

工業工程專題

以文件特徵值為基礎之文章風格評斷模式

指導教授 侯建良 博士
學生姓名 高漢佑 李松霖 呂冠昇

一、研究動機與目的

研究動機

文件，是一種表達思想、知識與價值觀的一種方式，也在人類的生活中扮演了不可或缺的角色。從文章的著作與發行，到求職者找工作所用到的履歷表，人們都希望利用文件來傳達重要的訊息，卻也伴隨著許多問題有待解決，如下：

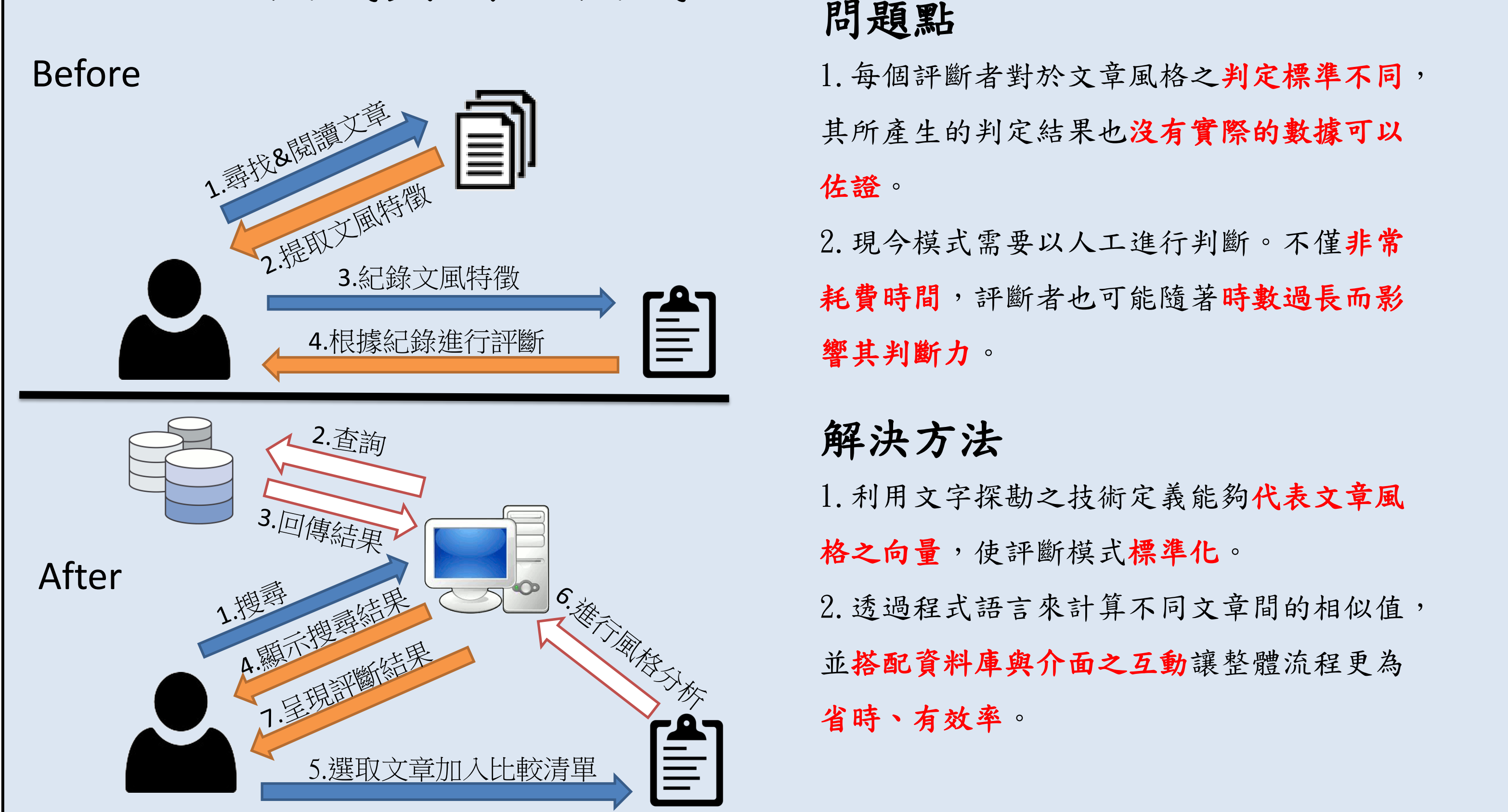
- 網路的便利性使得文章抄襲的現象增加
- 利用匿名的方式或冒用他人身分來發布文章

