

國立交通大學

管理學院（資訊管理學程）碩士班

碩 士 論 文

運用文字探勘技術在社群行為上之人格預測

Using text mining to predict personality based on social
behavior

研 究 生：張曉珍

指導教授：李永銘 博士

中 華 民 國 102 年 6 月

運用文字探勘技術在社群行為上之人格預測

Using text mining to predict personality based on social behavior

研 究 生：張曉珍

Student：Xiao- Zhen Chang

指導教授：李永銘

Advisor：Dr. Yung-Ming Li

國立交通大學

管理學院（資訊管理學程）碩士班

碩 士 論 文

A Thesis

Submitted to Institute of Information Management

College of Management

National Chiao Tung University

In Partial Fulfillment of the Requirements

For the Degree of

Master of Science

in

Information Management

June 2013

Hsinchu, Taiwan, the Republic of China

中華民國 102 年 6 月

運用文字探勘技術在社群行為上之人格預測

研究生：張曉珍

指導教授：李永銘 博士

國立交通大學管理學院（資訊管理學程）碩士班

摘要

現今網路無遠弗界，人與人溝通或社交行為，已由早前的書面作業或面對面交談，漸漸成為線上作業。因時代的變遷，人們已較常在社群網路上發表文章及抒發自己的情緒，這些個人行為會透過文字的表達呈現於文章中。本研究資料來源透過 Facebook 社群行為中文語料的部份進行人格分析，採用最被廣為接受 Costa & McCrae(1985)的五大人格特質構面（Five Factor Model, the Big Five）。此五大人格特質，分為神經質型、外向型、開放型、隨和型、嚴謹型五大類。

研究方法採用二種方法進行研究及比較，方法一為關鍵詞彙預測法，透過中研究的 BOW - WordNet 擴充詞彙；方法二為機器學習預測法，採用自行開發的程式利用變型的貝氏理論加以研究及實作。研究結果顯示，針對顯著人格加以分析比較，方法一有 61%的準確率，方法二有 80%的準確率。方法二的實驗結果所預測的準確率高於方法一的預測結果，二項方法當詞彙數夠多時，更可增進研究分析的準確性。另研究透過方法二學習訓練後的詞彙，用來自動擴充方法一的詞彙。在使用測試集的資料加以驗證後，結果顯示有效的增進方法一的預測結果，由 66.67%增至 73.33%。故本研究證實透過在 Facebook 的中文貼文可有效的分析個人在社群網路上的人格特質，未來可供後續研究者參考，以及企業應徵人員的參考依據等效益。

Using text mining to predict personality based on social behavior

Student : Xiao-Zhen Chang

Advisors : Dr. Yung-Ming Li

Master Program of Institute of Information Management
College of Management
National Chiao Tung University

Abstract

Nowadays Internet is used for communication widely. People prefer communicating via Internet Services over talking face-to-face or writing letters. They are more often writing blogs or posting messages on social networks and the personality will be presented by habitual vocabularies they used. This research is analyzing Chinese vocabularies to predict personality from posted contents by Facebook users. The personality classification is based on Five Factor Model (Costa & McCrae, 1985). The five categories are Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness.

This research compares two methods. Method one is key vocabulary prediction by using SINICA BOW-WordNet. Method two is machine learning prediction by using compact Bayes theorem. The results show that the accuracy of method two (80%) is better than method one (61%). The accuracy of method two will be better when the sample is enough. The result could be used to extend vocabularies of method one and improvements accuracy from 66.67% to 73.33%.

This research demonstrates a different way to analyze personality by analyzing posted contents on Facebook from traditional questionnaire and the contribution of this research can provide helpful reference to HR of enterprise when recruiting employees.

誌 謝

在校二年時間要忙於修課、忙於工作、忙於家庭。以前沒發生的事，似乎所有的事全在這段時間發生了，一週課程排得很滿，對於坐火車通勤的我，真得很難熬。尤其完課回到家已快 11 點，隔天又得趕坐 7 點的火車，僅有的空閒時間大概只剩在火車上及假日了。假日跟家人相處出遊的時間不多，大多是用來寫作業或參加班上活動的時間，對於家人的包容、體諒真得非常地感激。

求學過程雖然很辛苦，但這二年過得真得很充實，雖然很累，但學到很多，不管是在校所學的專業知識或是跟同學相處的互動。感受到交大資管所的師生真得很優秀，很專業，也一直趕在時代尖端，求新求變。指導教授李永銘老師的名言，眼高手黑、站在巨人的肩膀上等，能夠體會這些話的含意真得收益良多。

很感謝指導教授在論文上及課堂上的教導，也很感謝資管所各位教授的教学及指導，還有同學間這段時間的相處。能夠在工作一陣子又回到學校上課，這種經驗真得很難得，會是我人生中一個很重要的回憶，謝謝你們。

最後非常感謝論文口試委員陳安斌教授、劉敦仁教授、金必煌教授及指導教授李永銘教授的指導，使論文最後能呈現的更完整，謝謝你們。

目 錄

中文提要	I
英文提要	II
誌謝	III
目錄	IV
表目錄	V
圖目錄	VI
一、緒論	1
1.1 研究背景	1
1.2 研究動機與目的	1
1.3 研究範圍	2
1.4 論文架構	2
二、相關研究	4
2.1 社群網站	4
2.2 人格心理學	6
2.2.1 人格	6
2.2.2 人格特質	7
2.2.3 特質論	7
2.2.4 五大人格特質	9
2.3 習慣領域	11
2.3.1 習慣與習慣領域	11
2.3.2 習慣領域與五大人格關係	13
2.4 資料探勘	14
2.4.1 資料探勘的起源	17
2.4.2 資料探勘與知識發現	18
2.4.3 資料探勘功能	18
2.4.4 資料探勘分類模式	19
2.4.5 貝氏分類法	20
2.4.6 單純貝氏分類法	20
三、設計與方法	22
3.1 系統架構(方法一：關鍵字詞預測法)	22
3.1.1 Big Five 特徵詞彙	23
3.1.2 BOW - WordNet	23
3.1.3 特徵詞彙擴充	23
3.1.4 特徵詞彙資料庫	25

3.1.5 FB 資料截取	26
3.1.6 CKIP 斷詞斷句	26
3.1.7 TF- IDF	28
3.1.8 五大人格預測	28
3.2 系統架構(方法二：機器學習預測法)	29
3.2.1 系統架構(方法二) 流程說明	30
3.2.2 方法二(機器學習預測法) 演算	30
3.3 問卷來源	32
3.4 軟體及平台使用	32
四、系統實作及資料驗證	34
4.1 Facebook 資料截取及處理	34
4.2 資料驗證 - 心理測驗結果	37
4.3 方法一(關鍵字詞預測法)	40
4.4 方法二(機器學習預測法)	42
五、實驗結果與分析	48
5.1 方法一(關鍵字詞預測法)	48
5.2 方法二(機器學習預測法)	49
5.3 方法二支援方法一	53
六、結論及建議	57
6.1 結論	57
6.2 建議	59
6.3 效益	60
參考資料	61
<u>自</u> <u>傳</u>	63
附錄一	65
附錄二	67
附錄三	69

表 目 錄

表 1：人格定義.....	7
表 2：習慣領域四概念.....	12
表 3：例子：習慣領域特性與五大人格關係.....	14
表 4：截取 FB 資料之程式開發步驟.....	26
表 5：FACEBOOK 資料截取及處理步驟.....	34
表 6：FB STORY、MESSAGE、PLACE 功能描述	35
表 7：方法二實作前置步驟.....	43
表 8：方法二實作步驟.....	43
表 9：方法二詞彙分類範例.....	45
表 10：方法一擴充結果比較表.....	55
表 11：方法一擴充前後準確率比對.....	56
表 12：中研院平衡語料庫詞類標記集.....	66
表 13：心理測驗衡量評分表.....	69



圖 目 錄

圖 1：系統架構.....	3
圖 2：社群網站用戶數據排名.....	5
圖 3：全球社群網路使用分佈圖.....	5
圖 4：五大人格特質.....	10
圖 5：習慣與習慣領域關聯圖.....	13
圖 6：習慣、習慣領域、五大人格關聯圖.....	14
圖 7：KDUNGGETS 票選於 2012 年使用資料探勘技術的行業.....	16
圖 8：KDUNGGETS 票選於 2012 年使用資料探勘熱門主題.....	16
圖 9：資料探勘與其它領域關係圖.....	17
圖 10：資料探勘四大核心技術.....	17
圖 11：知識發現過程.....	18
圖 12：分類模式圖.....	20
圖 13：系統架構(方法一).....	22
圖 14：WORDNET - 中文查詢介面.....	24
圖 15：WORDNET- 英文查詢介面.....	25
圖 16：中研院中文斷詞系統.....	27
圖 17：語料庫詞類標記取得.....	27
圖 18：系統架構(方法二).....	29
圖 19：FACEBOOK 中文貼文資料截取.....	35
圖 20：FACEBOOK 中文貼文資料檔.....	35
圖 21：截取 MESSAGE、STORY、PLACE 對應內文.....	36
圖 22：CKIPCLIENT 字詞切割.....	36
圖 23：COUNTWORDFREQ 詞頻計算.....	37
圖 24：心理測驗問卷格式.....	38
圖 25：心理測驗問卷回覆.....	39
圖 26：心理測驗結果分類.....	39
圖 27：方法一個人用詞資料檔.....	40
圖 28：方法一之 IMPORTER 程式.....	41
圖 29：方法一特徵詞資料庫.....	41
圖 30：方法一人格分類實例.....	41
圖 31：方法二實作介面.....	42
圖 32：方法二匯入文字檔內容.....	44
圖 33：方法二實作.....	45
圖 34：方法二詞彙分類.....	46
圖 35：方法二詞彙 NORMALIZE 及權重值.....	47
圖 36：方法二詞彙 GRADE 計算.....	47
圖 37：方法二類別預測結果.....	47
圖 38：方法一研究結果(1).....	48

圖 39：方法一研究結果(2).....	49
圖 40：方法二測試集個人預測結果(1).....	50
圖 41：方法二測試集個人預測結果(2).....	50
圖 42：方法二測試集個人預測結果(3).....	51
圖 43：方法二測試集個人預測結果(4).....	51
圖 44：方法二五大人格分類統計表.....	52
圖 45：方法二（機器學習預測法）預測結果.....	53
圖 46：方法二擴充詞彙截取.....	54
圖 47：方法一擴充詞彙資料產生.....	54
圖 48：方法一擴充結果比較圖.....	55
圖 49：二方法顯著人格預測比較.....	57
圖 50：擴充詞彙方法關聯圖.....	58
圖 51：擴充詞彙人格預測比較趨勢圖.....	58



一、緒論

1.1 研究背景

目前社群網站的興起，以 Facebook 而言，截至 2013 年一月全球註冊人數已達到九億八千萬，而台灣註冊人數約為一千三百三十萬，而大多數上網有臉書帳號的人比比皆是。使用範圍無論是社群照片分享、發表文章、尋找資訊、推薦、尋找朋友等已廣泛被使用，且融入人們的生活中，可說是不可或缺的朋友或為具影響力的媒體了。

1.2 研究動機與目的

隨著時代變遷，伴隨著資訊科技的蓬勃發展，網路世界已成為人與人溝通的媒介，所以社群網路因此而展開。由於透過社群網路有其方便性及可讀性，所以人們較常在社群網路上發表文章及抒發自己的情緒。現今社群網路可說已達到成熟的階段，社群行為分析及預測亦已漸漸被重視及研究。

Facebook 檔案能顯露出應徵者未來工作的表現，在近期的一項研究中，學者邀請人資專員人工瀏覽應徵者的 Facebook 檔案，並問卷依照內容所顯現的人格特質來評比未來工作表現。結果顯示與當初人資專員依照 Facebook 檔案所作出的預測極為接近，準確度甚至超過了性向測驗結果。

另在 2011 年的研究顯示有 90% 的聘雇者曾經檢閱過工作應徵者的 Facebook 檔案。學者表示運用 Facebook 在聘雇員工時有其價值和實用性，但需進行更多的實證[3]。故本研究開發經 Facebook 授權允許，程式自動截取 Facebook 中文內文，依據五大人格特質構面來分析此用戶的五大人格。

本研究針對 Facebook 社群行為中文語料的部份進行人格分析，採用最被廣為接受 Costa & McCrae(1985)的五大人格特質構面 (Five Factor Model, the Big Five)。此五大人格特質，分為神經質型、外向型、開放型、隨和型、嚴謹型五大類。網路世界無遠弗界，從使用 Facebook 的文字行為可分析每個人在社群網路上的人格特質。本研究運用文字探勘技術在社群行為上預測人格特質，可供後續研究參考及企業應徵人員的參考依據。

1.3 研究範圍

研究對象針對個人及朋友臉書的中文貼文，開發程式截取臉書資料，再以資料探勘分析方法，預測人格特質。

1.4 論文架構

本論文研究設計採用關鍵字詞預測法及學習預測法二種方式，第一種方式之關鍵字詞預測法，主要依據五大人格特質之五大類別中的各個特質，挑選可具代表此五類的相關特徵詞，接著利用中央研究院的中英雙語知識本體詞網依據此特徵詞再加以擴充更多的特徵詞，以增加研究分析結果的準確度。再來從 Facebook 裡截取貼文後，再利用中研院 CKIP 中文斷詞系統，進行斷詞動作及詞頻計算，以取得五大人格特質的歸類。第二種方式之機器學習預測法，透過自動開發機器學習的方式，以百分之八十為訓練資料，由 FaceBook 之 Big Data 中抓取動詞及形容詞，依據已知類別對應相關之特徵字詞加以訓練，用以計算權重，而百分之二十之測試資料，則依據訓練集訓練後的特徵詞歸類，依據個人常用特徵詞預測個人於五大人格之類別。

本論文之系統架構如下圖 1，分為緒論、相關研究、設計與方法、系統實作及資料驗證、實驗結果及分析、結論及建議共六大部份。

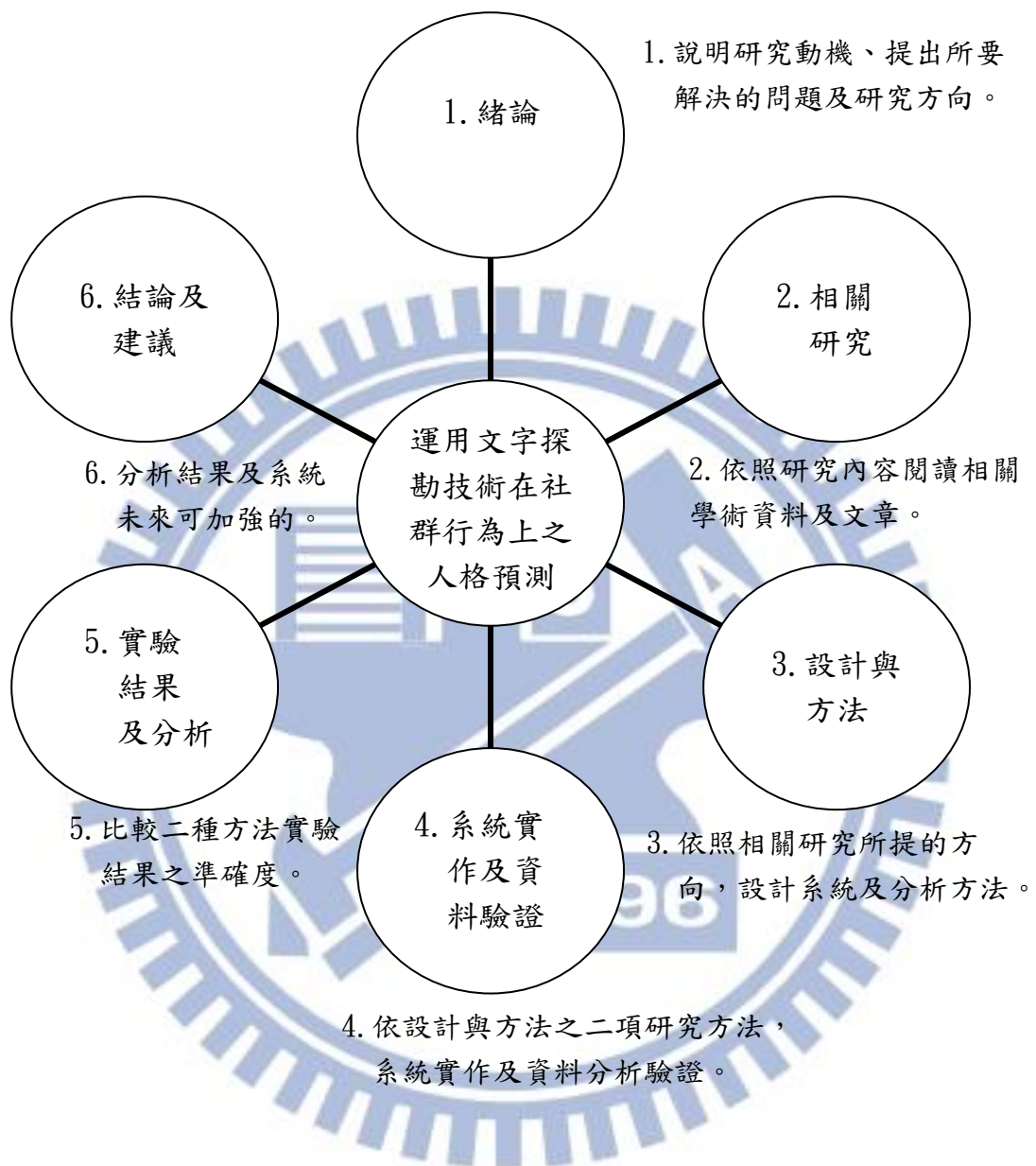


圖 1：系統架構

資料來源：本研究自行整理

二、相關研究

本研究結合社群網站，並依據人格心理學及習慣領域的理論，使用資料探勘的分析方法，預測 Facebook 用戶於社群行為之人格特質。

2.1 社群網站

社群網站是以建立線上社群的方式，透過虛擬空間，提供平台讓志同道合的人群聚在一起，交流互動、維繫情感、分享、討論資訊等功能。

較廣為人知的社群網站如 Facebook、Google+、Plurk、微博、Twitter 等，而以 Facebook 為較多用戶使用。

根據全球著名的人力資源網站 Glassdoor 調查指出，全球 50 大最佳職場的企業中 Facebook 排名第 1 名，Facebook 公司的標語為” Challenged every day to do your best work”。該公司領導層真正相信 Facebook 的使命，是使世界變得更加開放和連接。Facebook 廣受歡迎及使用的原因正是如此，也是成功的原因之一。

使用 Facebook 的動機包含有好奇心、跟上時代潮流、免費、方便操作、人性化等。而使用者使用 Facebook 所得到的好處則有歸屬感、商機、炫耀、虛榮心、增加曝光、聊天、展現自我、聯繫感情、拉近彼此關係、分享生活即時動態及資訊、拓展交友圈、調劑時間、休閒愉樂、尋找朋友、宣揚理念、推動社會運動等等，這些優勢已讓 Facebook 成為大多數人民生活的一部份[16][17]。

Globalwebindex 數據估計，Facebook 全球 31 個市場在 16-65 歲中有九億三百萬用戶。而於 2012 年 12 月止臉書使用戶亦佔全球社群網站之冠，相關統計數據，如下圖 2。

