，使之成為事後機率分配（posterior probability），再由其中選出機率最高的可能性來作為預測的依據。說明如下

$$P(H \mid d) = \frac{P(d \mid H)P(H)}{P(d)}$$

其中：

P(d)：prior probability of obtaining data d

P(H)：prior probability that H is correct

P(d|H)：probability of obtaining data d if H is correct

P(H|d)：posterior probability that h is correct

例：某校畢業校友之性別與從事職業類別調查結果顯示男性有 80%從商，女性有 50%從商，而調查樣本中男性比例佔 30%，女性比例佔 70%，請問若已知某校友從事行業為商業，則此人為男性的機率為何？

依此問題，各項機率可化簡如下：

obtaining data d：從商

P(d)：所有人從商的機率

$$P(d) = 30\% \cdot 80\% + 70\% \cdot 50\% = 59\%$$

P(H)：該校友為男性的機率

$$P(H) = 30\%$$

P(d|H)：該校友為男性且從商的機率

$$P(d|H) = 80\%$$

P(H|d)：該校友從商且為男性的機率

$$P(H|d) = \frac{80\% \times 30\%}{59\%} = 40.7\%$$

2.3 中文斷詞處理

所謂「中文斷詞」，是將一連串的中文「字」，轉換成「詞」的組合，其中每個「詞」是由「一個字」或者「多個字」所組成。如何將由「字」所組成的「句子」切割成一個個的「詞」，則是中文斷詞的主要使命。

斷詞是中文語言處理中最基礎的工作。中文的句法(syntactic)和語意(semantic)基本單位是「詞」，單獨的中文字未必是語句分析的最小單位。因此，任何中文自然語言處理，例如：檔案檢索、中文輸入、光學字體辨識、語音辨識、機器翻譯等，都需先對中文句子進行「斷詞」，才能進行下一步的處理。由於斷詞結果的正確性及完整性對後續的處理動作有關鍵性的影響，這使得中文斷詞變成一件非常重要的工作。相對於歐美語系國家，句法和語意基本單位是「字(word)」而非「詞」，雖然每個字都是由多個字母組成，但在字與字之間都有明顯的空白作為分隔，所以如何判斷出單字則屬較為簡單的工作。換言之，這類型的語言並不存在斷詞的困擾。

第三章、特徵擷取

作文中的特徵可分為直接特徵與間接特徵兩種,其中間接特徵又稱為表面特徵。閱卷老師在評閱作文時所直接觀察的特徵謂之直接特徵，包括文章的修辭、內容、語義等特徵皆稱之，此類特徵較能直接反應文章語義層面上的好壞。然而從閱卷者的角度來看，間接特徵雖不能直接用以評定文章內容好壞，但對於評閱者及評分系統卻是相當重要的參考數據。因此本論文採用兩項直接特徵與四項間接特徵作為貝氏學習法所需的類別屬性。我們將在本章中說明擷取特徵的方式及理由，詳細的系統架構，將詳述於下一章節。

3.1 直接特徵的擷取

本系統所使用的直接特徵共有兩種，其一為文章內概念數的使用數量，其二為文章所使用的口語化程度，將分述如下：

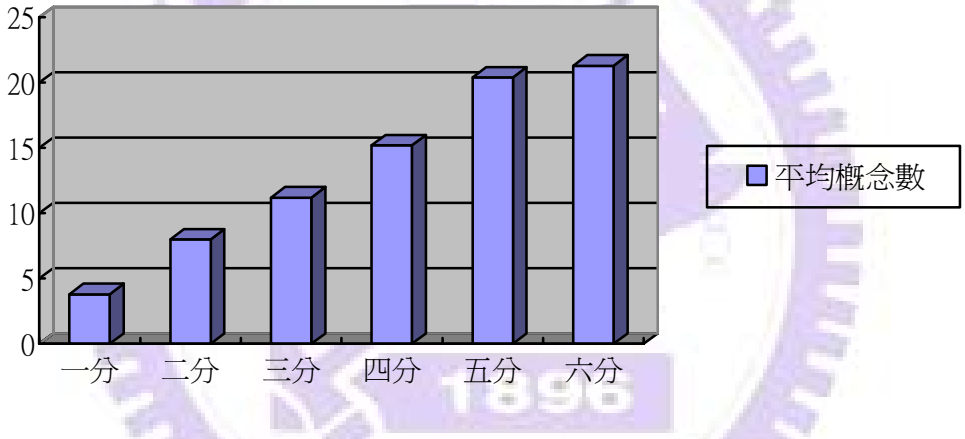
3.1.1 概念數

在[7]中，作者利用學生撰寫作文過程中所運用的概念組織，來決定文章內容的優劣，而用於描述一個「概念」的最小意義單位謂之「義原」。在本系統中，利用該方法，在訓練文章中產生一組具有鑑別力的義原集合，又稱「概念子集合」，其中，經過篩選出來的義原，意謂著高分文章中常見且鮮於低分文章中出現之義原，這些義原又稱為「好義原」。接著將所有的測試文章經過中文斷詞後，將內容「概念化」(Conceptualize)，令所有斷好的詞轉換為義原，並根據各篇文章所產生的義原集合，進行統計該集合中出現好義原的數量，以作為文章概念數的特徵；但由於義原是根據詞彙所進行的轉換，不同的詞彙也有可能會對應到同一個義原，如「教室」、「客廳」等中文詞其義原皆為{room|房間}；這樣的統計方式並無法確定該篇文章的好義原是由哪些詞彙所對應，因此某些文章中雖出現好義原，但其對應的詞彙卻為不雅用詞：例如，某測試文章中有一句子為，『(有時候)(在)(十)(分鐘)(的)(下課)(裡)，(大家)(互摸)(胸部)』，其中「胸部」一

詞可對應至好義原{part部件}，但該詞在一般用語上並非適當用法。有鑑於此，本系統對於統計文章好意原數量的方式加以修整；在訓練產生概念子集合時，同時紀錄所有轉換為好義原的詞彙，因此，對於測試文章好義原的定義存在一必要條件：該詞彙所對應的義原須存在於概念子集合中，且該詞彙必須存在於此義原的詞彙列表中，如下所示：

$$Good\_Def = \{\ DEF(x) \mid DEF(x) \in ConceptSubset,\ x:term\ of\ essay \in termList\ of\ DEF(x)\ \}$$