為了解決以上的問題，我們需要以一定的標準評斷文章的風格，但一般以人工閱讀文章後進行判斷的方法，時間花費過多且可能受到主觀影響，因此我們希望以程式語言建構一套系統，有效率且公平地對大量文章進行風格評斷。

研究目的

本研究利用文字探勘的技術，將屬於非結構化資料的「文章」轉換成結構化的「文件特徵向量」來代表一篇文章之「風格」。並透過「風格」相似度之比較來篩選出較可能來自同一作者或是較有抄襲嫌疑之文章，最後將結果提供給使用者參考。

二、一般模式與期望模式



問題點

1. 每個評斷者對於文章風格之判定標準不同，其所產生的判定結果也沒有實際的數據可以佐證。
2. 現今模式需要以人工進行判斷。不僅非常耗費時間，評斷者也可能隨著時數過長而影響其判斷力。

解決方法

1. 利用文字探勘之技術定義能夠代表文章風格之向量，使評斷模式標準化。
2. 透過程式語言來計算不同文章間的相似值，並搭配資料庫與介面之互動讓整體流程更為省時、有效率。

三、研究方法

取出特徵向量

向量一：停止詞
利用tf-idf演算法計算停止詞在各篇文章的權重。
$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$
$$idf(t) = \log \frac{1+n_d}{1+d(t)} + 1$$