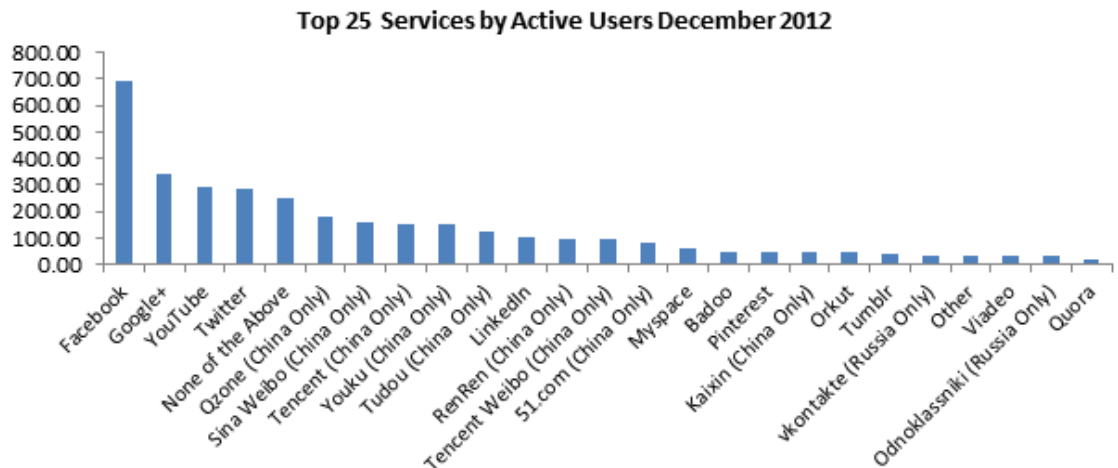


圖 2：社群網站用戶數據排名

資料來源：Globalwebindex

根據 VINCOSBLOG 分析指出截至 2012/12，臉書以十億的有效使用戶數，在 137 個國家中佔了 127 個領先地位。其中以亞洲為最重要的成長關鍵，使用者高達 2 億 7 千 8 百萬，超過了歐洲的 2 億 5 千 1 百萬使用者，成為臉書最大的版圖，如下圖 3。

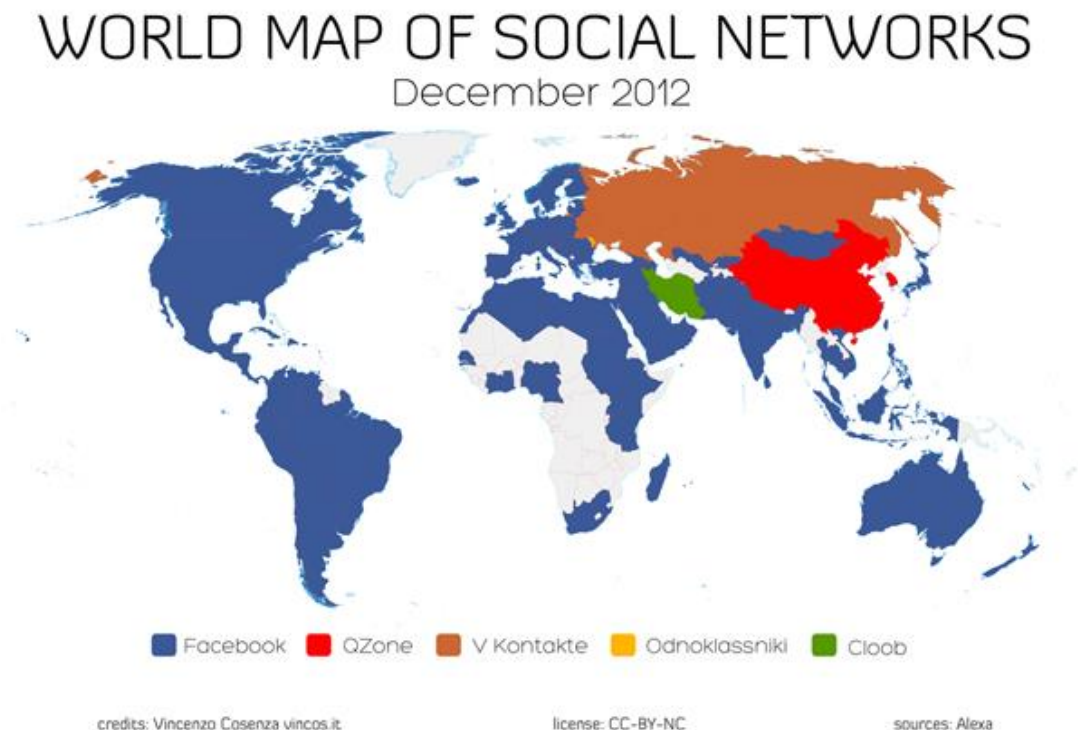


圖 3：全球社群網路使用分佈圖

資料來源：VINCOSBLOG

2.2 人格心理學

若今天你去面試一家公司時，或許有人問你，請描述自己是怎樣一個性格的人呢？這問題你會如何回答？遇到這種問題時，其實很多人對自己都不是很了解，因為一個人可能會有多重性格存在，對於回答的答案準確性有多高？其實有待確認，但可確定的是，每個人的性格表現可由平常的行為中所展露。Christopher Nave(1960)研究指出人格特質從小時候至年老基本上是相同的。顯示小學一年級的個性是關鍵期，從那時開始往後反映至成人時之行為，可說性格伴隨人的一生，不因時空及成長背景而有所改變。當然各方心理學家對於人格有不同的定義，但最終存在兩個基本概念，一為獨特性，另一為行為模式，也就是所謂的個人獨特的行為模式，這人與人之間的差異性正是人格心理學研究之處。

2.2.1 人格

人格為性格傾向呈現穩定的表徵，意指「一個人內容相當持久的特質、性格或特徵，使得這個人的行為顯示一致性」。人格心理學家已經找出許多研究人格的方法，某些人格理論家致力於長時間的、深入的研究個體的人格。另一些則努力發展心理測驗，來測量許多人的性格，還有另一些理論家研究某些人們所共有的人格特質[1]。

人格廣義指存在於個體內的特質、動機、情緒、自我價值、因應策略等內涵(Ryff, Kwan, & Singer, 2001)。因此，可用某些特質及動機的強弱來描述一個人的人格，因人格可以是存在於個體身心系統中的一種動力組織，及是決定個體行為與思想特質所在(Allport, 1961) [2]。人格可分為類型及特質，類型指類似五大人格分類的開放型，外向型，嚴謹型，內向型，隨和型。特質指描述五大人格類型，例如快樂的，開心的，憂鬱的等等。

而性格的本質具有一致性及特殊性二種，也就是說，一個人在不同時間及情境下仍然表現出與以往相同的行為，以及在同一情境或不同情境下會有不同以往的表現，而因先天或後天造成，會影響一個人的性格特質，而所謂人格特質，是用以區別個人心理特色的綜合體。心理學家主張人格特質的測量是研究人格的首要工作，且多採用標準化的測驗方式來測量人格特質[2]。

至於人格心理學大致可分為五大理論：心理動力論、行為論、特質論、人本論、以及互動論等，特質可以發揮統括、預測、解釋個人的行為舉止三種主要的功能。

自 1980 年代以來，許多學者研究證實，選擇適當特質，可利用特質來預測行為（Darley, 1994）。以下表 1 為多位學者對人格的定義，從學者對於人格的定義中可得之。人格特質的行為表現為一個人成長過程所形成的個人獨特性，在在與個人的生活習性及基因有關，而本論文以特質論為研究重心，排除以往的人工標準化測驗，採用專家知識庫及機器訓練的方式，結合社群，截取日常行為的活動記載以預測人格特質。

提出者	年度	定義
Carl Jung	1921	原型：一種遺傳傾向、人們普遍共有、特定行動等
Allport	1937	即一個人之心理及生理系統所形成的內在動態組織，決定個人對於環境獨特的適應性
Karen Horney	1941	人格結構是真我、實我和理想我的組合
Cattell	1965	可以預測一個人在某一個情境中將表現的行為
Eysenck	1970	指個人性格、氣質、智慧及體質等，一個穩定而有持續性的組織
Scott & Mitchell	1972	是一種人類心理成長及發展的過程
Pervin & John	1997	為形成一個人經常性的行為、思想等特性

表 1：人格定義

資料來源：本研究整理[14][15]

2.2.2 人格特質

人格特質是指個人在成年後的一切行事作風、人際關係、語言表達上，皆有每個人的表現模式，此稱為人格特質。比方會說某人平常待人很熱心、不情緒化、很開朗等，種種皆為形容一個人的人格特質[4]。

2.2.3 特質論

針對人格特質之相關研究學者眾多，且各方皆有其一定的研究領域，本研究僅

以目前較受注意的特質論來研究個人之人格特質。而特質論對人格的觀點如下：

特質：指人與人之間之不同，構成獨特自我特性[1]。

來源：1. 先天遺傳性格。

2. 後天環境的影響。

特質區分二種派別：

1. 普遍共有：每個人皆擁有相同一組特質，每個人會有差異於所表現的所有特質程度的多少，所以可解釋特質是普遍共有的[1]。
2. 個人不一：每個人有不同的特質組合，所以是有差異的，因此有可能每一個人存在的特質皆不相同[1]。

以下於特質論中針對 Gordon Allport、Carl Jung 及 Karen Horney 三位心理學家的理論加以介紹如下[1]：

Gordon Allport (1961)將個人特質分為三大類：

1. 首要特質：指最具代表性的個人獨特性質。此特質是一種單一特質，且強烈支配個體的人格，反應於個人行為上，例如：泰瑞莎修女首要特質為富有憐憫心。
2. 中心特質：針對個人的人格及行為是影響力重大，且此中心特質是指構成人格特質的核心部份。如同形容個人負責任、樂觀、誠實等皆是歸屬於中心特質。
3. 次要特質：此特質指在某種情況下才會表現出來的特徵，對人格及行為的影響不甚明顯，亦較不重要的人格特質。

Carl Jung 的分析心理學將心智可分為二層面，顯示人是有潛意識的[1]：

1. 意識層面

2. 潛意識層面

第一層：個人潛意識

為個人累積的經驗形成，包含被壓抑及被忽略的記憶。

第二層：集體潛意識

為人類所共享的記憶及行為的相關記憶。

Carl Jung 從傳說、信仰及文化習俗中，找到一種傾向，這傾向是人們普遍都有的知覺或特有方式或行動等相關的遺傳因子，此為 Carl Jung 所稱的原型，而這個原型是存在於人們所謂的集體潛意識之中[1]。

Carl Jung 認為對人類生活而言，特別具重要的原型包括以下幾點[1]：

1. 面具：也是指人們將人格表現於外，願意公開且與人分享。
2. 陰影：隱藏自己的一些恐懼、怨恨、邪惡等的念頭，不對外公開，僅存於個人內心裡，或許連自己都不清楚自己有此念頭。
3. 陰性基質：指男性人格中含括部分女性特質，在男性中顯示出溫柔的一面等正是陰性基質的呈現。
4. 陽性基質：指女性人格中含括部分男性特質，在女性中顯示出理性的一面等正是陽性基質的呈現。

綜上可得，每個人從生來就具有每個人一些獨特特質，由遺傳得到這些特質或由成長過程中發展出這些特質，這些特質是潛藏於個人潛意識中，會不由自主的第一個反應於個人的行為之中。而佛洛伊德：潛意識決定重要決策；狄克斯特霍斯、凡歐登：潛意識思考理論，讓潛意識來思考決定問題。所以有研究顯示，個人特質由基因及童年的經驗決定，通常不太會變，且會影響各方面的行為，包含了休閒活動、創意、健康、言語、行動及幽默感等行為[9]。所以本研究由個人於社群行為中的文字書寫內容，運用文字探勘來分析及預測此人之性格分類。

另外人們會好奇想問，心理學家皆有一套心理學理論，但他們都是如何衡鑑人格的呢？心理學家一般以人格投影測驗來衡鑑人格；也就是以心理動力論（心理動力論者認為行為由內部力量所驅動）為基礎的方式，利用人格測驗，試著探索人們潛意識的部份，以完成對個案人格之評估。所以傳統皆以人格分析表為測量工具，現今可由科學的方式，利用開發工具截取相關個人訊息，亦可達到衡鑑人格的目的。

2.2.4 五大人格特質

從 20 世紀 80 年代以來，人格研究者們在人格描述模式上達成了比較一致的共識，提出了人格五因素模式，被稱為“Big Five”。近年來相當多研究顯示在許多學者所歸納人格特質以五種廣泛的因素解釋五大人格，五大人格模式為目前為止最

被廣泛接受的 (Digman, 1990; Goldberg, 1992)。現今五大人格模式已成為研究人格的典範 (Blue, 2000; Costa & McCrae 1992; Howard & Howard, 1995)。

故歷經多年的研究證實五大人格的確存在，而這些特質可了解自己及他人，因而科學研究人員研究務實且有效的方法，他們相信人心的祕密結構是潛藏在語言中的，他們猜測人類用於形容自己與他人的詞彙，應是反應於人的個性的基本。而於一九三〇年代，一群專業的研究者挑選可用於形容人的個性的字詞，共整理了一萬八千多字，接著找出重要字眼，總共有四千多個，至一九四〇年代，另一批研究者利用電腦分析降低維度，將原先四千多個字詞，縮減為二百字左右，而往後的四十年來，數千人因此利用此形容詞來分析及評估人類的五大人格特質[9]。

五大人格特質分為外向型、隨和型、神經質型、嚴謹型及開放型五大類別，如圖 4。



圖 4：五大人格特質

資料來源：本研究整理

五大人格特質描述如下：McCrae 和 Costa (1987, 1997, 1999) [4][9]

外向型

指受外界與他人的刺激的需要。直率、喜歡社交、生氣勃勃、友善、有自信且合群等是這類型的表徵。

神經質型

指情緒穩定與抗壓的程度。容易焦慮、具有敵意、自我意識較高、容易

有不安全感、脆弱易受傷。

開放型

指一個人追求與欣賞新奇、有趣、罕見經驗的程度。開放性跟好奇心、較具彈性、有豐富的想像力、有藝術家的敏感以及不墨守成規的態度有關。

隨和型

此類型的人常表現於關懷他人的程度，比如有同情心、值得信任、較合群、謙虛且正直。

嚴謹型

指做事有條理、堅持不懈、自律的程度。謹慎的人比較勤奮、有規律、計畫周詳、守時且較可靠。

五大人格有五個面向，每個面向皆有高低面向，而本研究以五大人格模式之顯著人格，也就是取高分者為主要人格的研究方向，透過人類行為反應於文字，分析語句以找出人類個性的主要面向。

2.3 習慣領域

習慣領域指動態的念頭和思路（凡指做法、反應、觀念等），經一段時間會漸穩定，且會停於固定範圍內。而這些範圍所包含動態和組織，也就是所稱的習慣領域。而習慣領域的形成，除非有重大事情或新的訊息產生或刻意擴展習慣領域，否則此習慣領域皆是在穩定的固定範圍之中的[6]。

2.3.1 習慣與習慣領域

一個人平常的習慣及行為的養成是經年累月所產生於個人的大腦中的，人類的大腦有一千億個腦細胞，而大腦又是主宰人類行為的中樞，可說人類的行為是受大腦的指令所控制。大腦運作時會有明暗變化所組成的網狀，也就是所稱的電網，每個電網的變化，反應於一個行為，而這些動作是受長期習慣所養成，且不知不覺的控制我們的行為。

習慣與習慣領域不同。習慣指的是那些強有力的電網產生的行為，也就是潛在領域，指的是潛藏在你大腦裡經年累月所累積的思路及念頭。習慣是習慣領域的一小部份，習慣領域不僅包含習慣的潛在領域，亦包含實際領域、可發概率及可達領

域共四種，如圖 5 所示。

以下介紹將習慣領域四概念，及將習慣領域舉例應用於個人於 FaceBook 貼文時的對應關係，如表 2 所示。

習慣領域 四概念	定義	例子(當你在 FB 貼文時)
潛在領域	總和潛藏在腦海裡的所有電網	你腦海裡所有知道的詞彙
實際領域	目前時點正占領我們注意力的電網	此時此刻你會引用何詞彙 為你貼文的內容
可發概率	每個電網實際占有注意力的機率	此詞彙被你引用的機率
可達領域	由實際領域所延伸的電網	實際引用的詞彙後，會引 發一些其它的情緒及想法 等的詞彙產生

表 2：習慣領域四概念

資料來源：參考習慣領域[6]，本研究自行整理

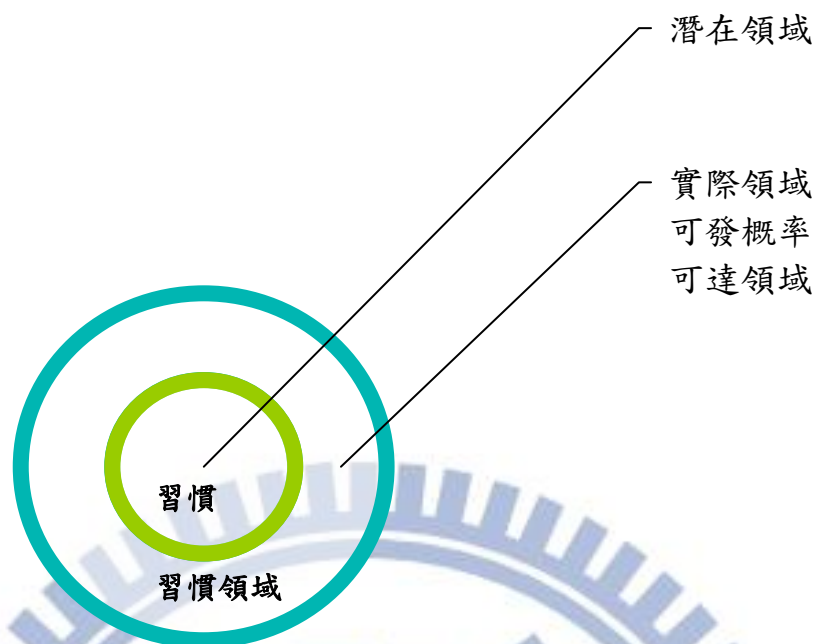


圖 5：習慣與習慣領域關聯圖
資料來源：本研究自行整理

2.3.2 習慣領域與五大人格關係

習慣領域中先天的電網是與生俱來的，當我們漸漸長大時，到十多歲時，電網會越漸加強，至近二十歲時，此時電網會變得更強而有力，這些先天的電網是不容易被改變的。而後天的電網是透過學習及經驗累積而來，亦受環境的影響很大，故不同文化不同國家的人種會有所不同。但可知的是，這些不管是先天或後天形成的電網，都可由個人的行為反應表現出來，只是個人受哪個時期的電網影響較大，即表現在個人特質上會越明顯。

習慣影響一個人的行為，而由一個人的平常行為特徵可預測這個人的人格特質，當想法等電網被使用的越多時，其實越容易被取出，也就越容易表現於行為之中，所以預測人格的準確度更可提升，可更了解自己及他人。這些行為特徵，往往受個人的平常習慣所影響。好的修養習慣，會讓一個人自尊自愛、沈重穩重、有耐心。待人處世習慣好的人，會樂於助人、有親和力、坦率、有關愛的心亦富同情心。而有好的生活習慣的人，會有自信、樂觀、勇於前進的精神。