經此一調整後，我們由訓練文章中統計概念數量與寫作評閱之間的關係圖表，如圖 1 所示。



圖表 1：平均概念數

3.1.2 口語化程度

作文與演講最大的差異在於「口語」的表現方式；在演講的過程中，為了能讓所有聽眾清楚瞭解演講者所傳達的訊息，演講方式大多以「談話性」的演講技巧作為與聽眾間溝通的方式。但相對於作文來說，寫作是一嚴謹的思考過程，文章的內容必須條理分明並避免口語化的使用，以下表1中『今天天氣真是熱斃了』（以下簡稱A句）與『今天天氣非常炎熱』（以下簡稱B句）兩句子為例，A句的寫法較貼切於一般大眾的平時對話，因此較易引起聽眾的興趣，適於談話性質高的演講場合中；相較之下，B句的寫法在修辭上較A句顯得嚴謹且完整，因此適於結構完整的作文寫作中。表格1所示為演講與寫作方式不同的比較：

演講方式	寫作方式
今天天氣真是『熱斃了』。	今天天氣非常炎熱。
平時都在玩，考試卻考了第一名，真是『電電吃三碗公』，看不出來。	鮮少讀書的他，在這一次的考試中得到了第一名，真是令人大感意外。
他在班上的人緣『超好』。	他在班上的人際關係非常好。
他這人，『超搞笑』。	他是個非常幽默的男生。

表格 1：演講與寫作方式比較

由表1得知，在演講的過程中，為了引起聽眾的興趣及注意，演講方式是以較為輕鬆、活潑、詼諧的口語方式為主，但這樣的作法較不適用於寫作技巧中。在本系統中，為了分辨口語化的寫作技巧，本系統利用在斷詞上一簡單概念來加以區別。我們認為在一個句子中，最小的單位詞字數至少為二字以上，以A、B兩句為例，此二句型經過中文斷詞後所得結果為：

(a)今天 天氣 真是 熱 斃 了

(b)今天 天氣 非常 炎熱

依上述結果所示，我們發現在口語化的句型中，斷詞後的結果所含有的『單字詞』比例較結構完整的敘述句來得高；根據觀察指出，這類單字詞的種類主要可分為「流行語」、「錯別字」、「特定詞性」三類：

I. 流行語：隨著傳播媒體的發達與網路的普遍應用，在日常生活對話中出現許多流行術語。例如：「粉好笑」、「伊媚兒」、「很俗」等，這類新形態的語言呈現方式，除了打破正式作文的文法與句型外，尚加入許多象徵性符號來描繪說話的語氣或表情，也因而使得文章內容呈現高度口語化的現象。但這類非正式書寫文字的流行新名詞並不存在於傳統字典中，因此經由斷詞系統處理後，這類詞彙皆會被斷為單字詞。以上述為例，經由斷詞系統處理後，結果分別

為(粉)(好笑)、(伊)(媚)(兒)、(很)(俗)；這樣的斷詞結果顯示出單字詞與流行語之間的相互關係，因此在本系統中，透過單字詞的補捉，以作為口語化特徵。

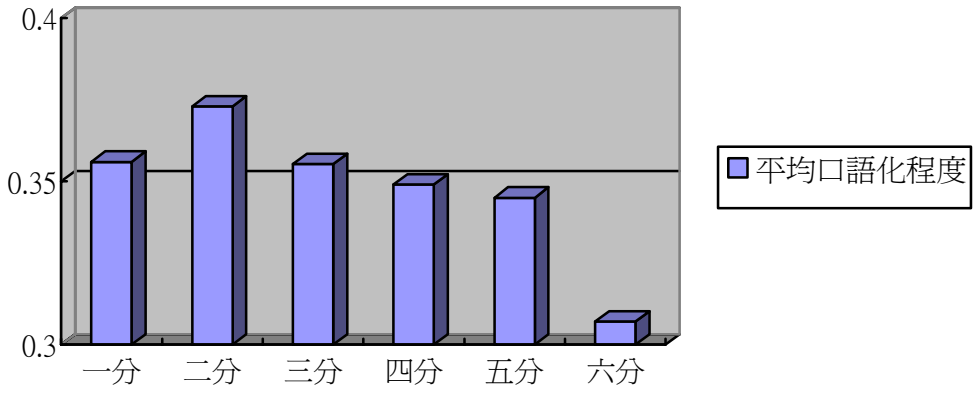
II. 錯別字：錯別字是學生寫作時常見的語文弊病，當文章中出現錯別字時，即間接影響斷詞的結果。例如：「唾手可得」本為一正確成語寫法，當寫作者將其誤寫為「垂手可得」時，斷詞的結果會因錯別字的出現而展現出不同的結果，分別為(唾手可得)與(垂)(手)(可)(得)。

III. 特定詞性：根據斷詞結果顯示，連接詞、介係詞、位置詞及代名詞等語義層面較低的詞類，普遍且大量出現於各類文章中，而此類詞性經由斷詞處理後，大多屬於單字詞。例如：「我在教室裡」一句，在經由斷詞處理後，斷詞結果為(我)(在)(教室)(裡)，其中根據中研院平衡語料庫詞類標記集所示，在此斷詞結果中，被斷為單字詞的(我)、(在)、(裡)分別為語義層面較低的代名詞、介係詞、位置詞；然而此類詞性皆為常見的寫作用法，在各類文章中隨處可見，因此當本系統加入此類單字詞作為口語化特徵計算時，並不影響其優異效果。

具此觀念，我們定義一口語化程度公式如下，作為本系統所使用的學習特徵之一：

$$\text{口語化程度} = \text{單字詞個數} \div \text{文章字數}$$

依上述公式可知，在斷詞文章中，若出現單字詞的比例愈高，則顯示口語化程度愈深，換言之，寫作手法即愈差。依此推論，我們由訓練文章中統計口語化程度與寫作評閱之間的關係圖表，如下圖2所示。



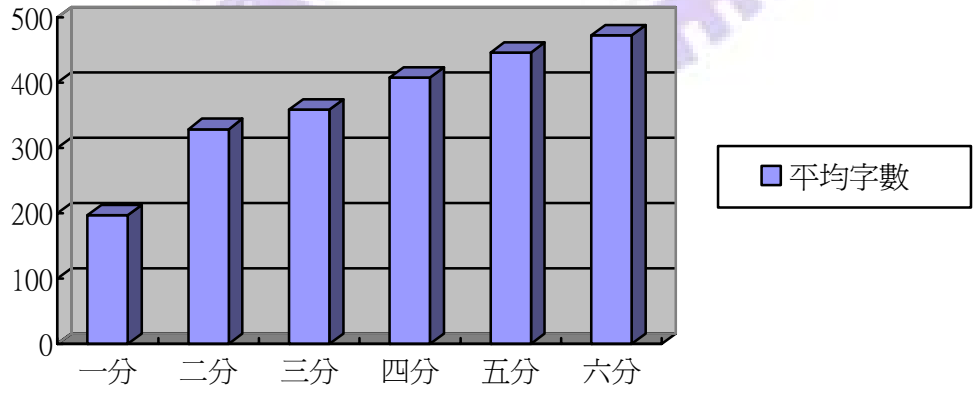
圖表 2：平均口語化程度

3.2 間接特徵的擷取

本系統所使用的間接特徵共有四種，分別為文章字數、主題數、名詞數量以及平均段落字數，各項特徵擷取方法將分述如下：

3.2.1 文章字數

文章字數的多寡，雖不能成為評分的主要訴求，但在閱卷者評分的準則中，欲成為高分文章，對於字數，卻往往存在著一定程度的要求。因此，本系統以文章字數作為一類特徵，並由訓練資料中觀察文章字數與評閱分數之間的關係。如下圖 3 所示：