向量二：句子長度

向量三：句子銜接

前一句	短	中	長
短	短-短	短-中	短-長
中	中-短	中-中	中-長
長	長-短	長-中	長-長

向量四：段落、引言、最大詞數
文章中段落、引言、頻率最高之詞的使用次數。
向量五：標點符號
文章中問號、驚嘆號、冒號、分號、引號與括號之使用次數。

取出特徵向量

1. 連接詞
2. 句長
3. 句子銜接
4. 段落、引言、最大詞數
5. 標點符號

餘弦相似度
$$sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

文件特徵	特徵相似值
停止詞	s_stpw
句子長度	s_sl
句子銜接	s_sc
段落、引言、最大詞數	s_pqm
標點符號	s_pt

綜合評價 = $s_stpw * s_sl * s_sc * s_pqm * s_pt$

文件相似值分群

文件相似值分群
K-means 分群法：

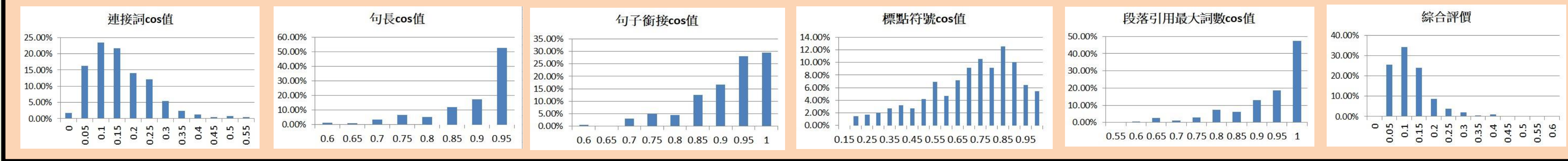
下圖為分群過程中，各群的變動。

與作者相似度分析

與作者相似度分析

餘弦相似度

利用餘弦分析算出 C_2^n 個各項特徵值，相乘得到綜合評價。以下為實驗29篇散文之各特徵值與綜合評價之統計圖：

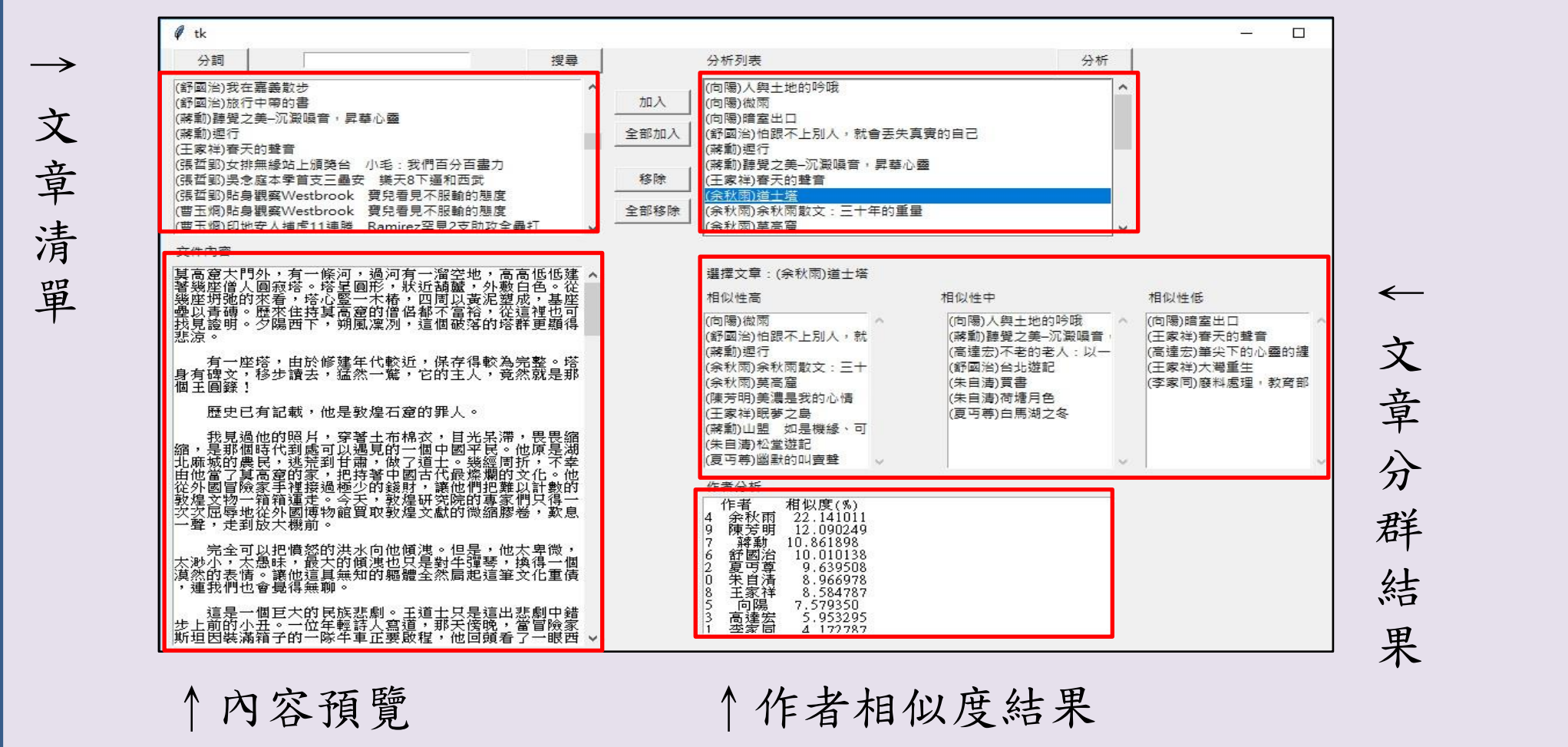