以上可了解每個人的行為習慣不同、習慣領域不同，潛在領域及實際領域的表現亦不同，因此呈現於個人的人格會有所差異。如表 3 所示，習慣領域最終可反應

於人格的表現，所以我們在研究五大人格特質的同時，其實需先了解人類的習慣、習慣領域，這三者是相輔相成的。

如圖 6 可呈現習慣→習慣領域→五大人格關係。也就是個人從小養成的習慣，這些潛在因素是造成人格發展的起因，再加上後天的環境的影響及重大事物的造成，皆看此重大影響層度的大小而定，以決定個人習慣領域的範圍大小。所以習慣再加上習慣領域會造就一個人的不同人格特質，本研究以五大人格特質加以分析。

習慣領域特性	五大人格特質
懂得感恩	隨和型
求新求變	開放型
熱心	外向型
自律性強	嚴謹型
易怒	神經質型

表 3：例子：習慣領域特性與五大人格關係

資料來源：本研究自行整理

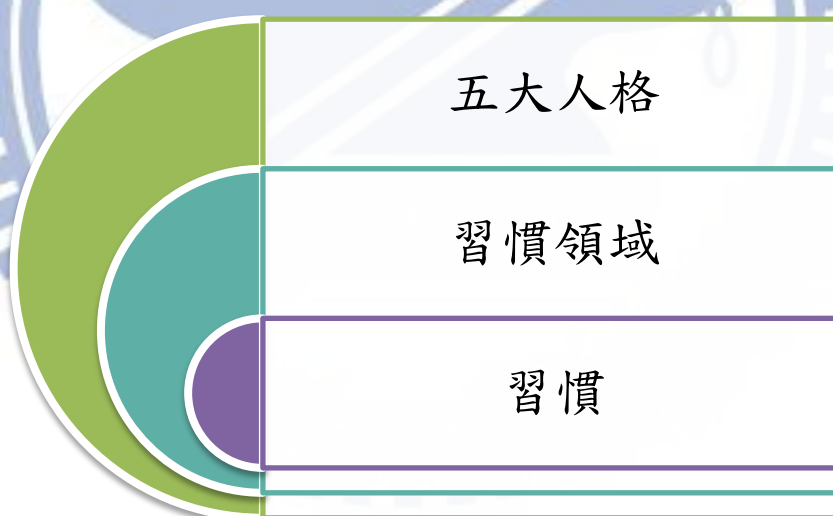


圖 6：習慣、習慣領域、五大人格關聯圖

資料來源：本研究自行整理

2.4 資料探勘

資料探勘，亦稱資料採礦，為近年來資料庫領域熱門的議題，是利用統計以及機械學習的演算法，啟發性地從大量資料中找尋隱藏具有商業價值的知識與規律，

以作為自動化商業策略之應用。另資料探勘是啟發性透過演算法主動搜尋有意義的規則，而讓資料說話，以及與傳統統計最大的差異在於它具有商業化以及行動的意涵[5]。

資料探勘其中包含文字探勘及網頁探勘等，文字探勘為處理非結構化資料，通常字句長短不一定，且訊息的記載文字是很自由的，可說每一筆資料沒有共通的結構性。關鍵詞是表現文件意義的最小單位，因此文字探勘技術所使用之自動處理，如自動分類等應用彼彼皆是。自由文句在做文件分類時，常將文件切成一個一個語意單位，稱為特徵詞彙，從這些特徵詞彙與類別來找出彼此對應關係。網頁探勘為從 Web 資料發掘本質關係(即有趣與有用資訊)的流程，以文字、連結或使用資訊形式表示，而 Web 為世上最大的資料儲存庫，其資料為 HTML、XML、文字格式[11]。

資料探勘領域含括資料庫技術及機器學習技術，許多研究人員及產業人員，認為資料探勘領域是增加企業潛能的重要指標，因透過資料探勘技術是可挖掘出具有意義的資訊，可做為決策之用，更可增加企業的競爭力，是企業的愛好者。圖 7 為 Kdunggets 票選於 2012 年使用資料探勘技術的行業，HR 為佔了第二名，本研究針對 Facebook 行為預測性格分析，可應用於 HR 招募時人格分析。

本研究以預測類別變數之社群文字分析為研究重心。如圖 8 所示於 Kdunggets 票選於 2012 年使用資料探勘熱門主題，社群分析佔前三排名，可顯示社群尚是目前熱門的研究對象之一。

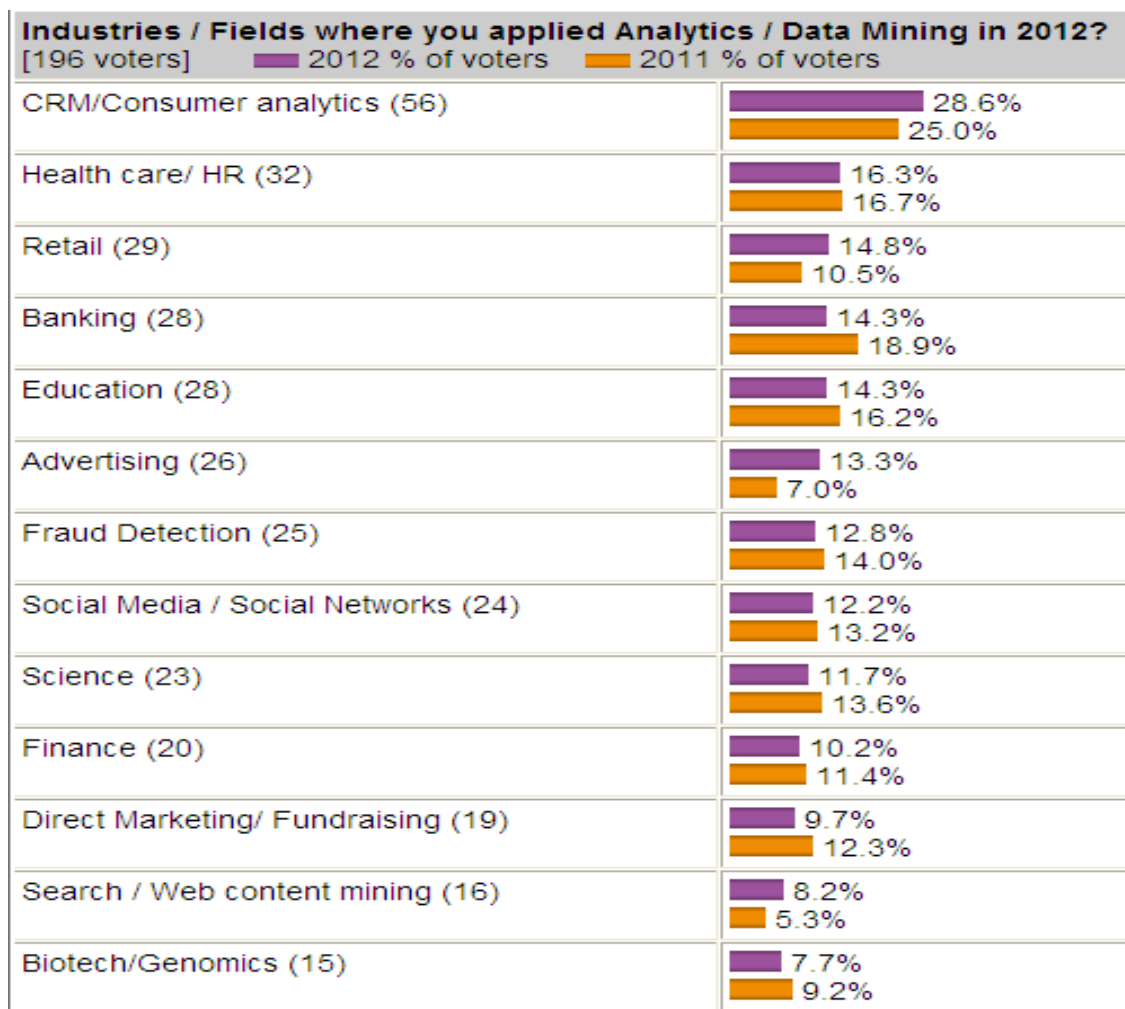


圖 7：Kdunggets 票選於 2012 年使用資料探勘技術的行業
資料來源：Kdunggets

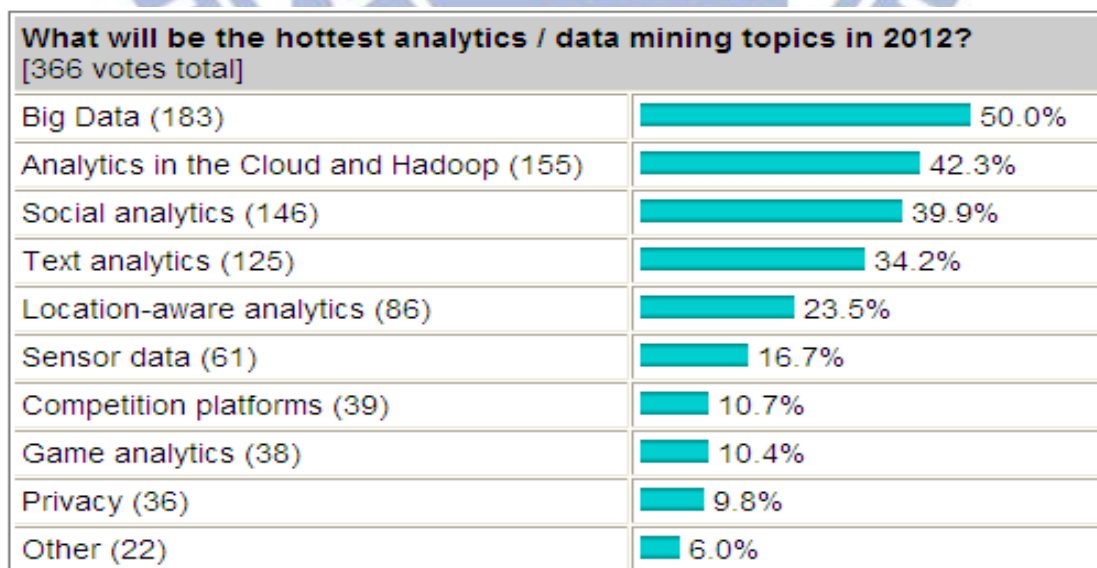


圖 8：Kdunggets 票選於 2012 年使用資料探勘熱門主題
資料來源：Kdunggets

2.4.1 資料探勘的起源

傳統資料分析技術常會遭遇無法處理的資料問題如大量資料、高維度問題、異常型態資料等問題，而資料探勘的起源就是研究者為了解決傳統資料分析所帶來的問題，故使用方法論及演算法發展出這資料探勘的分析工具，這工具可有效處理大量資料及不同資料型態的資料。

資料探勘其實包含了統計學的假設檢定、人工智慧、學習理論等，而與其它領域之間可說扮演很重要的關鍵角色，如圖 9 所示。研究者另針對資料探勘的工作分為四種核心技術，分別為分群分析、預測模式、關聯規則分析及異常偵測四種，如圖 10 所示，在此四種核心技術中，本研究以預測模式的分類模式為研究方法。



圖 9：資料探勘與其它領域關係圖

資料來源：本研究整理

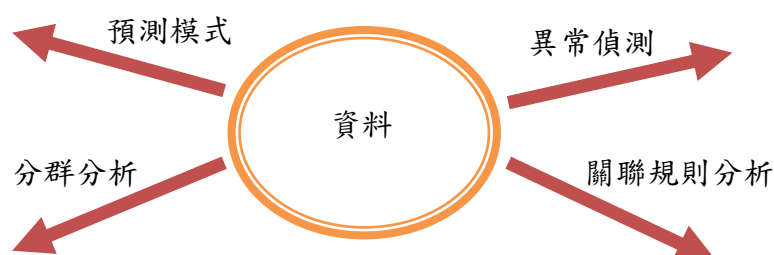


圖 10：資料探勘四大核心技術

資料來源：本研究整理

2.4.2 資料探勘與知識發現

資料探勘為知識發掘的過程之一，知識發展整個部份包含將資料轉換成有用的資訊的一個過程，也就是這部份包含了資料前置處理至資料採礦後的後處理的一個轉換過程，如圖 11 所示[19]。

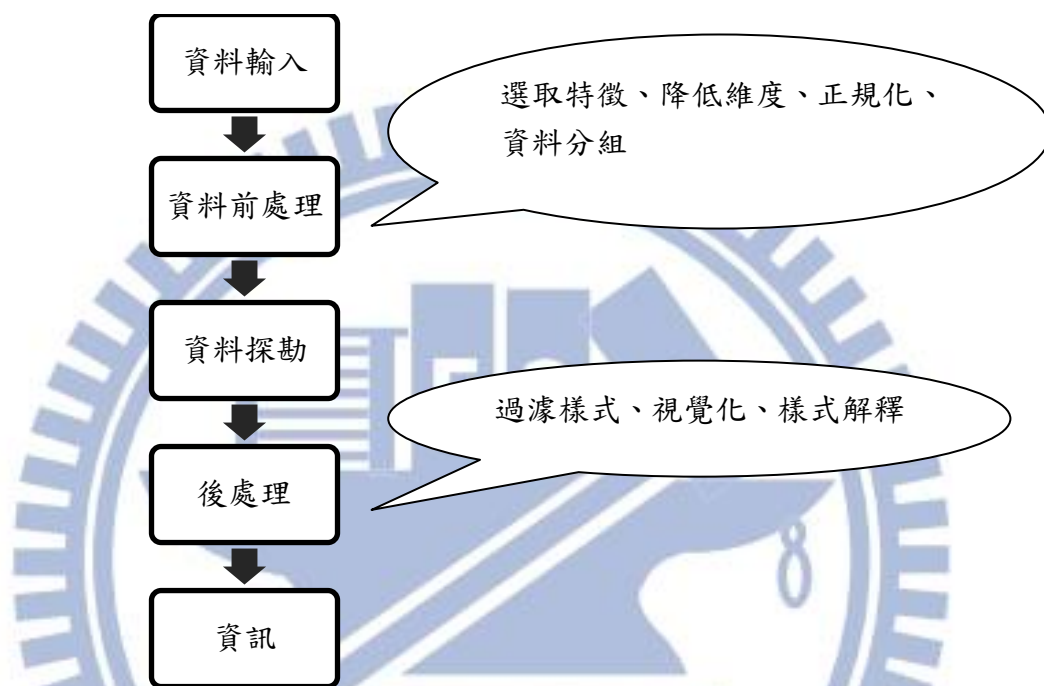


圖 11：知識發現過程

資料來源：資料探勘[7]P.1-3

2.4.3 資料探勘功能

資料探勘的規則型態可區分為六種模式[10]

分類(Classification)

預測類別變數的一種過程稱為分類，分類是資料探勘中最普遍的一種模式，也是人類為了溝通，會將事物加以分類，是人類在產生知識規則最初始的型態。在大多數所謂的分類技術中，是依據已知的結果來分類，但資料探勘中的分類則為預測的技術，也就是要在事物發生前，需事前預測分類結果，這是資料探勘分類與一般分類最大的不同點。分類模型可再細分為決策樹、叢集演算法、以及貝氏機率分類等演算方法。

推估(Estimation)

推估為預測連續數值之相關屬性資料。使用的演算方法，包含迴歸分析及類神經網路分析等方法。

群集化(Cluster)

依據相似度，將相似資訊歸於同群，可將原複雜且大量的資料，將以分群後可大幅簡化，此過程即稱為群集化，而物以類聚是形容群集化再好不過了。群集化是對未知的事實找出內部相似性加以分群，而分類則是依據已知的類別，來預測未知的事實。

同質分組(Affinity Group)

同質分組亦可稱為關聯規則，從眾多物件中，找出哪些物件是被關聯且應發生在一起的，故同質分組可找出各組之間的差異程度。最著名為啤酒與尿布例子。

序列(Sequential)

序列亦可稱為時間序列，為在同質分組中找出事物相互關聯之先後發生順序，可協助找出事物的生命週期，利用不同時點的因應措施，供決策分析使用。

描述(Description)

無需透過複雜演算法去計算，此描述指的是透過人類對資料的敏銳度及資料視覺化的呈現，亦可同時找出資料的潛在規則。

2.4.4 資料探勘分類模式

資料探勘原型為預測，預測類別變數(不連續)的過程稱為分類，而預測連續變數(連續數值)的過程稱為推估。本研究採預測類別變數，由於FaceBook中的中文貼文字詞為不連續值，故將以六大模式中的分類為研究重點以預測五大人格分類。

分類是指建立一個學習目標函數 f ，使得這學習函數可由 x 屬性對應至 y 類別，如圖12所示，而分類法的目的是將一個物件指定至其中一個已預設的分類中[7]。如同本研究中利用人的常用語，只取中文部份，而後加以分類，歸屬此人為五大人格中五大類的哪個類別。

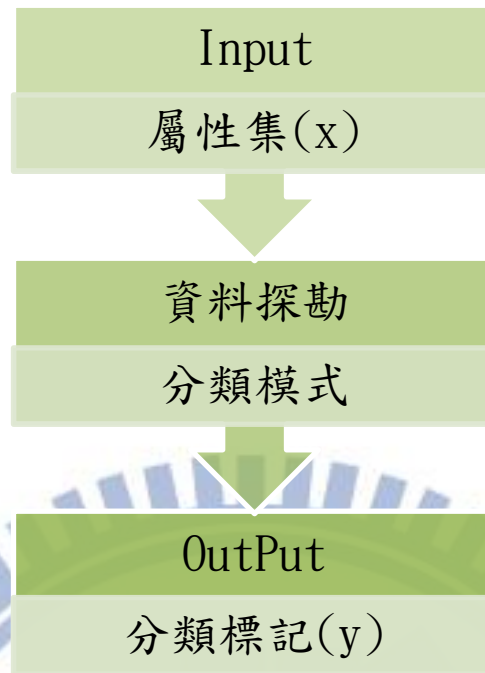


圖 12：分類模式圖
資料來源：本研究整理

2.4.5 貝氏分類法

貝氏定理由機率所推導而出，在假設條件獨立的情況下，依據已知的事件發生之機率來推測未知資料的類別。貝氏定理包含二種分類法，單純貝氏分類法及貝氏信念網路二種。

貝氏分類法中，其中一方法為單純貝氏定理也是本研究參考部份理論之研究範圍之一，因以往研究者研究顯示單純貝氏判別對於大量資料分析具有高效率與高準確率。

2.4.6 單純貝氏分類法

單純貝氏定理是獨立假說，也就是假設屬性間對同一類別之影響力是獨立的，而由於計算簡化，故稱為單純的原因[8]。

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

公式(1)表示 $P(H|X)$ 是在 X 條件下 H 的事後機率

$P(H)$ 為假設 H 的事前機率

要驗證事後機率的正確性不容易，需大量訓練資料或適量屬性資訊，貝氏理論有一定有效度及準確度的原因是因為貝氏允許以事前機率 $P(H)$ 、 $P(X)$ 及類別的條件機率 $P(H|X)$ 為主，來預測事後的機率[7]，如上公式(1)。

單純貝氏分類法為假設在類別 y 中，屬性與屬性間具有條件獨立的特性，依此來計算機率值[7]。條件獨立公式如下公式(2)。

$$P(X|Y=y) = \prod_{i=1}^d P(X_i|Y=y) \quad (2)$$

$X=\{X_1, X_2, \dots, X_d\}$ 包含 d 個屬性值

條件獨立

假設 X 、 Y 、 Z 是隨機變數集合， X 中的變數在 Z 情況下和 Y 是相互獨立的。條件如下公式(3)。

$$P(X|Y,Z) = P(X|Z) \quad (3)$$

本研究方法二的公式將採用變型之單純貝氏分類法的方式預測其五大人格分類，由於研究分析之條件欄位並非多個，僅以詞彙分類，故無需耗用貝氏分類法實作，但研究方法會參考貝氏分類法之機率原則加以分類，將於後續第三章之設計與方法中介紹。