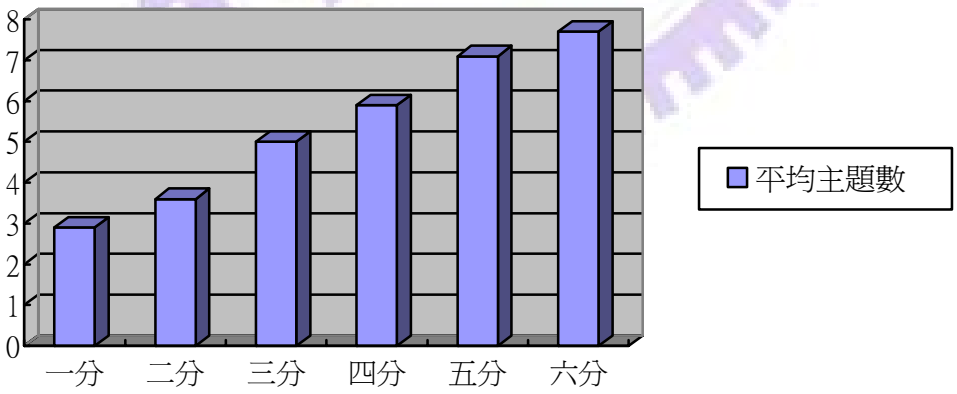


圖表 3：平均字數

3.2.2 主題數

從寫作者的角度來看，我們認為學生寫作時，思考的概念是以階層性觀念為主。仔細點說，當寫作者開始寫作，得知作文命題（topic）後，便開始構思佈局在文中所要撰寫的相關議題，這些相關議題主要是以作文命題為中心，一層一層慢慢的向外延伸擴張，當寫作者所能提及的相關議題越多，即意謂著作文的內容題材越趨豐富；反之，若寫作內容的相關議題過少，即表示文章內容老是繞著同樣的題材打轉而顯枯燥乏味。因此，我們設計一個方法來尋找文章中所描述的相關議題數量。

標點符號是書面語中用來表示停頓、語氣以及詞語性質作用的符號，是一種用於輔助文字、紀錄語言的符號。正確使用標點符號，對於準確表達文意有積極的正面意義。其中，句號的用法為表示陳述句（述說一種事情的句子）的停頓，即意謂著寫作者對於欲描述之議題結束呈述之用。因此本系統以文章中所出現的句號數量來代表文章中所描述的相關議題數，並由訓練資料中統計每篇作文所使用的主題數量與批閱者對其評分的相互關係。如下圖 4 所示：



圖表 4：平均主題數

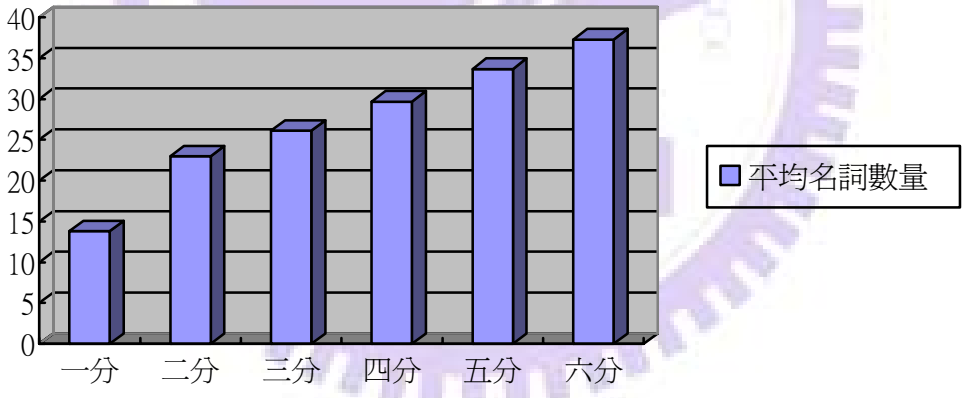
3.2.3 名詞數量

在主題數量的介紹中，我們認為寫作的手法具有階層性的特質，寫作者以作文命題為中心，依序撰寫其他議題，藉以抒發心中的構思。但是，若只有單純的

統計主題數量，卻不加以衡量該議題是否言之有物，這樣的作法猶如欠缺臨門一腳，無法令人信服。從寫作者的角度來看，文章中每一個不同的議題呈述，必定存在一個或一個以上的主要概念，根據觀察，這些主要概念絕大部份皆以名詞為主。因此，我們以文章中所提及的名詞主體為輔，用以檢驗主題數量是否足以搭配名詞主體的使用量。

根據中研院平衡語料庫詞類標記集所示，名詞的詳細分類共有十三種，但在本系統中，我們真正給予計量的只有普通名詞（Na）與地方詞（Nc）兩種，並規定屬於這兩種類型的詞類字數須為二字以上才予以計算。舉例來說，「老師(Na)、同學（Na）、教室（Nc）、人（Na）、事（Na）」等五個名詞皆屬於普通名詞與地方詞的類別，但相較之下，「人（Na）、事（Na）」這兩個單字詞則顯得較無鑑別力，因此不予計算。