四、系統開發

使用工具：

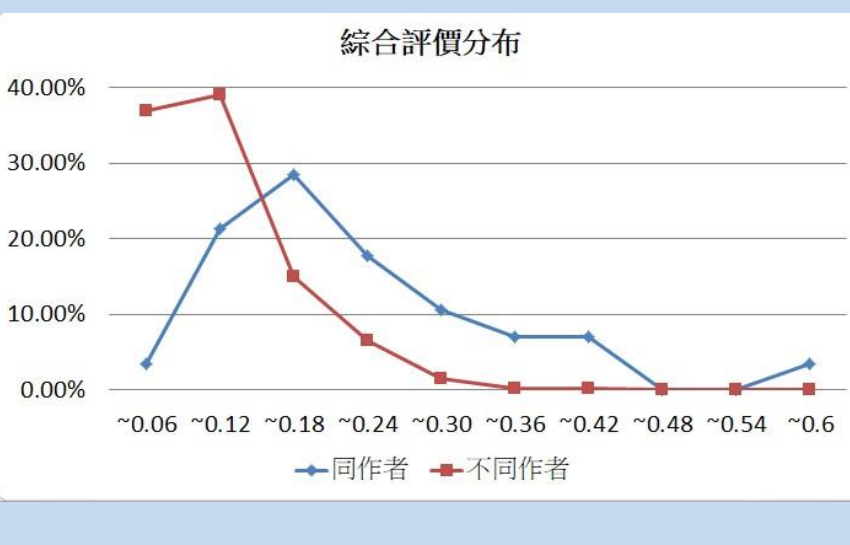


↓ 分析列表

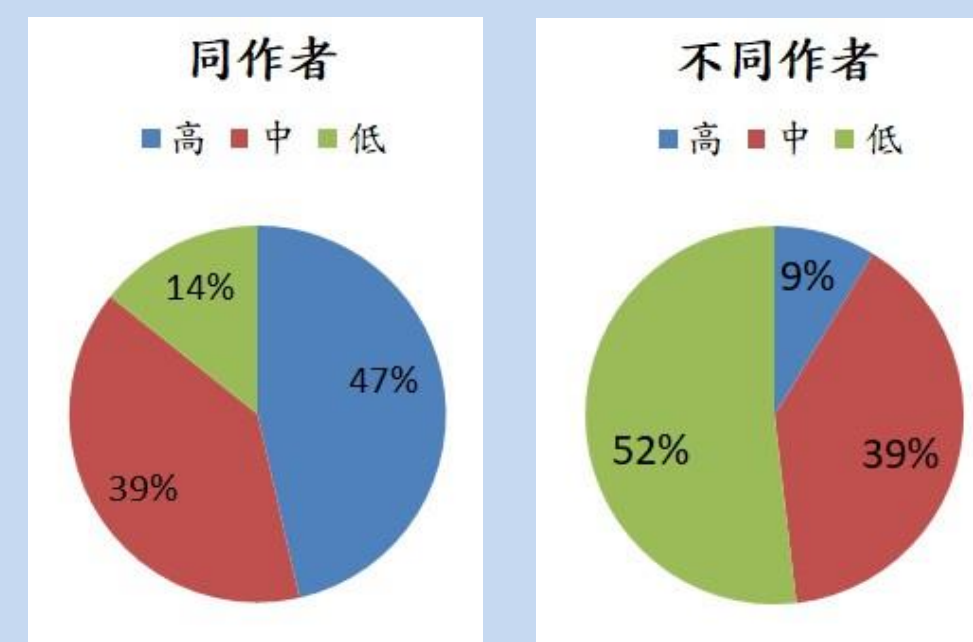


五、實驗結果

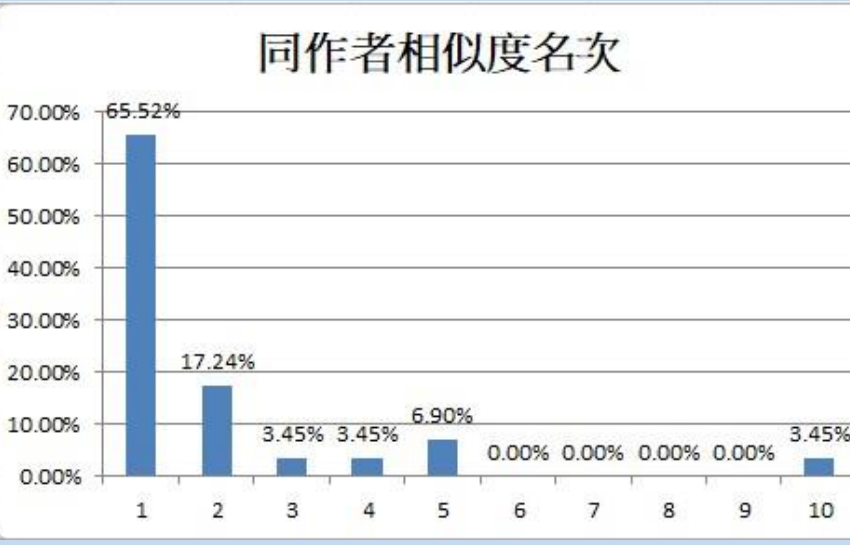
綜合評價分析



分群結果



作者分析結果



各圖所顯示的資訊都說明了本方法對於作者的判斷有一定的準確度，因此本方法可以有效地評斷文章的風格。

六、結論與未來展望

結論

本研究開發的系統明確定義了一套評斷文章風格相似度的模式，能夠將「文章風格」這種抽象的概念標準化。根據現有文章分析結果來看，該系統在評斷文章風格、作者身分與文章抄襲與否等方面有一定的成效。除此之外，最重要的莫過於該系統能夠在時間有限的情況下，對大量的資料進行分析並呈現給使用者。這也正符合「工業工程」所講求的成本最小化，效益最大化之概念。

未來展望

本研究中探討了文章風格與作者間確實有著一定的關聯性。而我們發現現今許多公司在雇用新員工時都會對其進行性向測驗來評斷其個性、適合的職位，而這樣的方法會使受測者能根據問題來進行相對應的回答來塑造「假個性」。因此，我們希望在未來能將文章風格與人格特質進行更深入的關聯性分析，並搭配現有的性向測驗來讓受測結果更加準確，讓每位員工分配到最適合他的位置。