三、設計與方法

3.1 系統架構(方法一：關鍵字詞預測法)

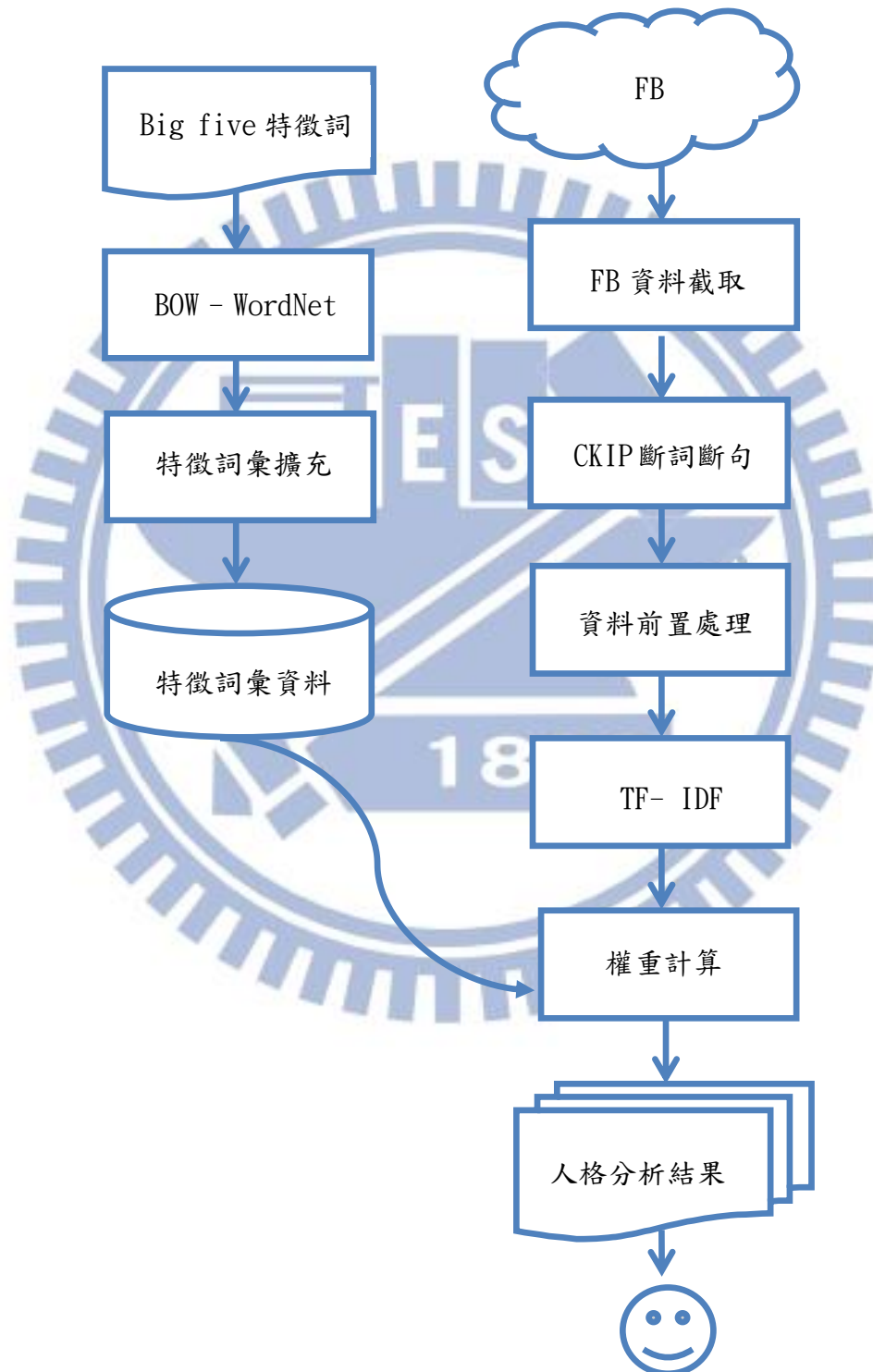


圖 13：系統架構(方法一)

資料來源：此圖本研究自行整理

3.1.1 Big Five 特徵詞彙

依據研究中最廣為接受的 McCrae 和 Costa (1985) 五大人格特質（開放型、嚴謹型、外向型、隨和型、神經質型）此五類人格特質依諸位研究人員所描述此五大類特徵，經整理後共 403 個特徵詞[12][13][21][22][23]。

3.1.2 BOW - WordNet

本研究之方法一選用「中央研究院中英雙語知識本體詞網」(The Academia Sinica Bilingual Ontological Wordnet (BOW))，因 BOW 之語言座標，是以台灣通用的中文為出發點。且 BOW 之開發成員及資料來源來自國內外知名之知識團隊，為本研究所用之因。目前此網站開放使用的系統功能是由「數位典藏」國家型計畫下的「語言座標」計畫所建構完成。而所謂語言座標指可讓不同來源的典藏知識內容，可以轉換成互通的訊息，將成為下一代語意網中不可或缺的基礎架構。BOW 引用的資料除了中央研究院詞庫小組（資訊所），文獻語料庫（語言所）及計算中心開發的資料外。國外則有 IEEE 批准執行的 SUMO 知識本體(teknowledge.com 管理)，及普林斯頓 (Princeton University) 的 WordNet。國內主要有來自遠見科技股份有限公司，包括該公司自有資料及與中研院共同開發資料，以及教育部國語會的辭典[25]。

3.1.3 特徵詞彙擴充

本研究依所整理之特徵詞，透過中央研究院中英雙語知識本體詞網- WordNet 網站，如圖 14 及圖 15 所示，利用中英對照，將 403 個特徵詞，依同義詞/上位詞/下位詞，擴充至 2731 個特徵詞。不考慮反義詞，因學者所提出之潛意識思考理論，人的第一反應出來的行為，即來自於大腦中潛意識所驅使，所呈現於行為、言語、文字中。例如一個外向型的人，平常用語會多以呈現快樂的字眼，當他不快樂時，會以呈現我” 不快樂”，而較不會呈現我很” 憂鬱” 等字眼，但平常就比較會想到負面情況的人，例如神經質型的人，他的使用字眼會以” 憂鬱” 呈現，而較不會以” 不快樂” 呈現。故本研究不考慮反義詞的原因在於此。

利用 BOW 特徵詞擴充，例如：“仁慈”於中文查詢，得到同義詞有“寬大”、“慈愛”等上位詞有“寬恕(leniency)”，而下位詞有“恩赦(amnesty)” [18]。

Click or Sense title for more information. 請點選 或詞義標題以獲取更多的訊息

Sense map with other version 同義詞集和其他版本對應	WordNet1.7.1
仁慈 map with synset of other version 該詞義中的仁慈和其他版本對應	仁慈@WordNet1.7.1
Domain 領域	General(一般) 建議領域值
POS 詞類	Noun(名詞)
Explanation 解釋	leniency and compassion shown toward offenders by a person or agency charged with administering justice
Translation 翻譯	mercy , mercifulness , clemency
Synset 同義詞集	仁慈, 寬大
Hypernym 上位詞	leniency ^[1.6] , lenience ^[1.6]
Hyponym 下位詞	amnesty ^[1.6] , re-sentencing ^[1.6] , reprieve ^[1.6] , respite ^[1.6] , pardon ^[1.6] , free_pardon ^[1.6] , commutation ^[1.6]

資料來源：中央研究院中英雙語知識本體詞網

詞彙網路

WordNet

[最新消息](#)
[系統簡介](#)
[詞網](#)
[知識本體](#)
[索引](#)
[檔案櫃](#)
[相關連結](#)
[回首頁](#)

英文查詢 Eng-Ch LookUp

檢索條件值 Keyword : leniency@WordNet1.6

請點選 ☐ 或詞義標題以獲取更多的訊息. Click ☐ or Sense title for more information.

詞義(Sense)1: 寬恕

同義詞集和其他版本對應 Sense map with other version	WordNet1.7.1
該詞義中的leniency和其他版本對應 leniency map with synset of other version	leniency@WordNet1.7.1
領域 Domain	General(一般) 建議領域值
POS 詞類	Noun(名詞)
解釋 Explanation	lightening a penalty or excusing from a chore by judges or parents or teachers
翻譯 Translation	寬恕
同義詞集 Synset	leniency , lenience
上位詞 Hypernym	tolerance [1.6]
下位詞 Hyponym	mercy [1.6] , mercifulness [1.6] , clemency [1.6]

圖 15：WordNet- 英文查詢介面
資料來源：中央研究院中英雙語知識本體詞網

3.1.4 特徵詞彙資料庫

將擴充後之特徵詞彙存入 SQL Server 2008 資料庫中，供後續抓取特徵詞來預測五大人格類別使用。特徵詞彙資料量越大，則分類準確度越大，故建議可擴充更多的特徵詞彙於資料庫中。

3.1.5 FB 資料截取

文字探勘，是以抓取文字為主，為將非結構化資料處理成有意義的資料，文字字句長短不一，資料的記載是很自由的，這些自由文句在做分類時，常將內容資料切成一個一個語意單位，這些語意單位被稱為特徵詞彙，可從這些特徵詞彙與類別來找出彼此特徵詞彙之間的對應關係，可應用於文字探勘技術自動處理，例如自動分類等之應用[11]。本研究以文字探勘為研究範圍。

本研究透過 C#.Net 程式開發，自動抓取 Facebook 朋友之相關資訊記載於 SQL Server 2008 資料庫中。截取 FB 資料之程式開發步驟，如表 4 所示。

截 取 FB 開 發 步 驟	1. 設定 APP ID
	2. 設定使用權限
	3. C# Web Browser 元件登入及取得 Access Token
	4. 透過 Graph API 找朋友
	5. 找朋友的貼文等資訊

表 4：截取 FB 資料之程式開發步驟

資料來源：本研究自行整理

3.1.6 CKIP 斷詞斷句

本研究使用中研院之 CKIP 斷詞斷句，截取資料利用所提供之工具計算各詞彙之詞頻[20][26]。形容五大人格分類詞性以動詞及形容詞的詞性為主，故只取動詞及形容詞的詞類為 A、ADV、Nv、Vi、Vt 五種詞類，詞類如下圖 17(刪除線為本研究排除之詞類，而詳細之中研院平衡語料庫詞類標記集請參照附錄一)。而資料前置處理則只取中文字詞，依據 CKIP 計算後結果，排除無需之詞類、單一字元、英文、標號後，由最初的 93636 筆的詞彙，減少為 45467 筆的詞彙，詞類排除非 A、ADV、Nv、Vi、Vt 五種詞類後，最終取 32880 筆詞彙，而後依據每筆詞彙計算 TFIDF 將資料存入資料庫中。中研院所提供之 CKIP 斷詞斷句功能，需先申請服務帳號後，才可執行所下載之 CKIPClient 及 CountWordFreq 程式，中研院所提供的中文斷詞系統之服務畫面如下圖 16。

- ➔ 簡介
- ➔ 未知詞擷取做法
- ➔ 詞類標記列表
- ➔ 線上展示
- ➔ 線上服務申請
- ➔ 線上資源
- ➔ 公告
- ➔ 聯絡我們

- 下載詞類統計程式
 - 最新更新日期：2011/11/30
 - 本程式目的是幫助統計斷詞結果
 - 請在命令列下輸入：CountWordFreq.exe [InputFile] [OutputFile] [0/1]
 - [InputFile]：輸入檔為以tab字元區隔的斷詞文字，詞性可(不)加註在詞後面，以括號括起來
 - [OutputFile]：輸出文字中的所有詞，以及其出現的次數
 - [0/1]：統計時是否要連同Part Of Speech一起考慮。0:Without Part Of Speech 1:With Part Of Speech
 - 如需排序可以將結果貼在EXCEL檔
- 斷詞服務客戶端程式
 - 下載斷詞服務客戶程式(console)
 - 最新更新日期：2012/01/01
 - 支援big5與utf-16le編碼文章
 - 需安裝Java runtime enviroment 1.5以上版本
 - 請在命令列下使用，使用說明請見壓縮檔內 ReadMe.txt 說明

圖 16：中研院中文斷詞系統

資料來源：中研院中文斷詞系統—線上資源

中研院平衡語料庫詞類標記集

精簡詞類	簡化標記	對應的 CKIP 詞類標記	
A	A	A	/*非謂形容詞*/
C	Caa	Caa	/*對等連接詞，如：和、跟*/
POST	Cab	Cab	/*連接詞，如：等等*/
POST	Cba	Cbab	/*連接詞，如：的話*/
C	Cbb	Cbaa, Cbba, Cbbb, Cbaa, Cbeb	/*關聯連接詞*/
ADV	Da	Daa	/*數量副詞*/
ADV	Dfa	Dfa	/*動詞前程度副詞*/
ADV	Dfb	Dfb	/*動詞後程度副詞*/
ASP	Di	Di	/*時態標記*/
ADV	Dk	Dk	/*句副詞*/
ADV	D	Dab, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	/*副詞*/
N	Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
N	Nb	Nba, Nbe	/*專有名稱*/
N	Ne	Nea, Neb, Nee, Nee	/*地方詞*/
N	Ned	Neda, Nedb	/*位置詞*/
N	Nd	Ndaa, Ndab, Nde, Ndd	/*時間詞*/
DET	Nau	Nau	/*數詞定詞*/
DET	Nee	Nee	/*特指定詞*/
DET	Nep	Nep	/*指代定詞*/
DET	Neqa	Nega	/*數量定詞*/
POST	Neqb	Negb	/*後置數量定詞*/
M	Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	/*量詞*/
POST	Ng	Ng	/*後置詞*/
N	Nh	Nhaa, Nhab, Nhae, Nhb, Nhe	/*代名詞*/
Nv	Nv	Nv1, Nv2, Nv3, Nv4	/*名物化動詞*/
T	I	I	/*感嘆詞*/
P	P	P	/*介詞*/
T	T	Ta, Tb, Te, Td	/*語助詞*/
Vi	VA	VA11,12,13,VA3,VA4	/*動作不及物動詞*/
Vt	VAC	VA2	/*動作使動動詞*/
Vi	VB	VB11,12,VB2	/*動作類及物動詞*/
Vt	VC	VC2, VC31,32,33	/*動作及物動詞*/
Vt	VCL	VC1	/*動作接地方宜語動詞*/
Vt	VD	VD1, VD2	/*雙賓動詞*/
Vt	VE	VE11, VE12, VE2	/*動作句賓動詞*/
Vt	VF	VF1, VF2	/*動作謂賓動詞*/
Vt	VG	VG1, VG2	/*分類動詞*/
Vi	VH	VH11,12,13,14,15,17,VH21	/*狀態不及物動詞*/
Vt	VHC	VH16, VH22	/*狀態使動動詞*/
Vi	VI	VI1,2,3	/*狀態類及物動詞*/
Vt	VJ	VJ1,2,3	/*狀態及物動詞*/
Vt	VK	VK1,2	/*狀態句賓動詞*/
Vt	VL	VL1,2,3,4	/*狀態謂賓動詞*/

圖 17：語料庫詞類標記取得

資料來源：中研院中文斷詞系統—詞類標記

3.1.7 TF- IDF

詞頻 (term frequency, TF) 指某一個詞彙出現在文章中的頻率。

TF 如下公式 (4)

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (4)$$

分子指詞彙在文件 d_j 中出現的次數，而分母為文件 d_j 中所有字詞出現次數總和

逆向文件頻率(Inverse Document Frequency, IDF) 指某詞彙出現於多少篇文章數，分之總文章數，某詞彙算出之 IDF 越大，則說明此詞彙具有很好的類別區分能力。

IDF 如下公式 (5)

$$idf_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|} \quad (5)$$

分子 D 表示語料庫中的文件總數

分母表示包含詞彙 t_i 的文件數目

TFIDF (term frequency - inverse document frequency) 是一種常用的加權技術，常被使用資訊檢索及本文挖掘上。表示就是如果某個詞彙在某篇文章中出現的頻率高，且在其他篇文章中較少出現時，則認為此詞彙具有很好的類別區分能力，較適合用來做分類，可用來計算某個詞彙在文章中的相對重要程度，防止的影響力過強。

TFIDF 如下公式 (6)

$$tfidf_{i,j} = tf_{i,j} \times idf_i \quad (6)$$

3.1.8 五大人格預測

五大人格預測公式如下公式(7)

$$P = \operatorname{argmax}_{c \in C} \left(\sum_{i=1}^n f(X_c, i) \right) \quad (7)$$

P 為計算後之最大類別。 C 為五大分類集合 $\{C, A, N, O, E\}$ 。 V_c 為某人格的特徵詞彙。 I 為受測者使用過的詞彙。 $\{X_{c1} \sim X_{cn}\} \in I \cap V_c$ 。 $f(X_c, i)$ 為各類別特徵詞彙之 TFIDF 值。 i 為詞彙數，由 1 至 n 個。

3.2 系統架構(方法二：機器學習預測法)

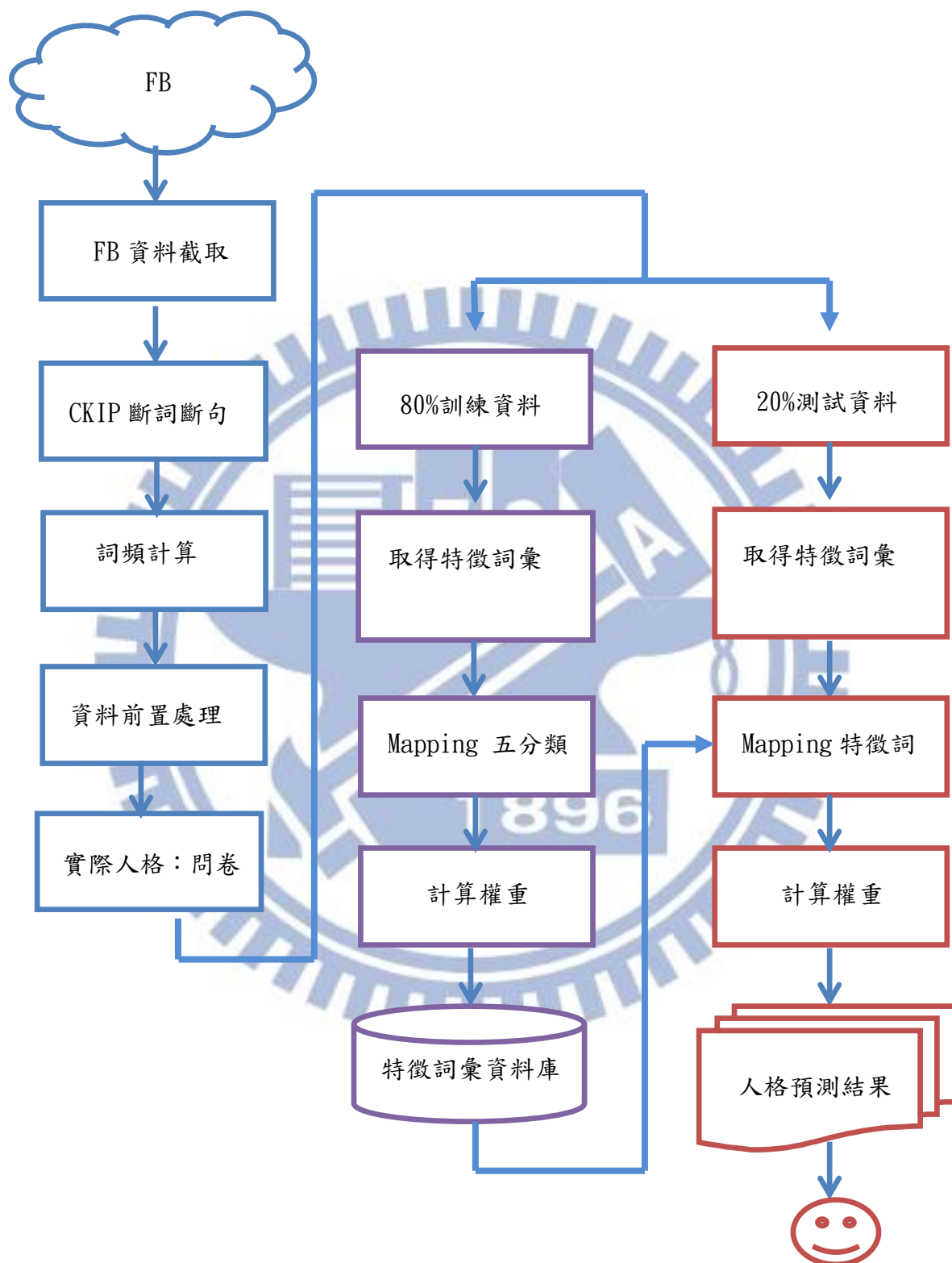


圖 18：系統架構(方法二)