下圖 5 所示為作文中所使用名詞主體數量與評閱分數之間的關係：



圖表 5：平均名詞數量

3.2.4 平均段落字數

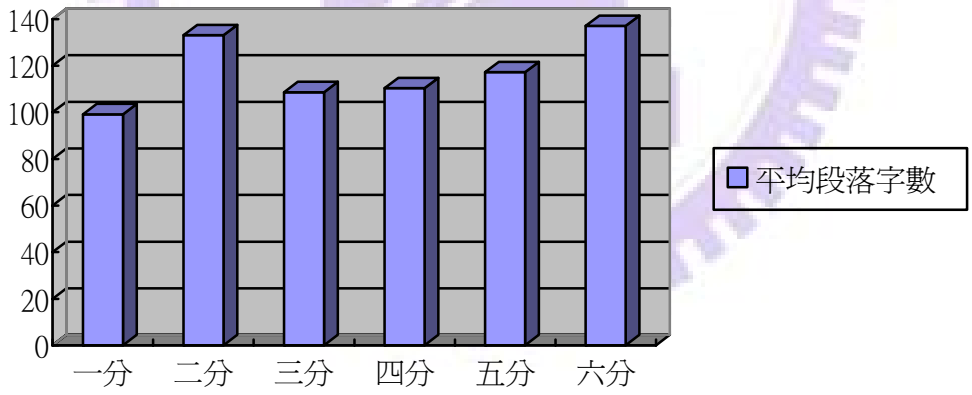
古人講究作文的章法，把文章的佈局，分做「起、承、轉、合」四個部分，又稱「四段章法」。而根據觀察，一篇字數約為四百字左右的文章，最好的分段方式為三段或四段，這樣的分段方式與古人講究的文章佈局不謀而合。但是，若只單純的以分段數量作為分類依據，著實為不明智作法，因此本系統進而改良以平均段落字數作為貝氏機器學習法的另一項類別屬性，希望達到文章的分數與平

均段落字數具有高度相關，因此設計一簡單公式來改善原有公式的盲點。

依照字面上的意思看來，平均段落字數計算公式應為：

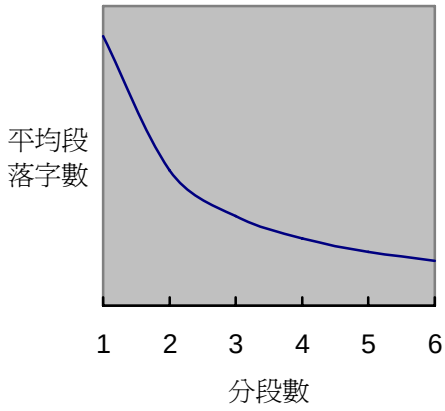
平均段落字數 = 字數 ÷ 分段數

但這樣的計算方法卻存在著一個矛盾的問題。當不同的文章，在字數皆相同的情況下，若文章分段數量越少，依此計算方法所得到的平均分段字數越多。例如，同為四百五十字的三篇文章，第一篇分段數為兩段，第二篇分段數為四段，第三篇分段數為六段，從分段方式的角度來看，分段數為四段的文章應具有最高的平均段落字數，但依照原計算方式，所得結果並非如此，因此須對原計算方法做適度的改良。圖 6 所示即為依照原公式計算所得統計圖表。根據此表，得知若依原公式計算，低分文章平均段落字數較高分文章高實屬不合情理；其原因在於，多數低分文章的分段方式較高分文章差，分段方式多屬一段或兩段，在平均段落字數的計算上因分母較小，故所得數值較高。



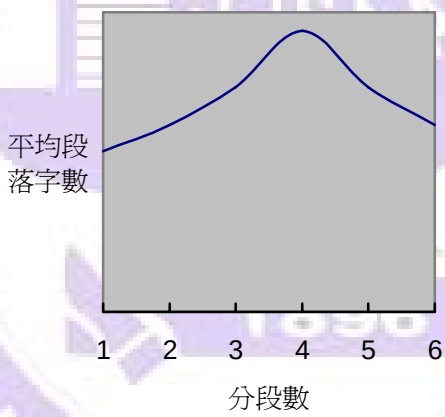
圖表 6：平均段落字數

依原本的公式計算方法，我們可定義 A(x) = w ÷ x，其中定義 A(X)為平均段落字數，w 為字數，x 為文章分段數；當字數(w)相同時，A(x)與 x 成反比關係，分段數越小，所得平均段落字數越大，其關係圖如圖 7 所示：



圖表 7：原始平均段落字數

但在作文的構思佈局中，我們認為，較佳的分段方式應將文章分為「起、承、轉、合」四個段落；換言之，具有較高的平均段落字數特徵應落在分段方式為四段的文章上，因此我們所希望得到的關係圖表如圖 8 所示。



圖表 8：理想平均段落字數

為達此一目的，本系統透過簡易的數學推理證明，對原本的公式進行適度的修改調整。在原本的計算方法中，分段數量小意謂著分母(x)較小，則依此計算方法所得平均段落字數大，欲改善此結果，需對分母做些許的修正。根據觀察，分段數小於兩段是一種較差的分段方式，較適當的方式為將文章分為四個段落來撰寫；因此當分段方式不理想時，必須將 x 調高至一適度數值，避免因分段數小造成平均段落字數高的反效果；我們以四段(x=4)作為最佳寫作分段方式的標準，過多或過少的分段方式皆屬較差的作法。

當分段數小於四段時，定義一常數 k，令 x’ = k-x（k > x）取代原有 x

作為新的分母，則新公式可改寫如下：

$$A(x)=\begin{cases} w\div x'\ldots if \quad x\le 4 \\ w\div x\ldots others \end{cases}$$

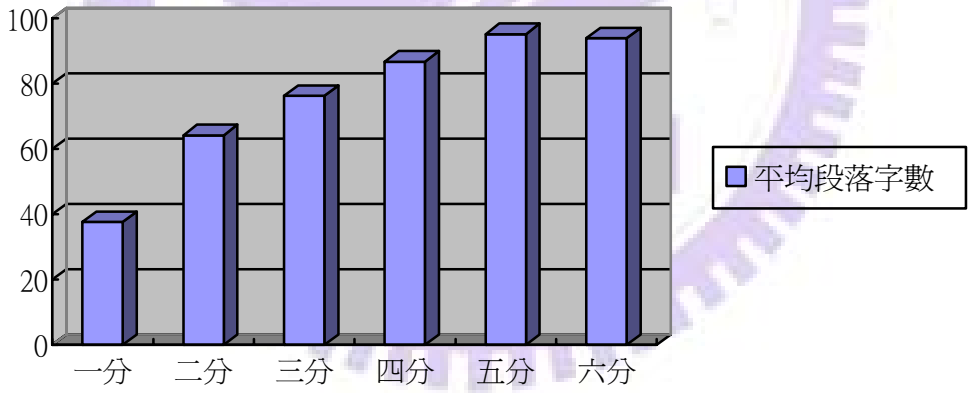
根據觀念，分段數為四段的文章應具有最高的平均段落字數特徵，此即意謂著 x 與 x’ 的最小值為 4，由此得知

$$\begin{cases} x\ge 4 \\ x'\ge 4 \\ x'=k-x \end{cases} \Rightarrow k\ge 8$$

依上述結果及實驗觀察得知，當 k=8 時，此特徵具有明顯鑑別力，因此經由修正後，在本系統中所提及的平均段落字數計算方法修正為

$$A(x)=\begin{cases} w\div(8-x)\ldots if \quad x\le 4 \\ w\div x\ldots\ldots\ldots others \end{cases}$$

根據上述計算公式，圖 9 所示為作文中所使用平均段落字數與評閱分數之間的相互關係：



圖表 9：平均段落字數

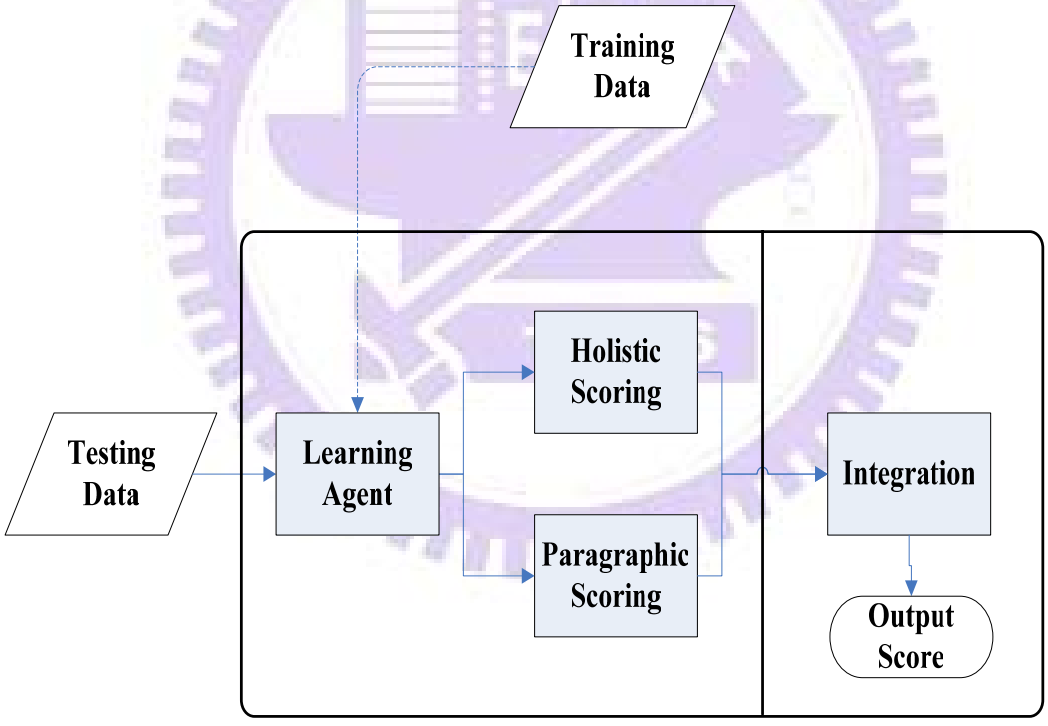
3.3 直接特徵與間接特徵的特性

由以上六個圖表可以明顯觀察到，其分佈圖的呈現與作文分數等級，除口語化特徵為高度負相關外，其餘均顯示出高度正相關的證據。但是，若單純的依據各類屬性的數量使用來評閱作文，預期結果可能是相當不理想的。因此，本系統以貝氏機器學習法為基礎，透過貝氏機率演算法，以上述六項獨立特徵作為系統的評分依據。

第四章、中文評分的系統設計

在本章節中，將詳細介紹系統架構及設計原理；在 4.1 節中，以系統架構圖來幫助瞭解系統主要的執行流程，其中主要的執行流程分為四個模組，1.Learning Agent-學習機制、2.Holistic Scoring-整體評鑑、3.Paragraphic Scoring-分段評鑑、4.Integration-評鑑整合，將分述其主要執行內容於以下四個小節中。

4.1 系統架構



圖表 10：系統架構

如圖 10 所示，在本系統中，主要包含四個模組：

1. Learning Agent - 學習機制
2. Holistic Scoring - 整體評鑑

3. Paragraphic Scoring - 分段評鑑

4. Integration - 評鑑整合

待蒐集完 Training Data(訓練資料，即為人工評閱過的作文資料)後，第一步，透過 Learning Agent（學習機制）訓練出貝氏機器學習法中各項特徵的門檻值及機率值。第二步，輸入測試文章，藉由 Holistic Scoring 及 Paragraphic Scoring 模組分別評鑑文章的整體分數與文章的各段分數。第三步，利用 Integration 模組針對文章的整體分數與各段分數做邏輯判別，並加以整合作為系統最後的評分依據。

4.2 Learning Agent - 學習機制

在此模組中，主要工作內容有二，首先是篩選特徵的門檻值 (Threshold)，接著依據所挑選出的門檻值，進行各類特徵屬性(attribute)的機率計算。

4.2.1 篩選門檻值

透過門檻值的篩選，分別挑選出五個門檻值 T_1、T_2、T_3、T_4、T_5，用以區分各文章類別層級範圍。在本系統中，採用一套簡單且迅速的挑選方法，首先擷取所有訓練資料的各項特徵，並將其特徵屬性值依大小排序後，再依各類文章所佔比例來設定門檻區間。以字數為例，若訓練文章中一至六分文章分別為 n_1、n_2、n_3、n_4、n_5、n_6 篇，則在由少至多排序好的字數陣列中，我們可依序得到五個門檻值，分別為：

T_1 = Sort_Attr[n_1]

T_2 = Sort_Attr[n_1+n_2]

T_3 = Sort_Attr[$n_1+n_2+n_3$]

T_4 = Sort_Attr[$n_1+n_2+n_3+n_4$]

T_5 = Sort_Attr[$n_1+n_2+n_3+n_4+n_5$]

因此，若某一文章字數為 w，欲判斷該文章等級為何，則查詢 w 所屬區間，

若 $T_{i-1} < w \leq T_i$，則此篇文章屬於 i 級分。

4.2.2 屬性機率值的計算

根據上述方式所訂定的門檻值，對於文章的評比並非絕對，我們只能認為「可能性較高」，但若單以這樣的方式作為分類依據，則略顯粗淺。以字數為例，根據 4.2.1 的方式，計算出屬於四分文章類別的字數門檻區間為 389~493 字，但在實際情況中，位於此門檻區間的文章不盡然全屬於四分類別，其中也包含了來自其他文章類別的可能性。