資料來源：此圖自行整理

機器學習理論主要是透過設計、分析使得計算機可以自動學習，可從數據中自動分析獲得規律，以預測未知數據，已廣泛的應用，例如：資料探勘[20]。

本研究運用資料探勘技術透過計算機可以自動學習，從未分類數據中訓練學習獲得分類數據，以預測未知人格分類，故方法二以機器學習預測法命名。

3.2.1 系統架構（方法二）流程說明

步驟一：透過自行開發程式，截取 Facebook 朋友資訊，接著透過中研院 CKIP 程式執行斷詞斷句後計算詞頻，之後去除單一字元、標點符號及只取動詞及形容詞。

步驟二：問卷採用心理測驗結果為抽樣者實際人格特質。

步驟三：將資料抽樣切成 80% 為訓練集及 20% 為測試集。

步驟四：將 80% 的抽樣訓練集資料，依據抽樣者個人實際人格類別結果，計算詞頻權重，取得每個特徵詞彙的詞類，將之維度降低為五大分類為 C、A、N、O、E，將結果寫入特徵詞彙資料庫中。

C：嚴謹型(Conscientiousness)

A：隨和型(Agreeableness)

N：神經質型(Neuroticism)

O：開放型(Openness)

E：外向型(Extraversion)

步驟五：接著將 20% 的測試集資料截取特徵詞對應於特徵詞彙資料庫中是何詞彙，取得特徵分類。

步驟六：依據詞頻計算權重後得出人格分析結果。

3.2.2 方法二（機器學習預測法）演算

步驟一：

依據抽樣訓練者個別詞彙及人格類別出現的次數，演算詞彙歸屬五大人格分類的次數，以及詞彙被每一抽樣者使用的次數。

V ：詞彙 (Vocabulary) $\in \{V_1, V_2, V_3, \dots, V_n\}$

T ：人格分類(Type) $\in \{C, A, N, O, E\}$

VC_t ：人格分類的詞計量，亦為詞彙在某一人格(Type)出現的次數，如下公式(8)

$$VC_t = \text{Count}(V_t) \quad (8)$$

$$t \in T$$

Count：抽樣者使用此詞彙次數

步驟二：

依據步驟一計算後的值，將數值正規化(Normalize)，使之值介於 0~1 之間，以及算出每個詞彙的權重值。

計算每個詞彙於各分類正規化後的值，如下公式(9)

$$VN(\text{Normalize 詞彙值}) : \frac{VC_t}{\sum_n VC_n} \quad (9)$$

$$n \in T$$

$$t \in T$$

例如： $VN_A = \frac{VC_A}{\sum_n VC_n}$ ，為取得類別 A 正規化後的數值。

而各詞彙於分類後正規化的值加總應為 1，如下公式(10)

$$\sum VN = 1 \quad (10)$$

計算每個詞彙權重值，為每一詞彙出現的次數佔總次數的比重，如下公式(11)

$$V_w = \frac{VC}{\sum_i V_i C} \quad (11)$$

W：權重

$$i \in \{\text{所有出現的詞彙}\}$$

VC = 每個詞彙出現的次數

步驟三：

測試集內的人員五大人格預測，取測試集內依每個人所用詞彙的詞頻、各類別正規化後的值、詞彙權重三者相乘積後的總和，如公式(12)。最後取最大值的類別，為預測結果。

五大人格預測如下公式(13)

$$Tg_t = \left(\sum_i (Count(V_i) \times VN(V_i) \times W_i) \right) \quad (12)$$

$$P_t = \text{MAX}(Tg_t) \quad (13)$$

P：人格預測(Personality prediction)

T：人格分類(Type), $t \in \{C、A、N、O、E\}$

$V = \{V_1, V_2, V_3, \dots, V_n\}$ 受測者使用過且出現在詞彙庫裡的詞彙集合。

T_g ：人格分類級別(grade of type)

$t \in T$

i：抽樣者所有使用過的詞彙

W：權重

$VN(V_i)$ ： V_i 正規化(Normalize)後的詞彙值

3.3 問卷來源

本研究依據 103 位朋友，實際回收 75 份心理測驗結果，回收率 73%，研究結果以五大人格模式之顯著人格為研究重心，也就是取心理測驗結果高分者之實際人格與於 Facebook 預測主要人格相驗證。

驗證數據之心理問卷資料，取自英國著名的心理學教授——李察·韋斯曼 (Richard Wiseman) 於 2009 年著作之書籍。Richard Wiseman 研究領域致力於運用科學方法，研究許多日常生活裡看似無法以理性角度去理解的一些行為，以個性這章節而言，則以直覺的方式用心理學科學方法進行研究探討。

Richard Wiseman 目前於英國賀福郡大學心理系研究單位任職，著有多本心理學相關暢銷書，Richard Wiseman 的研究領域在許多世界級的頂尖學期刊都報導過，亦在多個國際媒體以專題方式介紹過，如：《時代雜誌》、《每日電訊報》等。在 BBC 等著名媒體上對成千或上萬研究對象進行大規模實驗，也曾上過金氏世界記錄，為一本有信度的科學書籍[9]。心理問卷格式如附錄二。

3.4 軟體及平台使用

1. FaceBook
2. SQL Server 2008

3. Visual Studio 2008
4. 中研院 CKIPClient
5. 中研院 CountWordFreq
6. 中研院中英雙語知識本體詞網



四、系統實作及資料驗證

本研究利用二種研究方法來實作及驗證其準確率，採用方法一為關鍵字詞預測法及方法二為機器學習預測法二種。

方法一：關鍵字詞預測法，透過 BOW 手動擴充詞彙的方式進行研究，且因每個人有可能會有多重人格表現，故將資料驗證部份再細分為顯著人格分析及多重人格分析二種方式驗證其準確率。

顯著人格分析：取心理測驗分數，有達到高分者，且取最顯著之人格分類。

多重人格分析：取心理測驗分數，有達到高分之所有分類。

方法二：機器學習預測法，使用機器學習的方式，將抽樣資料分為 80% 為訓練集，20% 為測試集，訓練取 Facebook 中文貼文內容常用詞彙，存於詞彙資料庫中，以供測試資料預測其人格分類。

4.1 Facebook 資料截取及處理

Facebook 程式開發及資料處理儲存步驟，如下表 5 所示。

步驟	描述	範例圖示
步驟一	程式開發截取朋友 Facebook 之中文貼文。	GUI 畫面 如圖 19 所示
步驟二	依個人資料分別存至各個 Facebook 中文貼文資料檔 (.txt)。	Facebook 中文貼文 資料檔內容 如圖 20 所示
步驟三	Parser 中文貼文資料檔，截取關鍵字 message、story、place 對應之中文內文。	如圖 21 所示

表 5：Facebook 資料截取及處理步驟

資料來源：本研究自行整理

圖 19 截取朋友 Facebook 之中文貼文功能解說

Log In：Login 使用 Facebook graph API 授權。

Get Message：資料截取每個朋友 Facebook 貼文 story、place、message 三種行為內容，程式依每個朋友各自產生對應的檔案。

Save Dir：指定產生後的檔案儲存路徑。



圖 19：Facebook 中文貼文資料截取

資料來源：此圖自行整理

透過圖 19 程式產生之文字檔格式內容如下圖 20 所示

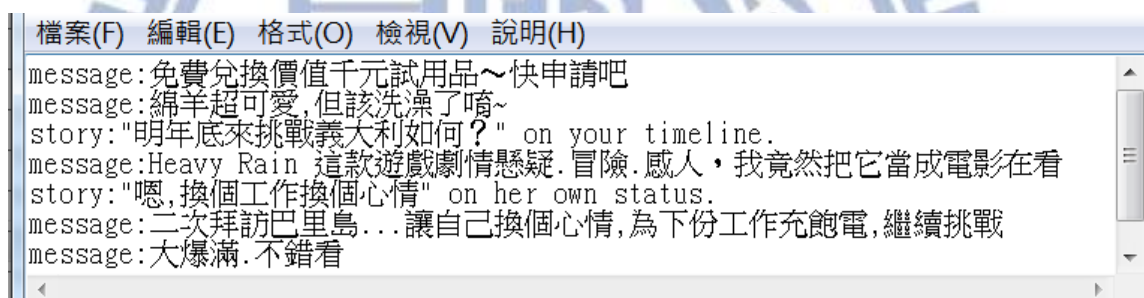


圖 20：Facebook 中文貼文資料檔

資料來源：此圖自行整理

Parser 資料，如圖 20 內容，有 story、message、place 三個關鍵字的內容，此三個關鍵字詞的內容，代表朋友曾貼文的所有文字記載，為本研究將研究的詞句，故將內容截取後存至 parser folder 依每個人獨自分開產生 parser 後的文字檔，parser 後的文字檔內容如圖 21 所示。

以下針對 story、message、place 三個關鍵字記載的內容加以描述如表 6 所示。

Graph API	描述
Story	主題貼文之下的貼文回覆
Message	主題貼文內容描述
Place	地方打卡的文字描述

表 6：FB story、message、place 功能描述

資料來源：本研究自行整理

免費兌換價值千元試用品~快申請吧
 綿羊超可愛,但該洗澡了唷~
 還在2月還來的及~Happy Birthday to you~
 還好今天還在2月,生日快樂唷~
 3/3 Happy Birthday~
 Happy birthday~
<http://www.ntsec.gov.tw/User/Article.aspx?a=1027>
 謝謝~
 謝謝學姐的祝福~
 明年底來挑戰義大利如何?
 果然是懷舊風~
 Heavy Rain 這款遊戲劇情懸疑.冒險.感人,我竟然把它當成電影在看

圖 21：截取 message、story、place 對應內文

資料來源：此圖自行整理

產生欲處理的檔案之後，開始呼叫 CKIPClient 字詞切割程式去執行每個文字檔的斷詞斷句，執行完後產生的檔案內容如圖 22 所示。

```
免費(Vi) 兌換(Vi) 價值(N) 千(DET) 元(M) 試用品(N) ?(QUESTIONCATEGORY) 快(N) 申請(Vi) 吧(T)|
綿羊(N) 超(Vi) 可愛(Vi) ,(COMMACATEGORY)
但(C) 該(ADV) 洗澡(Vi) 了(T) 唷(T) ~(FW)
還(ADV) 在(P) 2月(N) 還(ADV) 來(Vi) 的(T) 及(C) ?(QUESTIONCATEGORY) Happy(FW) Birthday(FW) to(FW) you(FW) ~(FW)
還好(ADV) 今天(N) 還(ADV) 在(P) 2月(N) ,(COMMACATEGORY)
生日快樂(Vi) 唷(T) ?(QUESTIONCATEGORY)
3(DET) /FW) 3(DET) Happy(FW) Birthday(FW) ~(FW)
Happy(FW) birthday(FW) ~(FW)
明年(N) 底(POST) 來(ADV) 挑戰(Vi) 義大利(N) 如何(Vi) ?(QUESTIONCATEGORY)
果然(ADV) 是(Vi) 懷舊(Vi) 風(N) ?(QUESTIONCATEGORY)
Heavy(FW) Rain(FW) 這(DET) 款(M) 遊戲(N) 劇情(N) 懸疑(Vi) ,(PERIODCATEGORY) 冒險(Vi) ,(PERIODCATEGORY) 感人(Vi) ,(COMMACATEGORY)
我(N) 竟然(ADV) 把(P) 它(N) 當成(Vi) 電影(N) 在(P) 看(Vi)
2012(DET) 國慶(N) 煙火(N) 在(P) 苗栗(N)
嘿(T) ,(COMMACATEGORY)
換(Vi) 個(M) 工作(N) 換(Vi) 個(M) 心情(N)
二(DET) 次(M) 拜訪(Vi) 巴里島(N) ...,(ETOCATEGORY) 讓(Vi) 自己(N) 換(Vi) 個(M) 心情(N) ,(COMMACATEGORY)
為(P) 下(DET) 份(M) 工作(N) 充飽(Vi) 電(N) ,(COMMACATEGORY)
繼續(Vi) 挑戰(Vi)
這(DET) 圖(N) 代表(Vi) 了(ASP) "(FW) 普羅米修斯(N) "(FW) 的(T) 意含(N) ~(FW)
我(N) 昨天(N) 去(ADV) 看(Vi) 的(T) ,(COMMACATEGORY)
那(DET) 間(M) 是(Vi) 大廳(N) 的(T) ,(COMMACATEGORY)
連(C) 第一(DET) 排(M) 都(ADV) 坐滿(Vi) ,(COMMACATEGORY)
誇張(Vi) ~(FW)
大(Vi) 爆滿(Vi) ,(PERIODCATEGORY) 不錯(Vi) 看(Vi)
```

圖 22：CKIPClient 字詞切割

資料來源：此圖自行整理

接下來針對斷詞斷句後的結果，呼叫 CountWordFreq 程式計算各字詞詞頻，產生後之檔案內容如圖 23 所示。

檔案(F) 編輯(E) 格式(O) 檢視(V) 說明(H)	
免費(Ui)	1
冒險(Ui)	1
底(POST)	1
心情(N) 2	
快(N) 1	
意含(N) 1	
感人(Ui)	1
懷舊(Ui)	1
懸疑(Ui)	1
我(N) 2	
把(P) 1	
拜訪(Ut)	1
挑戰(Ut)	2
排(M) 1	
換(Ut) 3	
明年(N) 1	
昨天(N) 1	
是(Ut) 2	
果然(ADU)	1

圖 23：CountWordFreq 詞頻計算

資料來源：此圖自行整理

4.2 資料驗證 - 心理測驗結果

驗證部份透過 Richard Wiseman (2009) 著作之一本有信度之心理測驗試題，依此測驗結果為依據計算其準確率。

問卷平台使用，選用 Google Doc 來製作，因 Google Doc 是簡易設計且易上手的線上問卷製作平台。運用此設計心理測驗，供朋友快速完成測驗，不需超過 1 分鐘即可完成一份 10 題的心理測驗，且 Google Doc 於使用者點選”提交”時，即可即時回傳至設計者 server 端，供即時查閱，可快速達到心理測驗的結果供後續研究分析。

本研究資料驗證採用 Google Doc 製作心理測驗試題，其格式如下圖 24，至於試題完整內容，請參照附錄二。

心理測驗

****以下10題測驗問題：請衡量你目前的概況，點選較符合的程度****
以“我覺得我是...”回答以下問題

註：測驗結果僅供學術分析統計使用，不外流個人測驗結果供其它用途，
亦不提供個人測驗結果於學術文件中，僅以全數統計結果顯示。

***必填**

1. 派對上的靈魂人物 *

- ☒ 極不同意
- ☐ 不同意
- ☐ 有點不同意
- ☐ 沒意見
- ☐ 有點同意
- ☐ 同意
- ☐ 非常同意

2. 對他人不太關心 *

- ☐ 極不同意
- ☐ 不同意
- ☐ 有點不同意
- ☐ 沒意見
- ☐ 有點同意
- ☐ 同意
- ☐ 非常同意

圖 24：心理測驗問卷格式

資料來源：此圖參照[9]後自行整理

心理測驗結果所回覆的內容及格式如下圖 25 所示。

1. 派對上的表演	2. 對他人不太關心	3. 總是做好準備	4. 容易感到壓力大	5. 有部佳點子	6. 話不多	7. 對人感興趣	8. 常忘記東西放回原位	9. 大多時間都很放鬆	10. 難以理解抽象概念
極不同意	極不同意	極不同意	極不同意	極不同意	極不同意	極不同意	極不同意	極不同意	極不同意
極不同意	極不同意	同意	同意	沒意見	同意	有點同意	不同意	不同意	不同意
有點不同意	極不同意	有點同意	沒意見	沒意見	極不同意	非常同意	沒意見	沒意見	沒意見
沒意見	有點同意	有點不同意	有點同意	沒意見	沒意見	沒意見	有點不同意	有點不同意	有點不同意
沒意見	不同意	沒意見	不同意	同意	不同意	同意	有點不同意	同意	極不同意
沒意見	極不同意	同意	不同意	同意	不同意	同意	有點同意	有點同意	有點同意
不同意	不同意	同意	不同意	同意	有點不同意	有點不同意	同意	有點同意	同意
沒意見	不同意	沒意見	有點同意	有點同意	有點同意	有點同意	有點不同意	有點不同意	有點同意
有點同意	不同意	有點同意	極不同意	同意	沒意見	沒意見	有點同意	非常同意	不同意
不同意	不同意	同意	非常同意	有點不同意	不同意	不同意	不同意	沒意見	沒意見
不同意	沒意見	有點同意	同意	沒意見	有點不同意	有點不同意	不同意	沒意見	非常同意
不同意	極不同意	同意	同意	沒意見	非常同意	非常同意	有點不同意	有點不同意	有點同意
不同意	不同意	同意	同意	非常同意	同意	非常同意	沒意見	有點不同意	不同意
同意	不同意	有點同意	不同意	有點同意	不同意	非常同意	沒意見	同意	同意
有點同意	不同意	有點同意	有點同意	沒意見	沒意見	同意	同意	沒意見	沒意見
不同意	有點同意	非常同意	有點不同意	有點不同意	有點不同意	有點同意	不同意	不同意	不同意
不同意	有點同意	有點不同意	同意	有點同意	同意	不同意	不同意	不同意	不同意
有點不同意	有點不同意	同意	有點不同意	同意	同意	同意	有點不同意	同意	不同意
不同意	不同意	同意	沒意見	非常同意	同意	沒意見	不同意	同意	沒意見
不同意	有點同意	同意	同意	有點不同意	同意	同意	不同意	沒意見	有點不同意
有點不同意	有點同意	同意	同意	同意	同意	同意	不同意	沒意見	有點不同意
不同意	有點同意	同意	同意	同意	同意	同意	不同意	沒意見	有點不同意
有點不同意	有點不同意	同意	同意	同意	同意	同意	不同意	沒意見	有點不同意
有點不同意	有點不同意	同意	有點同意	有點同意	有點同意	沒意見	沒意見	有點不同意	有點不同意
有點不同意	有點不同意	同意	有點同意	有點同意	有點同意	有點不同意	同意	同意	同意
有點同意	不同意	有點同意	同意	有點同意	有點同意	同意	同意	不同意	有點同意
有點同意	不同意	有點同意	同意	有點同意	有點同意	同意	同意	不同意	不同意
不同意	沒意見	有點同意	同意	同意	同意	同意	同意	同意	沒意見
有點同意	不同意	有點同意	同意	同意	沒意見	同意	沒意見	沒意見	有點不同意
沒意見	極不同意	同意	同意	同意	同意	同意	不同意	沒意見	不同意
同意	沒意見	同意	同意	沒意見	有點不同意	沒意見	不同意	有點不同意	不同意

圖 25：心理測驗問卷回覆

資料來源：此圖自行整理

依據心理測驗結果計算取得顯著人格及多重人格分類，如圖 26 所示。

O(開放型)	C(嚴謹型)	E(外向型)	A(隨和型)	N(神經質型)	顯著人格	多重人格
	1		2	3	,,N	,C,,A,N
		1	4	,,A		,,E,A
				1	,,N	,,,N
3		1	2	O,,,	O,,,	O,,E,A
			1	1	,,A,N	,,,A,N
2				O,,,		O,,,
	1			2	,,N	,C,,,N
			4	1	,,N	,,,N
				,,A		,,,A
3			3	2	O,,A	O,,A,N
		3	3	,,E,A		,,E,A
			2	,,A		,,,A
	2			,C,,		,C,,
1				3	,,N	O,,,N
3			1	O,,,		O,,A
	1			,C,,		,C,,
	1			2	,,N	,C,,,N
2	1		1	O,,,		O,C,,A
			2	1	,,N	,,,N
				3	,,N	,,,A,N
1		1	1	3	,,N	O,,E,N
				,,A		,,,A
1			2	1	,,A	O,,,A,N
2		1	3	1	,,A	O,,E,A,N
	1	2		2	,,E,N	,C,E,N
	1		2	,,A		,C,,A
			3	,,A		,,,A
			1	,,A		,,,A
			3	,,A		,,,A
1	1		1	O,C,,A		O,C,,A

圖 26：心理測驗結果分類

資料來源：此圖自行整理

4.3 方法一(關鍵字詞預測法)

依計算詞頻後的字詞，取詞性為動詞及形容詞，另外去除單一字元及標點符號之後計算 TFIDF，TFIDF 計算公式，請參照公式(6)。最後將 id、term、TFIDF 資料存至資料庫中，如圖 27 所示。

由於個資法，基於資料保護的原則，本研究中將以真實 id 以 Mapping 成流水號的方法顯示，以保障個人資料的安全性。



	ID	term	TF-IDF
1	000001	不錯	0.0204080000
2	000001	代表	0.0204080000
3	000001	充飽	0.0204080000
4	000001	兌換	0.0204080000
5	000001	免費	0.0204080000
6	000001	冒險	0.0204080000
7	000001	可愛	0.0204080000
8	000001	坐滿	0.0204080000
9	000001	如何	0.0204080000
10	000001	感人	0.0204080000
11	000001	懷舊	0.0204080000
12	000001	懸疑	0.0204080000
13	000001	拜訪	0.0204080000
14	000001	挑戰	0.0408160000
15	000001	果然	0.0204080000
16	000001	洗澡	0.0204080000
17	000001	爆滿	0.0204080000
18	000001	生日快樂	0.0204080000
19	000001	申請	0.0204080000
20	000001	當成	0.0204080000
21	000001	竟然	0.0204080000

圖 27：方法一個人用詞資料檔

資料來源：此圖自行整理

透過中央研究院中英雙語知識本體詞網- WordNet 網站特徵詞擴充，利用中英對照將 403 個特徵詞，依同義詞/上位詞/下位詞，擴充至 2731 個特徵詞，最後減低維度，將之歸屬五大類別。擴充後結果透過自行開發之 importer 程式，如圖 28，將檔案讀取及解析後，此支程式將會把結果存於特徵詞資料庫中，供後續研究使用。

圖 28 功能解說

File Path：選擇預匯入之檔案路徑

Connection String：資料庫之連結設定

Import：Parser 檔案資料後匯入資料庫中

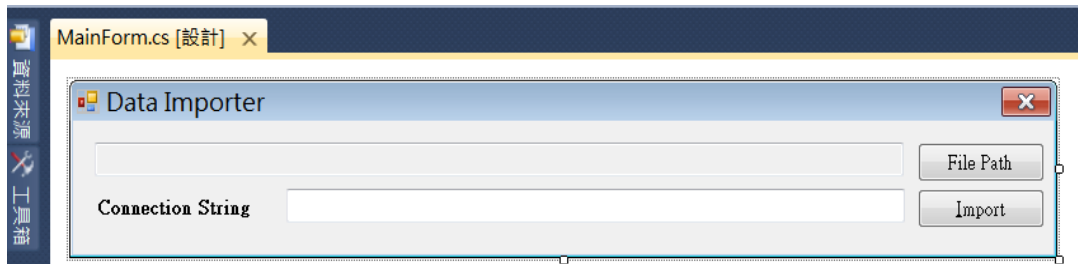


圖 28：方法一之 Importer 程式

資料來源：此圖自行整理

圖 29 為 Importer 程式匯入特徵詞資料庫的資料呈現。

結果		訊息	
	type	desc	term
2234	A	隨和型	相應
2235	A	隨和型	配合
2236	A	隨和型	照應
2237	A	隨和型	協調
2238	A	隨和型	符合
2239	A	隨和型	順應
2240	A	隨和型	適應
2241	A	隨和型	調適
2242	A	隨和型	配合
2243	A	隨和型	相合
2244	N	神經質型	憂慮
2245	N	神經質型	自憐
2246	N	神經質型	緊張
2247	N	神經質型	敏感
2248	N	神經質型	不穩定

圖 29：方法一特徵詞資料庫

資料來源：此圖自行整理

依個人用詞資料檔之字詞 Mapping 特徵詞資料庫中的字詞，以預測人格分類結果，如圖 30 所示，依個人類別取加總後之 TFIDF 最大者，則為此人之分類，此時預測出來的分類為 A(隨和型)。

結果		訊息	
	id	tfidf	type
1	000001	0.0816320000	A
2	000001	0.0612240000	O
3	000001	0.0408160000	C

圖 30：方法一人格分類實例

資料來源：此圖自行整理

依據回覆結果，計算人格分類，如圖 26 所示，由於一個人可能會有多重人格產生，故本研究將方法一之驗證結果，利用二種分析方式去驗證，採用顯著人格分析及多重人格分析二種方式驗證結果。

圖 30 為例，預測結果為五大人格分類之 A (隨和型)，於心理測驗後的人格分類，顯著人格分類為 A (隨和型)；多重人格分類為 C(嚴謹型)、A (隨和型)、N(神經質型)，如下圖。故顯著人格分類與預測結果相同，故屬預測正確；另 A (隨和型) 有落於多重人格分類之其中一分類，故多重人格分類屬預測正確。本研究方法一，以此類推算出所有朋友之人格分類後加以統計分類結果，以計算其準確率。

暱稱	O(開放型)	C(嚴謹型)	E(外向型)	A(隨和型)	N(神經質型)	顯著人格	多重人格
000001		2		3		2 ,,A,	,C,,A,N

4.4 方法二(機器學習預測法)

圖 31 為針對方法二(機器學習預測法)所開發的程式，此程式的設計原則為將資料依 80/20 法則，將有回填心理測驗問卷資料的 80%為訓練資料集，而將剩餘 20%的資料列為測試資料集，目的為了將 80%訓練後所得到的五大分類的詞彙記載於資料庫中，供 20%測試集資料預測五大人格分類。

圖 31 功能解說

Directory：選擇欲匯入方法二的文字檔目錄

Connection String：資料庫連結設定

Exec：開始執行機器學習預測法

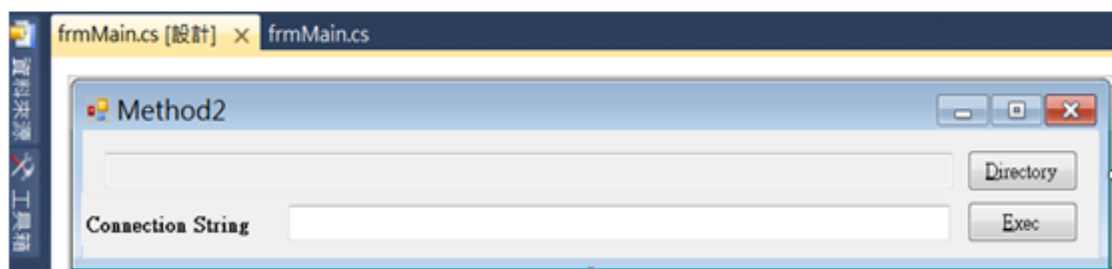


圖 31：方法二實作介面

資料來源：此圖自行整理

執行 Method2 程式前的前置動作，需先完成以下表 7 步驟。Method2 程式的 GUI 畫面如圖 31。

步驟	描述
步驟一	Facebook 資料截取
步驟二	CKIP 斷詞斷句
步驟三	詞頻計算
步驟四	心理測試問卷結果
步驟五	將心理測試問卷結果所得之人格分類，將之寫入斷詞斷句後的詞頻文字檔中，供後續 Method2 程式分析使用

表 7：方法二實作前置步驟

資料來源：本研究自行整理

首先將資料抽樣切分為 80%訓練集及 20%測試集，相關內部作業如下表 8。

資料集	步驟	描述
訓練集 80%	步驟一	批次讀取抽樣 80%已執行斷詞斷句後的文字檔案內容
	步驟二	Parser 每份訓練集文字檔，截取個人 ID、詞彙、詞頻、人格類別此四欄位寫入資料庫中
	步驟三	將維度降低為五大類別，依步驟二所產生的資料，再去計算訓練集每個詞彙分佈在五大人格分類 $T \in \{C、A、N、O、E\}$ 的總數
	步驟四	Normalize 個人五大類別權重，使數字介於 0~1 之間
測試集 20%	步驟一	批次讀取抽樣 20%已執行斷詞斷句後的文字檔案內容(不重覆訓練集資料)
	步驟二	Parser 每份測試集文字檔，截取個人 ID、詞彙、詞頻、人格類別此四欄位寫入資料庫中
	步驟三	依照步驟二該抽樣者所使用的詞彙出現次數 $\times$ 該詞彙個人人格的比重 $\times$ 詞彙出現的權重
	步驟四	依步驟三計算後的值取個人五大類中數值最大者為個人之人格分類預測結果

表 8：方法二實作步驟

資料來源：本研究自行整理

依每個人心理測驗回傳結果，經計算分數後，取最高分者，將符合的類別寫入已斷字斷詞後的文字檔的第一行，以供程式抓取後續判別使用，有可能個人最高分數有相同類別相同分數，如下圖 32 此抽樣者之最高分數 E 及 N 分數相同，故此人屬 E 及 N 二種人格。

文字檔內容資料及格式如圖 32。



圖 32：方法二匯入文字檔內容

資料來源：此圖自行整理

Method2 讀取每位朋友的文字檔內容資料，依序 Parser 相關字詞，取動詞及形容詞的詞類且資料內容以取中文字詞為主，排除無用之詞類、單一字元、英文、標號之後，寫入資料庫中的欄位包含個人 ID、詞彙、詞頻、五大人格類別，資料內容呈現如圖 33。

ID	Word	Count	Type
10...	傳回來	1	E,A
10...	吃起來	1	E,A
10...	坐月子	1	E,A
10...	多愁善感	1	E,A
10...	大排長龍	1	E,A
10...	天昏地暗	1	E,A
10...	奢人儉難	1	E,A
10...	庸人自擾	1	E,A
10...	從早到晚	1	E,A
10...	徹徹底底	1	E,A
10...	心想事成	2	E,A
10...	忙裡偷閒	1	E,A
10...	怎麼樣	1	E,A
10...	怎麼辦	1	E,A
10...	情何以堪	1	E,A
10...	打基礎	2	E,A
10...	找尋到	1	E,A
10...	曬太陽	1	E,A
10...	會不會	3	E,A
10...	有沒有	6	E,A
10...	有驚無險	1	E,A
10...	欲蓋彌彰	1	E,A

圖 33：方法二實作

資料來源：此圖自行整理

詞彙分類計算以下表 9 為例，” 快快樂樂” 這個詞彙，出現在 C 類人格有 4 次，出現在 A 類人格有 2 次，出現在 N 類人格有 3 次，出現在 O 類人格有 1 次，出現在 E 類人格有 2 次，Count 為” 快快樂樂” 詞彙總共出現的次數。

舉例來說，如果某一抽樣者同時有 E 及 N 的人格，則 VC_A 及 VC_N 各會被加 1，而 Count 當時只會被加 1，所以 $VC_C+VC_A+VC_N+VC_O+VC_E$ 與 Count 值不一定相同。

V	VC_C	VC_A	VC_N	VC_O	VC_E	Count
快快樂樂	4	2	3	1	2	8

表 9：方法二詞彙分類範例

資料來源：本研究自行整理

程式實際執行後所產生之詞彙分類結果如下圖 34 所呈現，例如詞彙為” 生日快樂” 這個字詞，在 A(隨和型)人格出現的次數是最多的，也就是隨和型最常引用此字詞祝福別人，故” 生日快樂” 於本研究中詞彙歸屬於 A 類，以此類推其它詞彙之分類統計。

結果		訊息					
	Word	N	E	O	A	C	Count
1	打電話	20	15	23	11	5	54
2	生日快樂	327	104	398	516	289	1269
3	在一起	31	11	7	25	5	44
4	有沒有	55	51	117	41	23	212
5	沒想到	68	59	7	17	2	96
6	為什麼	69	36	32	81	27	148
7	看起來	52	20	19	26	4	93
8	越來越	63	50	13	24	10	91
9	會不會	40	23	18	23	10	76

圖 34：方法二詞彙分類

資料來源：此圖自行整理

計算每個詞彙於各分類正規化後的值，及其權重值。

接續上例，以”生日快樂”這個詞彙為言，於詞彙分類計算後的數值為 N=327，E=104，O=398，A=516，C=289。

Normalize 詞彙分類數值，使之介於 0~1 之間，套用公式(9)計算如下：

詞彙”生日快樂”於五大類別加總後的值為：1634，以下針對 N、E、O、A、C 五大類計算後的數值各為

$$N=327/1634=0.200122399020808$$

$$E=104/1634=0.0636474908200734$$

$$O=398/1634=0.243574051407589$$

$$A=516/1634=0.315789473684211$$

$$C=289/1634=0.176866585067319$$

每個詞彙的權重值計算如公式(11)所示，為每個詞彙出現次數佔所有詞彙出現次數的比重，以詞彙”生日快樂”為例，生日快樂出現的次數為 1269 次，佔所有詞彙出現的總次數的比重值，算出來”生日快樂”的權重為 0.228525121555916。

計算後 Normalize 詞彙分類數值及每個詞彙的權重，如下圖 35 所示。

Word	N	E	O	A	C	Weight
生日快樂	0.200122399020808	0.0636474908200734	0.243574051407589	0.315789473684211	0.176866585067319	0.228525121555916
有沒有	0.191637630662021	0.177700348432056	0.407665505226481	0.142857142857143	0.0801393728222996	0.0381775616783721
為什麼	0.281632653061224	0.146938775510204	0.130612244897959	0.330612244897959	0.110204081632653	0.0266522600396182
沒想到	0.444444444444444	0.38562091503268	0.0457516339869281	0.111111111111111	0.0130718954248366	0.0172879524581307
看起來	0.429752066115702	0.165289256198347	0.15702479338843	0.214876033057851	0.0330578512396694	0.0167477039438142
越來越	0.39375	0.3125	0.08125	0.15	0.0625	0.0163875382676031
會不會	0.350877192982456	0.201754385964912	0.157894736842105	0.201754385964912	0.087719298245614	0.0136862956960202
打電話	0.27027027027027	0.202702702702703	0.310810810810811	0.148648648648649	0.0675675675675676	0.00972447325769854
在一起	0.392405063291139	0.139240506329114	0.0886075949367089	0.316455696202532	0.0632911392405063	0.00792364487664326
沒辦法	0.3125	0.041666666666667	0.166666666666667	0.291666666666667	0.1875	0.00720331352422114
不見了	0.385964912280702	0.333333333333333	0.140350877192982	0.105263157894737	0.0350877192982456	0.00612281649558797
怎麼辦	0.229166666666667	0.208333333333333	0.25	0.3125	0	0.00594273365748244
要不要	0.363636363636364	0.272727272727273	0.0681818181818182	0.25	0.0454545454545455	0.00540248514316586
還可以	0.567567567567568	0.243243243243243	0.108108108108108	0.0540540540540541	0.027027027027027	0.00486223662884927
沒關係	0.342857142857143	0.142857142857143	0.228571428571429	0.171428571428571	0.114285714285714	0.00468215379074374
忍不住	0.419354838709677	0.0967741935483871	0.258064516129032	0.193548387096774	0.032258064516129	0.00432198811453269
好好笑	0.416666666666667	0.277777777777778	0.027777777777778	0.25	0.027777777777778	0.00432198811453269
接下來	0.236842105263158	0.184210526315789	0.157894736842105	0.236842105263158	0.184210526315789	0.00432198811453269

圖 35：方法二詞彙 Normalize 及權重值

資料來源：此圖自行整理

由公式(12)計算出每位測試者之 N、E、O、A、C 五大類的 Grade 值，而 Type 為心理測驗之實際人格分類，如下圖 36 所示。

ID	Type	N Grade	E Grade	O Grade	A Grade	C Grade
144...	O,A	0.390084089234789	0.156001560910366	0.411770536844985	0.496866136432717	0.264348449132519
100...	A	1.38733136112142	0.495994207357368	1.55828087022792	2.01869731360424	1.11199950718842
100...	A	4.28077901443144	1.4074613210447	5.11574202012081	6.6345357652032	3.70104607873165
151...	A	3.63244871780677	1.20096846170618	4.40485498032602	5.63229188834786	3.14636373859509
140...	C,A	3.71332170908025	1.22133370889265	4.44486885748494	5.75836081974293	3.21723826154333
100...	E,A	1.28405826180125	0.466744868428379	1.51477830330834	1.8871126189611	1.04845847766481
178...	E,A	2.41909863079212	0.810714803742316	2.96068117938569	3.74183692256327	2.08891823841307
103...	A	2.17582991472651	0.703015187109083	2.6383831000072	3.40791011960498	1.90666430776167
178...	A	1.00364380717974	0.329642334916252	1.19013753465704	1.56471989012822	0.866835723592354
571...	A	0.479346257814713	0.161578678911053	0.56464571539765	0.738130208086552	0.40962166815308
646...	N	0.143846175972894	0.0470687957987324	0.169627185843866	0.222827912322922	0.123635152463891
100...	A	1.29949560336647	0.42264937097529	1.57787171244457	2.03033865746908	1.13713970346654
100...	N	0.183892424052396	0.0587804784325637	0.223611600626374	0.289143532340666	0.161913941857563
100...	N	0.0476975356143519	0.0160184556154315	0.0572098649967836	0.0738111301533634	0.0408113658621009
100...	N	0.00196454005206031	0.00147340503904523	0.000368351259761308	0.00135062128579146	0.000245567506507539