為了改善此一缺點，在計算各類特徵門檻值時，同時統計各類文章特徵屬性的最大與最小值，用以作為屬性機率值計算的重要依據。根據下表 2，以概念數為例：

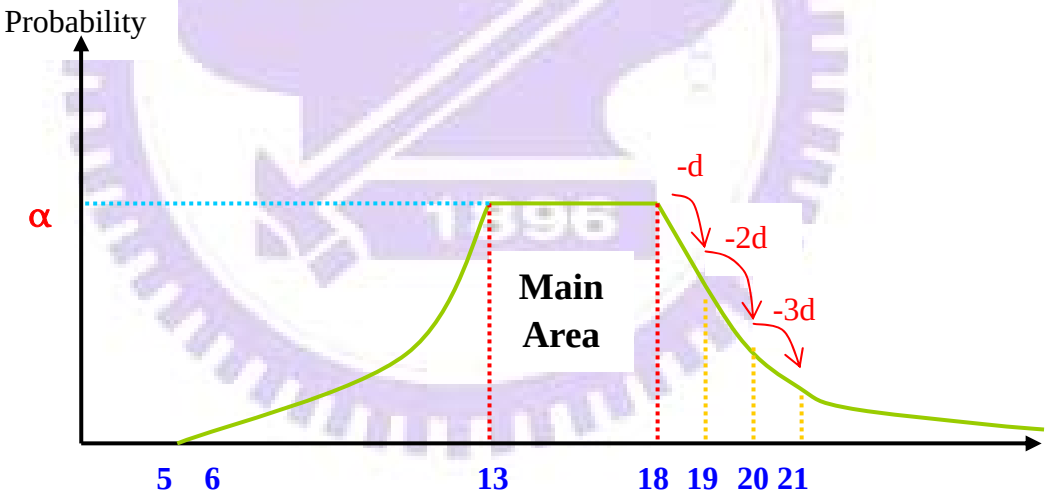
經過計算，如表 2 所示，以一分類別文章為例，其概念數門檻區間為 0 至 3，但在一分類別的訓練文章中，概念數的最小及最大值分別為 1 和 11；依此類推則可得知各文章類別所屬門檻區間及其對應的 min-Max 值。

	概念數門檻區間	min - Max
1-point	0 ~ 3	1 ~ 11
2-point	4 ~ 8	2 ~ 18
3-point	9 ~ 12	4 ~ 29
4-point	13 ~ 18	6 ~ 28
5-point	19 ~ 29	11 ~ 35
6-point	30 ~	21 ~ 30

表格 2：概念數門檻與 min-Mx 區間表

下圖 11 所示為預測該篇文章為四分機率值之原理。在四分類別的文章中，min-Max 區間為 6~28，因此當文章概念屬性數量為 6~28 者皆具有成為四分類別

的可能性。且根據表 2，四分文章類別概念數量是以 13~18 為主要門檻區間，當文章概念屬性數量位於其中時，應具有最高機率值 α，此時 α 為一經驗參數；相對地，當概念屬性數量離中心區域越遠則成為該文章等級的機率值越低，但其機率的遞減方式並非呈等差數列，因每當屬性數量多一個等級，其差異越顯明顯，因此，在本系統中機率值的遞減是採取快速遞減的方式。以下圖為例，當概念數量為 19 時，其預測機率值為 α-d，若概念數量為 20 時，其預測機率值為 α-3d，若概念數量為 21 時，其預測機率值為 α-6d，以此類推。依此方式計算，當文章概念數量為 7 時，依序可得六個類別機率預測值，分別為 P_1=0.52，P_2=1，P_3=0.428，P_4=0.133，P_5=0，P_6=0，接著再對此六個機率值做正規化(Normalize)處理，可依序得到 P_1=0.239，P_2=0.44，P_3=0.2，P_4=0.07，P_5=0.02，P_6=0.02，其中以 P_i 作為預測該測試文章為 i 類別的機率值。

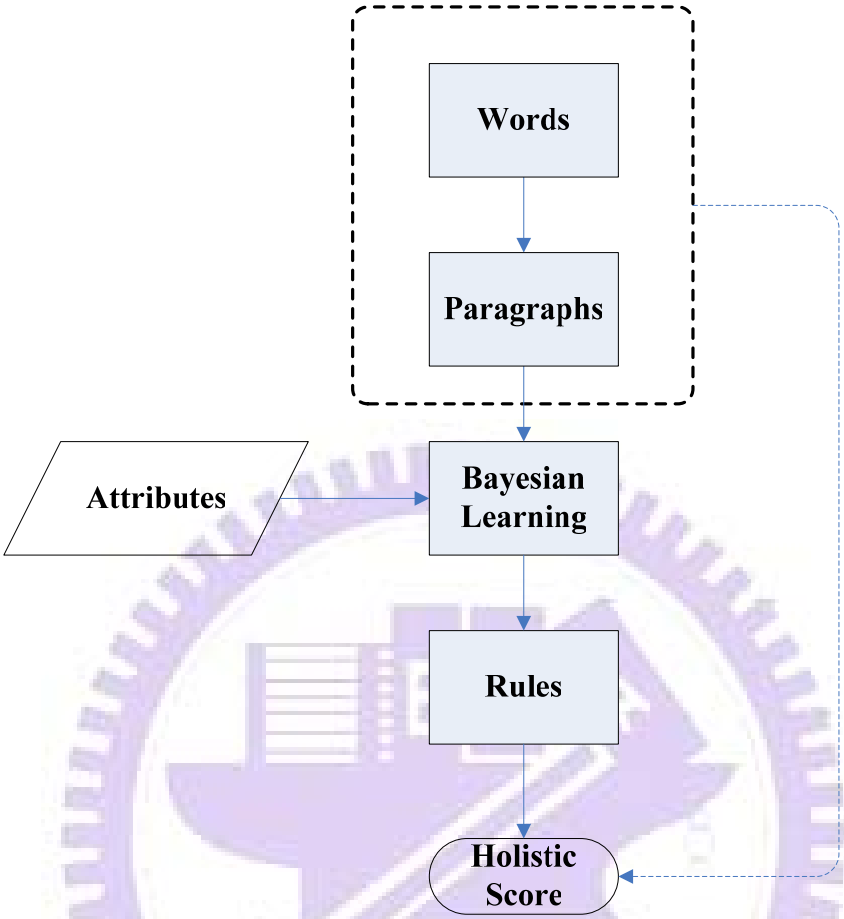


圖表 11：預測該篇文章爲四分機率值之原理

4.3 Holistic Scoring－整體評鑑

如下圖 12 所示，在此模組中，文章的整體評鑑分為兩部份：首先擷取出測試文章的各類特徵屬性(attribute)，並依貝氏機器學習法的機率值來評估文章的層級，接著再以系統制定的規則(rules)來做更進一步的文章層級確認，最後

則可得到一整體評鑑分數。



圖表 12：整體評鑑架構圖

4.3.1 貝氏機器學習法(6 項特徵)

在 2.2 節中提到，在貝氏機器學習法中，假設所有的輸入屬性間彼此獨立，且各項屬性皆具同等重要性。因此，當測試文章輸入後，即擷取系統所需的直接與間接特徵做為貝氏機器學習法的輸入屬性，包括概念數、字數、名詞數、主題數、平均段落字數及口語化程度等六項特徵。