圖 36：方法二詞彙 Grade 計算

資料來源：此圖自行整理

由圖 36 中的資料取得每位測試者之 N、E、O、A、C 五大類的 Grade 加總值中取最大者，為預測結果。

以圖 37 為例，N Grade 的值為此位測試者最大值，故預測結果屬 N(神經質型)，而與 Type(實際人格分類)N 比對是相同的，故預測結果正確。

結果 訊息

	ID	Type	N Grade	E Grade	O Grade	A Grade	C Grade
1	10...	N	0.00196454005206031	0.00147340503904523	0.000368351259761308	0.00135062128579146	0.000245567506507539

圖 37：方法二類別預測結果

資料來源：此圖自行整理

五、實驗結果與分析

5.1 方法一(關鍵字詞預測法)

方法一：關鍵字詞預測法，經統計後，其顯著人格分析(Method1)有 61%準確率，多重人格分析(Method2)有 70%準確率，如圖 38 所示。

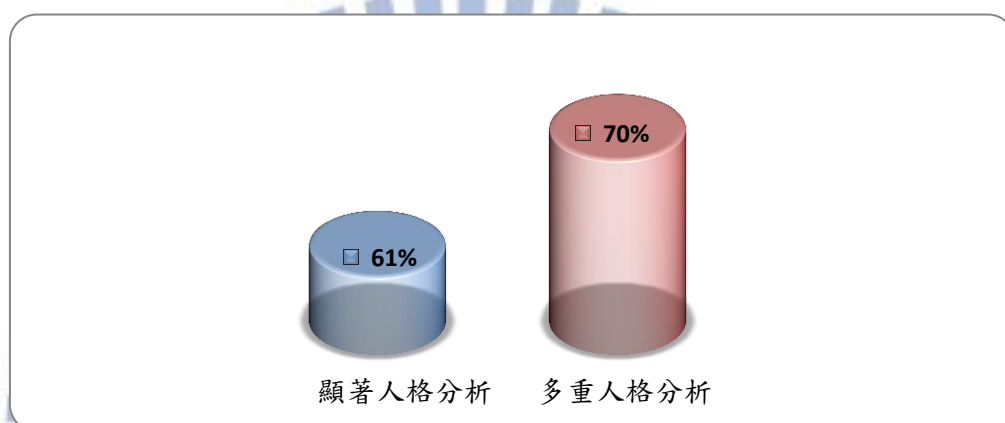


圖 38：方法一研究結果(1)

資料來源：此圖自行整理

由於朋友大多以 A(隨和型)的朋友居多，可以由顯著人格分析(Method1)及多重人格分析(Method2)的心理測驗結果可得知，故預測後的結果以 A(隨和型)所佔的比率最高，如圖 39 所示。此研究亦可證實物以類聚這個情況，也就如圖 30 所示，本人之五大人格分類屬 A(隨和型)，相對自己的朋友亦以類別為 A(隨和型)的居多，相同性格的人、志同道合的人會相聚成群。

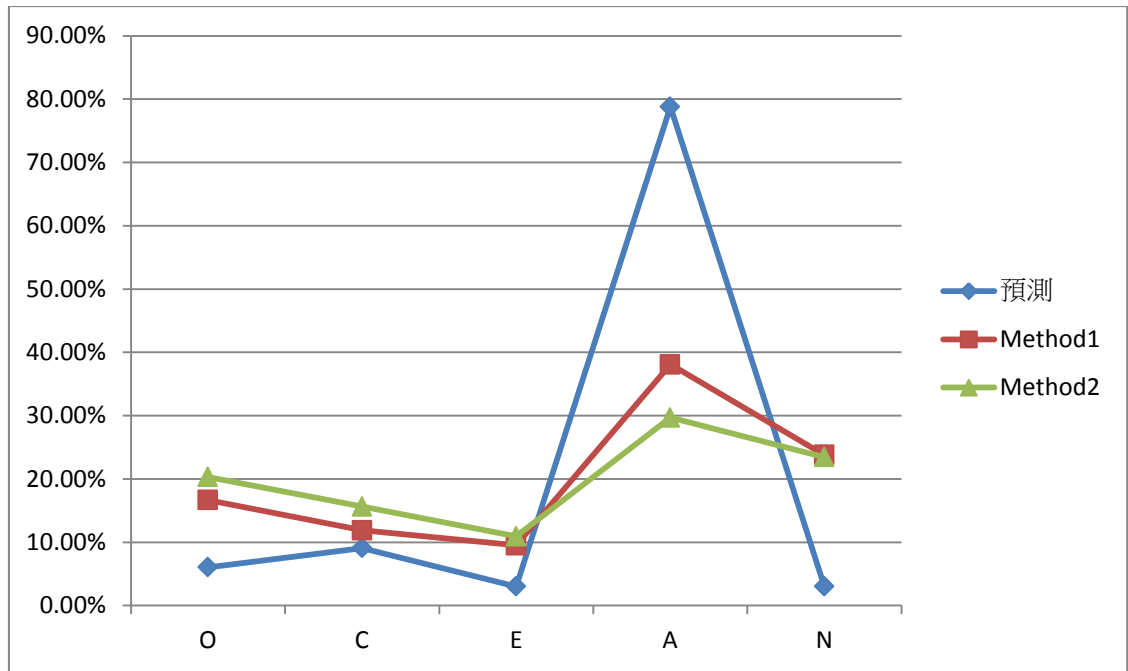


圖 39：方法一研究結果(2)

資料來源：此圖自行整理

5.2 方法二(機器學習預測法)

方法二之機器學習預測法，以機器學習的方式，將抽樣數的 80%為訓練資料集，透過學習將詞彙與已知之人格類別，加以訓練後，將每一詞彙歸屬於五大人格類別 {C、A、N、O、E} 之中，而五大類別所對應的名稱為，C：嚴謹型、A：隨和型、N：神經質型、O：開放型、E：外向型。對應後的詞彙及類別為用來預測 20%測試集資料的分類結果。

圖 40~43 舉例測試集中其中三個人的預測結果，加以說明此三個人實際人格與預測人格比對，以驗證比對後為預測正確或預測錯誤。

以圖 44 統計測試集所有抽樣者之人格分佈，可代表本研究方法二的預測分類結果分佈情況，以五大人格分類統計表顯示，可清楚顯示預測結果，而預測出的結果以 A(隨和型)的朋友居多，與方法一的結果相同。

本研究之方法二的預測結果，以顯著人格分析的預測為主，取人格分數之最高分者為顯著人格表現，預測準確率有 80%，高於方法一 61%的顯著人格預測準確率，方法二的預測結果如圖 45 所示。

方法二透過訓練學習的方式，所取得之五大人格詞彙表，這部份的資料可用於知識庫中詞彙資料的擴充，當詞彙量越大時，更可增加其分類的準確度，對預測五大人格類別有更大的幫助。

ID 為 000001 這個朋友的實際人格計算出來最高分者，同時有二個，屬 A(隨和型)及 O(開放型)二種人格表現，本研究預測值只會取最大值者為此預測分類結果，故預測後結果為 A(隨和型)，A(隨和型)於 A(隨和型)及 O(開放型)其中一類，故本研究將此預測歸屬為預測正確，ID(000001)預測結果如圖 40 所示。

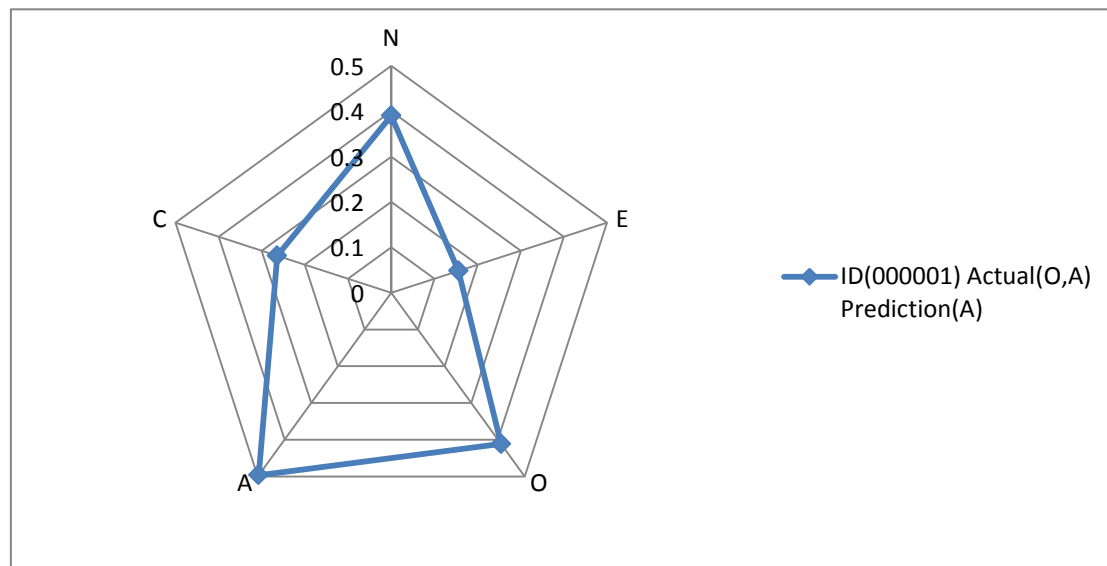


圖 40：方法二測試集個人預測結果(1)

資料來源：此圖自行整理

ID 為 000002 這個朋友的實際人格計算出來取最高分的值，屬 A(隨和型)，而預測後結果亦為 A(隨和型)，預測與實際一致，故歸屬為預測正確，ID(000002)預測結果如圖 41 所示。

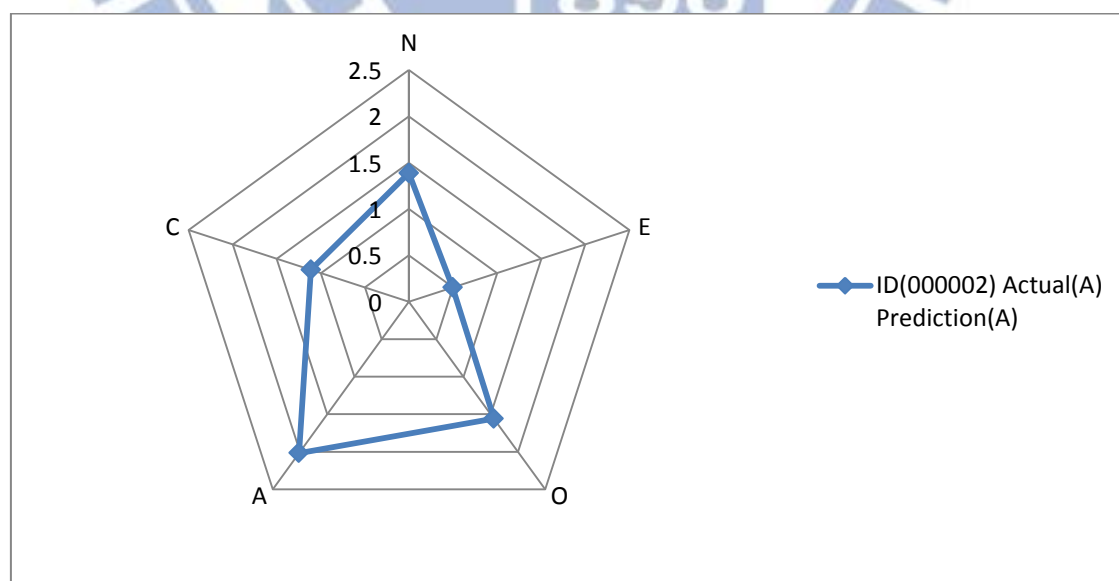


圖 41：方法二測試集個人預測結果(2)

資料來源：此圖自行整理

ID 為 000015 這個朋友的實際人格計算出來取最高分的值，屬 N(神經質型)，而預測後結果亦為 N(神經質型)，預測與實際一致，故歸屬為預測正確，ID(000015) 預測結果如圖 42 所示。

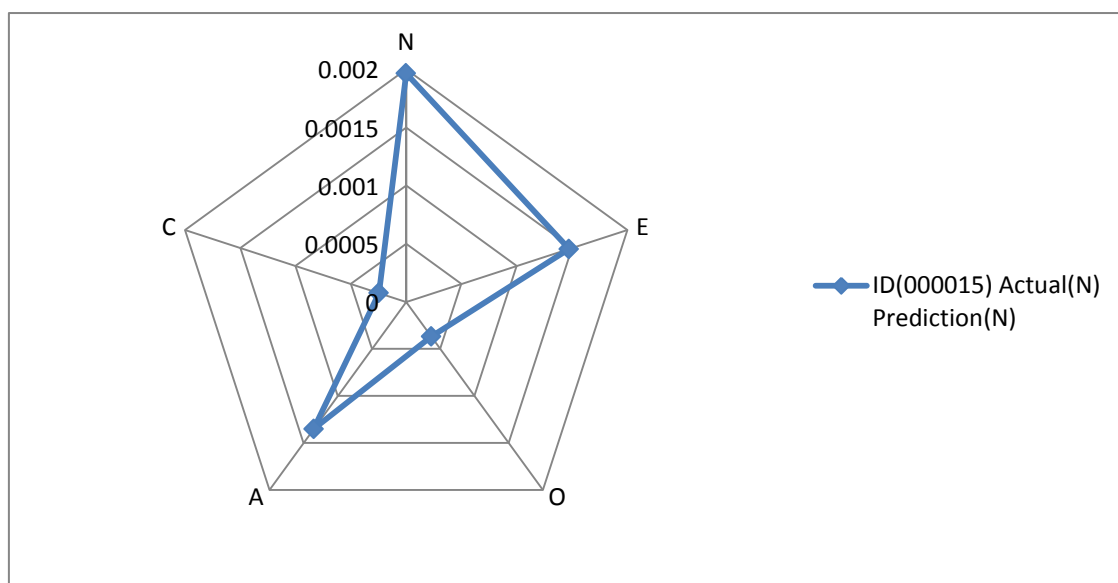


圖 42：方法二測試集個人預測結果(3)

資料來源：此圖自行整理

ID 為 000011 這個朋友的實際人格計算出來取最高分的值，屬 N(神經質型)，而預測後結果卻為 A(隨和型)，預測與實際不一致，故歸屬為預測錯誤，ID(000011) 預測結果如圖 43 所示。

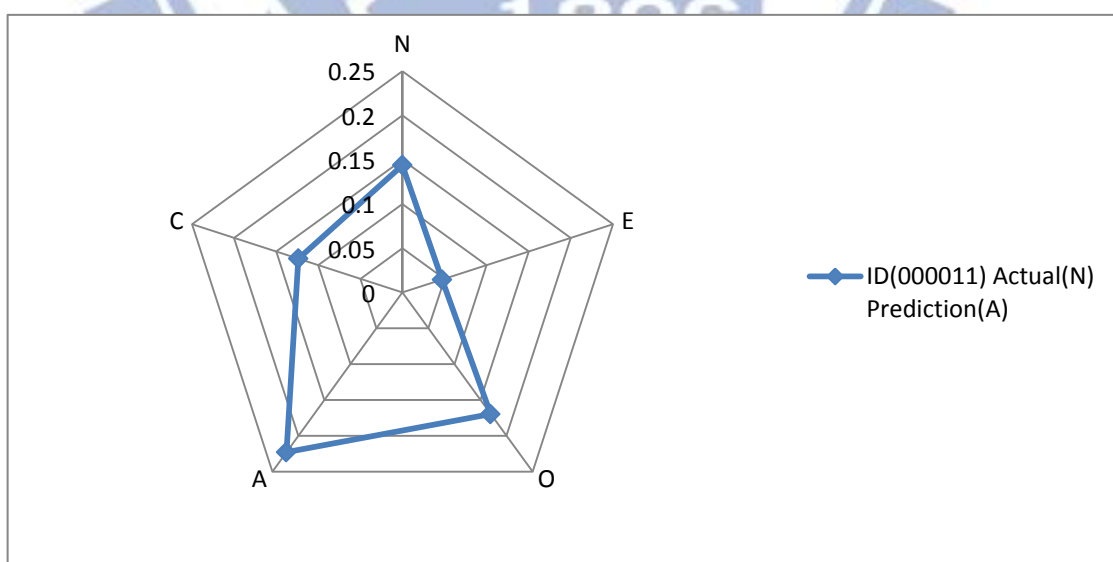


圖 43：方法二測試集個人預測結果(4)

資料來源：此圖自行整理

方法二於測試集中的所有分類資料加以統計其預測結果，統計出的結果以 A(隨和型)所佔的比例最高，圖 44 可顯示測試集中的抽樣者實際顯著人格，以隨和型的人數居多，而預測後的結果亦以 A(隨和型)居多。

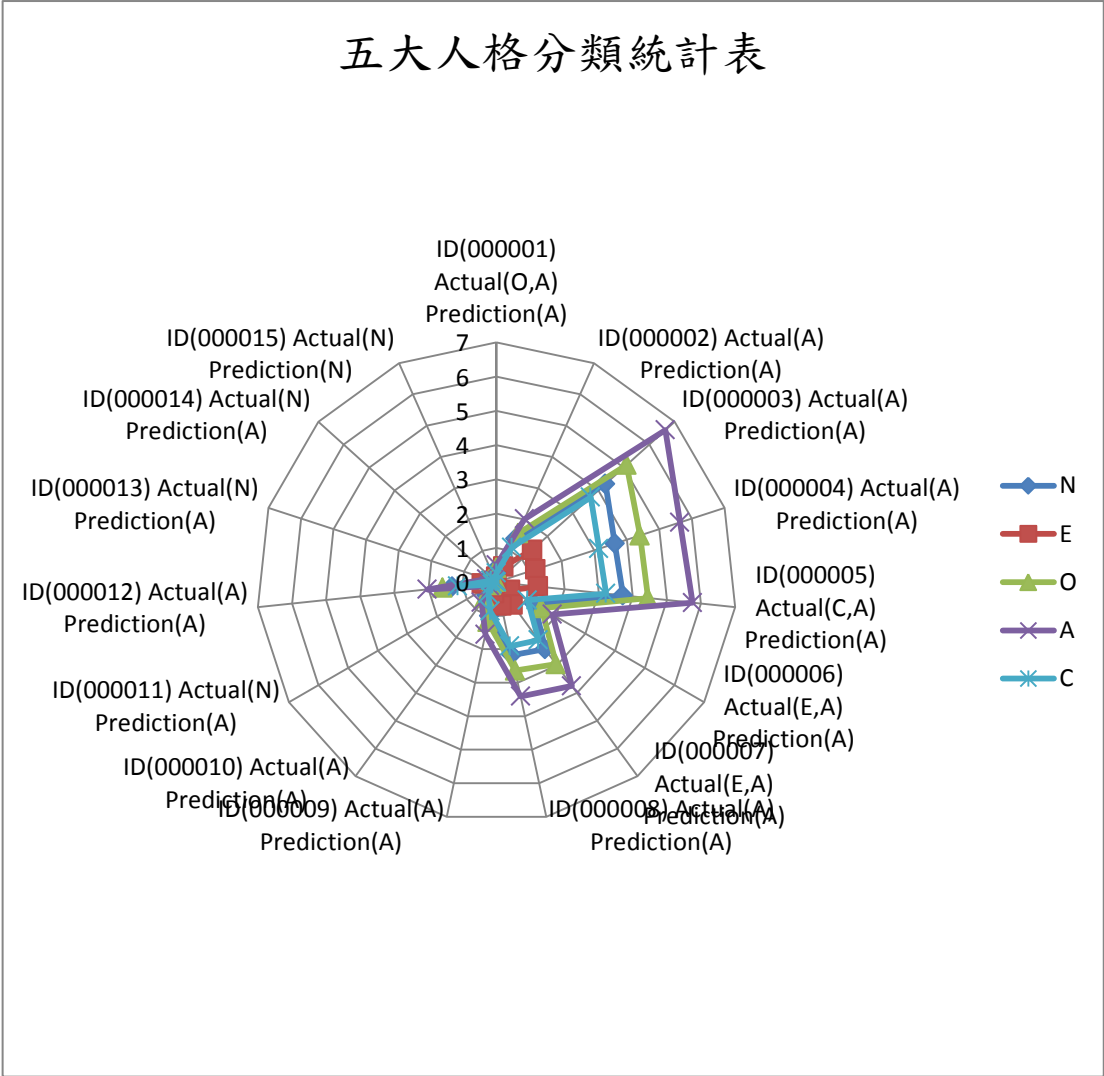


圖 44：方法二五大人格分類統計表
資料來源：此圖自行整理

方法二透過機器學習預測法實作後的結果，可達到 80%的預測準確率，此準確率計算以實際人格類別與預測人格類別完全符合時則為預測正確。另一則當實際人格有一個以上同分者，預測人格類別只要符合其中一個實際人格類別，則本研究將此結果歸屬為預測正確，反之則為預測錯誤。