根據公式，

$$P(H_i \mid d) = \frac{P(d \mid H_i)P(H_i)}{P(d)}，1 \le i \le 6$$

在本系統中，各項次代表意義分述如下：

$P(d)$：在所有訓練文章中，出現 d 屬性的機率

$P(H_i)$：測試文章屬於 i 類別的事前機率(prior probability)

$P(d|H_i)$：得知 d 屬性資訊後，預測該測試文章為 i 類別的機率值

$P(H_i|d)$：加入 d 屬性資訊後，該測試文章屬於 i 類別的機率值

例如：某一測試文章，屬於一至六分的事前分配機率分別為 0.1, 0.2, 0.5, 0.1, 0.05, 0.05，即 $P(H_1)=0.1$, $P(H_2)=0.2$, $P(H_3)=0.5$, $P(H_4)=0.1$, $P(H_5)=0.05$, $P(H_6)=0.05$；當得知其概念屬性數量為 7 時，預測此篇文章為一至六分的機率分別為 0.19, 0.41, 0.24, 0.13, 0.02, 0.01，即 $P(d|H_1)=0.19$, $P(d|H_2)=0.41$, $P(d|H_3)=0.24$, $P(d|H_4)=0.13$, $P(d|H_5)=0.02$, $P(d|H_6)=0.01$；在所有文章中，概念數量為 7 的文章佔全部數量文章的 0.15，即 $P(d)=0.15$；因此，當輸入其概念屬性時，此文章屬各類別的機率須更改為 $P(H_1|d)=\frac{0.19\times0.1}{0.15}=0.127$，$P(H_2|d)=\frac{0.2\times0.41}{0.15}=0.547$，$P(H_3|d)=\frac{0.24\times0.5}{0.15}=0.8$，$P(H_4|d)=\frac{0.13\times0.1}{0.15}=0.087$，$P(H_5|d)=\frac{0.02\times0.05}{0.15}=0.007$，$P(H_6|d)=\frac{0.01\times0.05}{0.15}=0.003$；最後將此六類機率值經過正規化處後可得，$P(H_1|d)=0.082$, $P(H_2|d)=0.348$, $P(H_3|d)=0.509$, $P(H_4|d)=0.055$, $P(H_5|d)=0.004$, $P(H_6|d)=0.002$。

依此方式輸入六項特徵屬性後，對於測試文章可得一機率分配值 P_1 ~ P_6，並由 P_1 ~ P_6 中挑選出最大機率值 P_i，以 i 類別作為該測試文章的預測分數。

4.3.2 規則

經由貝式機器學習法的機率計算，對於各篇測試文章，皆可得到一預測分數，但為了更進一步確定該篇文章是否確實屬於其預測類別，因此制定了下列規則以檢驗該文章是否符合該類別的特質：

I. 從作文的分段角度來看，較佳的寫作分段方式為『起、承、轉、合』四段，其中首段『起』代表文章的開頭，主要用意為提綱挈領，引領閱讀者進入主題，字數不宜過多。因此，本系統由高分文章的訓練資料中，訓練出一首段字數的門檻值(threshold)，當測試文章的首段字數超過此一門檻值，則酌情扣分。而以分段數量的角度來看，當分段方式過多或過少，皆為較差的寫作方式；而根據訓練資料的觀察，這樣的極端分段方式僅出現於低分文章，因此，當測試文章的分段數量過多或過少，則一律視為低分文章。

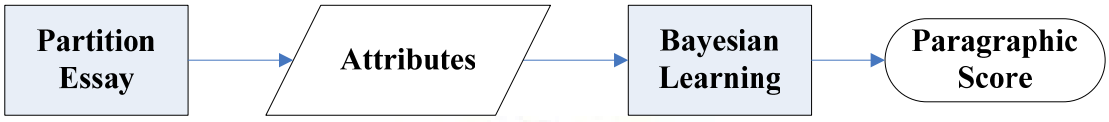
II. 文章內容的好壞，可由其名詞分佈略知一二；好的文章，對於各個主題間的組織描述，必為條理式的安排，而非毫無章法、胡亂拼湊一番；因此，在好的文章中，用以描述的主體名詞必定是較均勻分佈於各個議題中，而並非將所有的名詞全部擠壓於某一段落中。因此本系統由高分文章的訓練資料中，訓練出一段落與名詞間該有的比例界限，用以作為成為高分文章的必要條件。

III. 好的文章內容，其描述的手法必為多樣化，若能利用許多不同的角色串場則能豐富文章內容；但描述的手法若只是一直繞著同樣的主體打轉，文章則顯枯燥乏味。當文章中，同樣的名詞反覆出現多次，往往顯示該篇文章內容單調，仍待加強。因此本系統由高分文章的訓練資料中訓練出名詞使用重覆的比例界限，用以作為判定內容好壞的準繩。

IV. 一篇文章的字數多寡，雖不是閱卷老師最重要的評分依據，但卻深深影響閱卷老師的第一印象。在觀察的過程中，我們發現，字數多的文章雖不見得必為高分，但字數過少的文章大都屬於較差的文章。有鑑於此，本系統由低分文章中統計出一字數門檻，當文章字數低於此最低限度者，皆視為低分文章。

4.4 Paragraphic Scoring－分段評鑑

圖 13 所示為分段評鑑模組所進行的評分流程，首先將文章分成四段，並分別由此四段文章中擷取所須特徵，接著依其特徵屬性以貝式機器學習法來對各段文章做評估，最後即可得到該篇文章的分段評鑑結果。



圖表 13：分段評鑑架構圖

4.4.1 Partition

在文章分段的過程中，是以擷取各段 150 個字為目標，而擷取的段落終點是以標點符號為單位。換言之，所取得段落文章之第 150 字若非標點符號，則繼續往後取至標點符號後即截止。

其中，分段的方法如下：

第一段：由文章第 1 個字開始向後擷取；

第二段：以標點符號為單位，由文章起始處算起約略 75 個字後，開始向後擷取文章；

第三段：以標點符號為單位，由文章末端往回計算約略 75 個字後，往回擷取文章；

第四段：以標點符號為單位，由文章末端往回擷取約略 150 個字為止。

4.4.2 貝氏機器學習法(4 項特徵)

此一模組與 4.3.1 節中的設計原理相同，但所使用的特徵卻有些許的減少，原因為 Partition 模組中已將文章拆成字數約略相等的四個段落，因此『字數』與『平均段落字數』這兩樣特徵在此模組中並無作用，其餘的過程原理皆相同。

4.5 Integration－評鑑整合

在Integration模組中，最主要的工作為鑑別整體評鑑與分段評鑑間的搭配關係。從評閱者的角度來看，一篇結構完整的文章，必定具有良好的概念數量、主題數量、名詞數量等多項良好特徵，其整體評鑑分數也多屬高分，若抽取此篇文章任一段落觀察之，其分段評鑑結果也多屬高分；相對地，當文章整體評鑑屬低分類別時，其分段評鑑大多同為低分層級。據此想法，透過訓練資料，尋找高分與低分文章中，其整體評鑑與分段評鑑之間的搭配關係；經此訓練結果，對所有測試文章進行整體與分段評鑑分數的整合，以決定該篇測試文章之最後評比。

第五章、實驗過程與結果討論

在本章節中，將於 5.1 節說明本系統所使用的實驗資料來源。並於 5.2 節中說明主要實驗流程。最後於 5.3 節中探討實驗的數據結果。

5.1 實驗資料

本實驗中所使用的資料為臺北市敦化國中二年級學生所撰寫的作文，作文題目為『下課十分鐘』，可用資料共有 689 篇。這些作文是由人工建立電子檔，建檔的過程中保留學生原本的錯字及標點符號，以維持文章原貌。作文分數等級仿照 GMAT 作文測驗，採六級分制。每篇作文皆由二至三位老師所評閱，以維持閱卷分數可靠性。其中一至六分文章分別有 45 篇、128 篇、210 篇、208 篇、91 篇、7 篇，在本實驗中，選定一至六分文章各二分之一作為訓練資料，其餘二分之一作為測試資料用以評估系統效能。

5.2 實驗流程

所有作文皆經由「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」[5]進行文章的斷詞與詞性標記後，開始擷取各項特徵。本實驗共分兩大階段，首先，在系統訓練的階段，以 343 篇作文作為訓練資料，接著將訓練文章中各類特徵屬性依序排序完成，同時紀錄各類別間的 min-Max 區間，用以計算各類特徵屬性預測機率值。接著在測試階段中，以 346 篇的測試文章作為測試資料，此時系統根據訓練階段所產生的屬性預測機率值，將文章分為整體與分段評鑑進行評分後再加以整合，最後比較系統評閱的等級與原先所受的評閱等級差異，來計算系統的正確率。

5.3 實驗結果與討論

本次實驗中共計算二種正確率 Adjacent Rate、Exact Rate：

Adjacent Rate：允許一分誤差的整體正確率

Exact Rate　：毫無誤差的精準正確率

因受限於每位閱卷老師的背景知識、主觀認知與評量標準不盡相同之下，本實驗認為相差一分為可容許的誤差，在這一分的誤差範圍之下皆視為正確的評斷。

在本實驗中，針對現有的 689 篇作文中，於一至六分等級中選定各類文章之二分之一作為訓練資料，共計 343 篇訓練資料，其餘尚有 346 篇作文用以作為測試資料，實驗結果如下表 3 所示，並與[6][7]方法比較結果如表 4 所示：

評分結果 / 文章類別	1-pt	2-pt	3-pt	4-pt	5-pt	6-pt	Adjacent Rate	Exact Rate
1-pt (23 篇)	**18**	**4**	**1**	0	0	0	**95.7%**	**78.3%**
2-pt (64 篇)	**10**	**33**	**21**	0	0	0	**100%**	**51.6%**
3-pt (105 篇)	0	**17**	**53**	**32**	**3**	0	**97.1%**	**50.5%**
4-pt (104 篇)	0	0	**30**	**65**	**9**	0	**100%**	**62.5%**
5-pt (46 篇)	0	0	**1**	**27**	**18**	0	**97.8%**	**39.1%**
6-pt (4 篇)	0	0	0	**2**	**2**	0	**50%**	**0%**

表格 3：本系統實驗結果

	Modified ID3	Concept Method	Bayesian Method
Adjacent Rate	91.1%	92.30%	97.98%
Exact Rate	38.9%	46.89%	54.05%

表格 4：實驗結果

如表 3 所示，一分類別的測試文章共有 23 篇，其中 18 篇經由本評閱系統評為一分、4 篇評為二分、1 篇評為三分；二分類別的測試文章共有 64 篇，其中 10 篇評為一分、33 篇評為二分、21 篇評為三分；依此類推，可分別算出各類文章的 Adjacent Rate 及 Exact Rate。在允許一分的誤差下，一至五分類別文章的 Adjacent Rate 皆可達 95%以上的準確率；相較之下，在六分類別文章中，因有效樣本數較少，因此所得準確率較低；但從系統的整體效能觀之，在允許一分的誤差下，其整體的準確率可達 97.98%。根據表 4 所示，本系統與 Modified ID3 及 Concept Method 兩者相比，在整體的效能表現上，不論是 Adjacent Rate 或 Exact Rate，皆較另二種評閱系統優異，足以顯示本系統的可信度，因此適合用以作為老師評閱作文時的參考工具之一。

第六章、結論與展望

在本論文中，我們提出以貝氏機器學習法為基礎的作文評分方式，經由觀察統計文章中的直接與間接關係，我們發現文章的單一特徵雖無法直接決定作文分數，但將多樣性的特徵透過貝氏學習法的整合卻能凸顯其優異的評分效能。根據實驗結果所示，本系統對於作文的評閱擁有相當高的正確率，在允許一分誤差下的正確率(Adjacent Rate)可高達將近 98%，意謂著本系統與實際閱卷老師的批閱結果相當接近，可提供作為閱卷老師批改作文時的協助工具之一。

本系統提出一以貝氏機器學習法為基礎的中文評分系統，其中所需的訓練資料量過多，若欲應用於大考之中，稍嫌不便。盼未來能提出一套僅須少量訓練樣本即可達到同樣高正確率的系統模型。

參考文獻

[1] L. M. Rudner & L. Liang, Automated essay scoring using Bayes' theorem, National Council on Measurement in Education, New Orleans, LA.(2002)

[2] Jill Burstein, Karen Kukich, Susanne Wolf, Chi Lu, Martin Chodorow, Lisa Braden-Harder, Mary Dee Harris, Automated scoring using a hybrid feature identification technique, Proceedings of the 17th international conference on Computational linguistics(1998)

[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] Tsunenori ISHIOKA, Masayuki KAMEDA, Automated Japanese Essay Scoring System : Jess.(2003)

[5] 中央研究院資訊科學研究所詞庫小組中文斷詞系統
URL：http://ckipsvr.iis.sinica.edu.tw/

[6] 張佑銘, 中文自動作文修辭評分系統設計(2005)

[7] 蔡沛言, 自動建構中文作文評分系統：產生、篩選與評估(2005)