方法二(機器學習預測法)預測結果以圖 45 顯示。

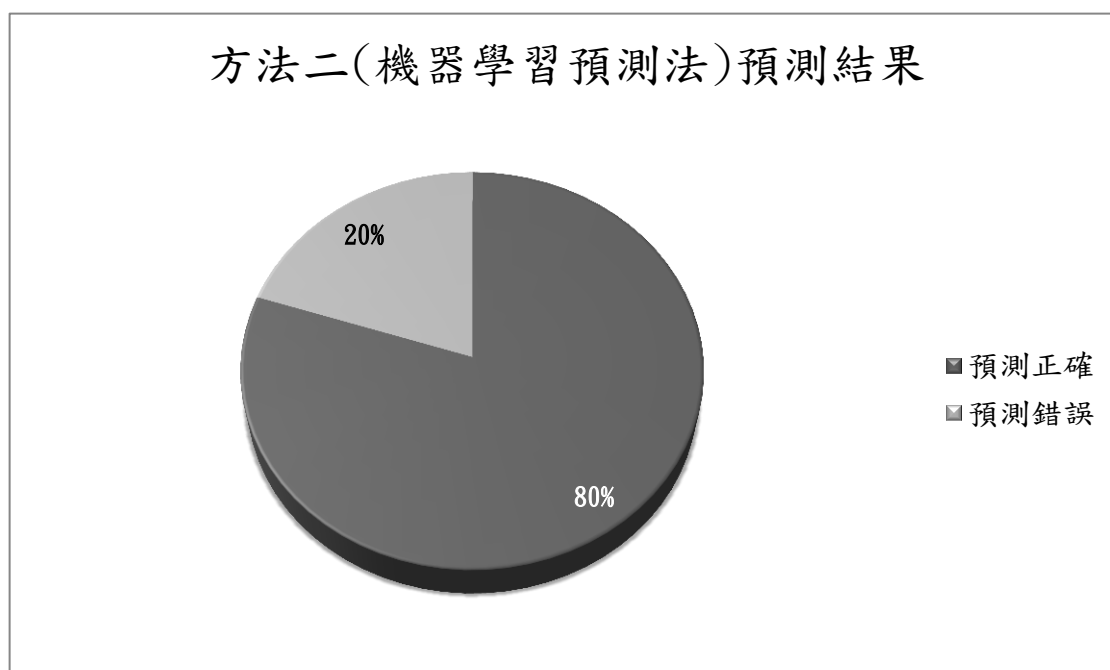


圖 45：方法二（機器學習預測法）預測結果

資料來源：此圖自行整理

5.3 方法二支援方法一

方法一(關鍵字詞預測法)可透過方法二(機器學習預測法)所訓練出的詞彙來擴充方法一的詞彙，此為自動擴充方法一的詞彙量，可使方法一的特徵詞彙資料庫中之詞彙增加，更可有效的解決手動擴充的問題，達到自動擴充的效果，當詞彙量多時亦可增進預測的準確性。

此自動擴充的方式為取方法二中之 N、E、O、A、C 五類人格之 Weight 與 stdev 的值較高者為重點詞彙，將以此為欲擴充方法一的詞彙。

取 Weight 的值較大者，表示此詞彙權重較大，也就是較常被引用且被引用的次數較多，而取 stdev（標準差）值較大者，表示此詞彙在各人格使用上，有較顯著差異，所以每一類別各取 $\text{Weight} \times \text{stdev}$ 值較高者的前 100 個詞彙，最後共取出 401 個較有影響力的詞彙，做為擴充方法一的詞彙。如圖 46 所示，例如取 N（神經質型）之重點詞彙。

	Word	N	E	O	A	C	Weight	stdev	diff
1	沒想到	0.44444...	0.38562091503...	0.0457516339869281	0.111111111111111	0.0130718954248366	0.0172879524581307	0.179356694763808	0.00310071001212418
2	看起來	0.42975...	0.16528925619...	0.15702479338843	0.214876033057851	0.0330578512396694	0.0167477039438142	0.129559788048036	0.00216982897325182
3	越來越	0.39375	0.3125	0.08125	0.15	0.0625	0.0163875382676031	0.130922209727762	0.00214549272199285
4	會不會	0.35087...	0.20175438596...	0.157894736842105	0.201754385964912	0.087719298245614	0.0136862956960202	0.0861973227827245	0.00117972204780966
5	在一起	0.39240...	0.13924050632...	0.0896075949367089	0.316455696202532	0.0632911392405063	0.00782364487664326	0.130668231904445	0.00103536866626969
6	還可以	0.56756...	0.24324324324...	0.108108108108108	0.0540540540540541	0.027027027027027	0.00486223662884927	0.198312825165883	0.000964243882492136
7	不見了	0.38596...	0.33333333333...	0.140350877192982	0.105263157894737	0.0350877192982456	0.00612281649558797	0.135712839339044	0.000830944811368176
8	沒辦法	0.3125	0.04166666666...	0.166666666666667	0.291666666666667	0.1875	0.00720331352422114	0.0973610120462327	0.000701321894804485
9	要不要	0.36363...	0.27272727272...	0.0681818181818182	0.25	0.0454545454545455	0.00540248514316586	0.123147428946764	0.000665302155303967
10	好好笑	0.41666...	0.27777777777...	0.027777777777778	0.25	0.027777777777778	0.00432198811453269	0.151535352188732	0.000654933991091223
11	莫名其妙	0.43243...	0.40540540540...	0.0540540540540541	0.0810810810810811	0.027027027027027	0.00360165676211057	0.179765288089902	0.000647452865441749
12	忍不住	0.41935...	0.09677419354...	0.258064516129032	0.193548387096774	0.032258064516129	0.00432198811453269	0.134404301006449	0.000580893791491948
13	好不容易	0.4375	0.3125	0.03125	0.1875	0.03125	0.00360165676211057	0.158853312839235	0.000572135108371098
14	買一送一	0.5	0.27272727272...	0	0.0454545454545455	0.181818181818182	0.00288132540968846	0.17860802458535	0.00051462783961203
15	不可能	0.4375	0.125	0.125	0.3125	0	0.00324149108589951	0.155120920574886	0.000502823081280019
16	不自覺	0.47826...	0.34782608695...	0.0869565217391304	0.0869565217391304	0	0.00270124257158293	0.181570547981053	0.000490466093952061

圖 46：方法二擴充詞彙截取

資料來源：此圖自行整理

依此類推共取出 401 個具影響力之特徵詞彙用於擴充方法一的詞彙資料庫，擴充後的詞彙內容如圖 47 所示，例如 0(開放型)的部份詞彙資料。

	type	desc	term
195	O	開放型	大刺刺
196	O	開放型	士別三日
197	O	開放型	千呼萬喚
198	O	開放型	千里迢迢
199	O	開放型	也就是說
200	O	開放型	上不了
201	O	開放型	人才濟濟
202	O	開放型	一探究竟
203	O	開放型	一哄而散
204	O	開放型	功成身退
205	O	開放型	比不過
206	O	開放型	比下去
207	O	開放型	引以為傲
208	O	開放型	生下來
209	O	開放型	民以食為天
210	O	開放型	多災多難
211	O	開放型	多才多藝
212	O	開放型	忙翻天
213	O	開放型	好不好看
214	O	開放型	狹路相逢
215	O	開放型	晉升為

圖 47：方法一擴充詞彙資料產生

資料來源：此圖自行整理

取原方法二的測試集人員資料，套用於方法一中，在尚未擴充方法二的詞彙前的預測結果記載於 Method1(Original)欄位中，擴充詞彙後的預測結果記載於 Method1(Expansion)中，Result_Method1(Original)為擴充前的預測結果，正確(True)或錯誤(False)，Result_Method1(Expansion)為擴充後的預測結果，正確(True)或錯誤(False)。如表 10 所示。

ID	Actual	Method1 (Original)	Method1 (Expansion)	Result_Method1 (Original)	Result_Method1 (Expansion)
000001	O, A	A	A	True	True
000002	E, A	C	C	False	False
000003	A	A	A	True	True
000004	N	C	N	False	True
000005	A	A	A	True	True
000006	A	A	A	True	True
000007	A	C	N	False	False
000008	A	A	A	True	True
000009	A	A	A	True	True
000010	C, A	A	A	True	True
000011	N	E	E	False	False
000012	A	C	N	False	False
000013	A	A	A	True	True
000014	E, A	A	A	True	True
000015	A	A	A	True	True

表 10：方法一擴充結果比較表

資料來源：本研究自行整理

由圖 48 可明顯看出，對於 ID:4 這位人員而言，擴充詞彙後的預測較未擴充詞彙前的預測來的準確，圖 48 中，0 表示預測錯誤，1 表示預測正確。

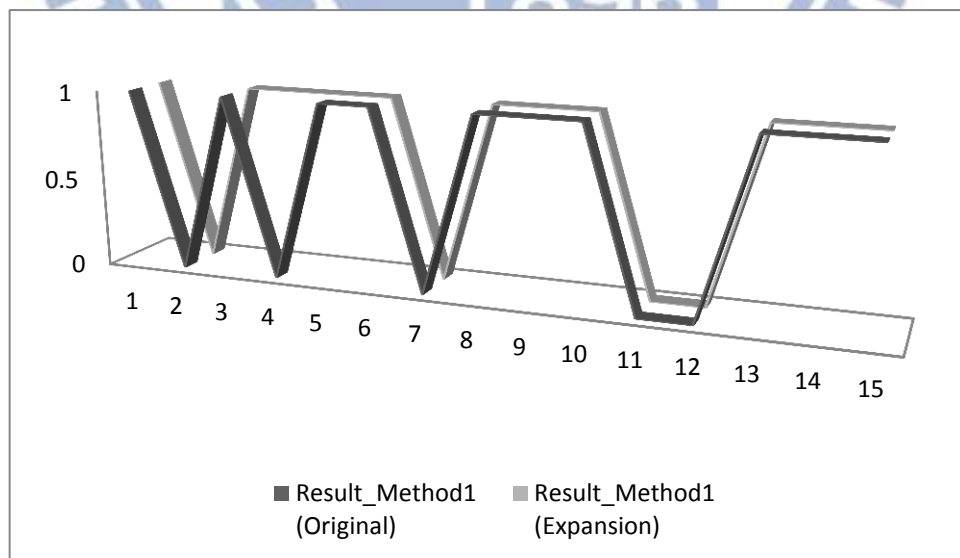


圖 48：方法一擴充結果比較圖

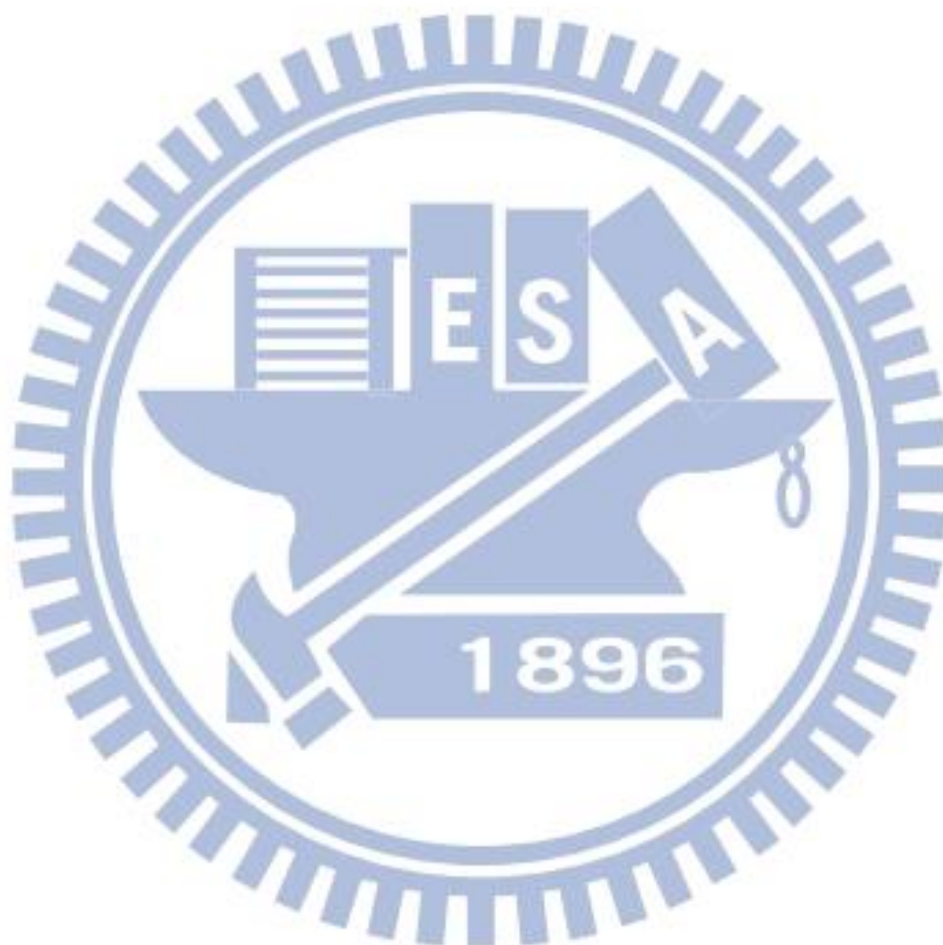
資料來源：此圖自行整理

方法二支援方法一的詞彙數，將原方法一詞彙資料庫由 2731 筆增至 3132 筆，擴充前方法一預測準確率為 66.67%，擴充後預測準確率為 73.33%，結果可證明方法二可有效的擴充方法一的詞彙，並且可有效的增進方法一的預測準確率。如表 11 所示。

Result_Method1	(Original)	(Expansion)
準確率(Precision)	66.67%	73.33%

表 11：方法一擴充前後準確率比對

資料來源：本研究自行整理



六、結論及建議

6.1 結論

本研究排除以往的人工標準化測驗，創新結合心理學、習慣領域、資料探勘、統計學、社群網路，運用文字探勘技術，進行五大人格分類預測。研究方法採用關鍵字詞預測法及機器學習預測法二種方法，於社群網路上日常行為的活動記載以預測人格特質，總結如下：

1. 排除人工制式化的心理測驗，結合社群行為以分析人格特質，為企業有效的選用對的人才，選到對的人才可替企業帶來價值的創造及效益的提昇。
2. 人類平常潛意識裡的文字習慣用語，證實可透過社群網站分析語句來找出用戶人格的主要面向。
3. 習慣領域中，人類的每一個行為、每一個動作是受長期習慣所養成，且不知不覺的控制我們的行為。本研究證實這些習慣行為會第一個反應於平常的行為當中，透過文字的表達可分析五大人格特質所在。
4. 方法一(關鍵字詞預測法)之研究結果準確率 61%，而方法二(機器學習預測法)之研究結果準確率 80%，顯示方法二預測效果較方法一顯著且較準確，如圖 49 所示。

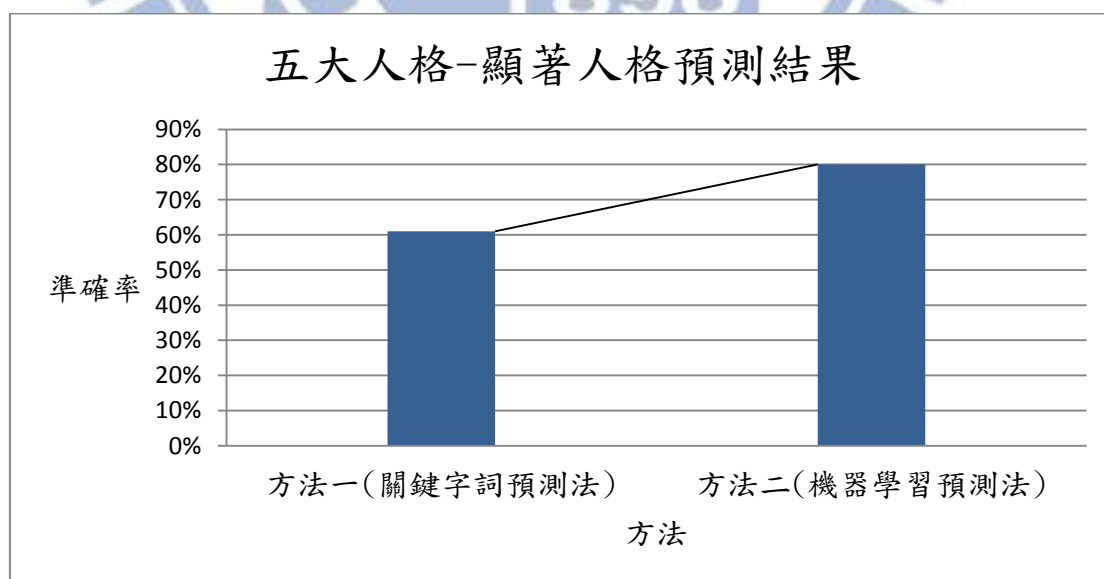


圖 49：二方法顯著人格預測比較

資料來源：此圖自行整理

5. 詞彙資料庫中之擴充詞彙數多寡，與研究結果的準確率成正比。
6. 方法二自動擴充方法一的方式分析研究成果，也就是方法一透過方法二的擴充詞彙，取具影響力之 401 個詞彙，由原特徵詞彙數為 2731 擴充至 3132。採用方法二的測試集人員的資料進行比較，其結果為方法一的準確率由 66.67% 增至 73.33%，可顯示方法二有效的擴充方法一的詞彙，使提高方法一的預測結果，如圖 51 所示。而方法一透過方法二的機器學習自動擴充詞彙，詞彙的有效增加使方法一的預測效果提升，就像齒狀圖 50 所示，其彼此環環相扣，因方法二滾動而帶動詞彙擴充，進而影響方法一的預測結果。

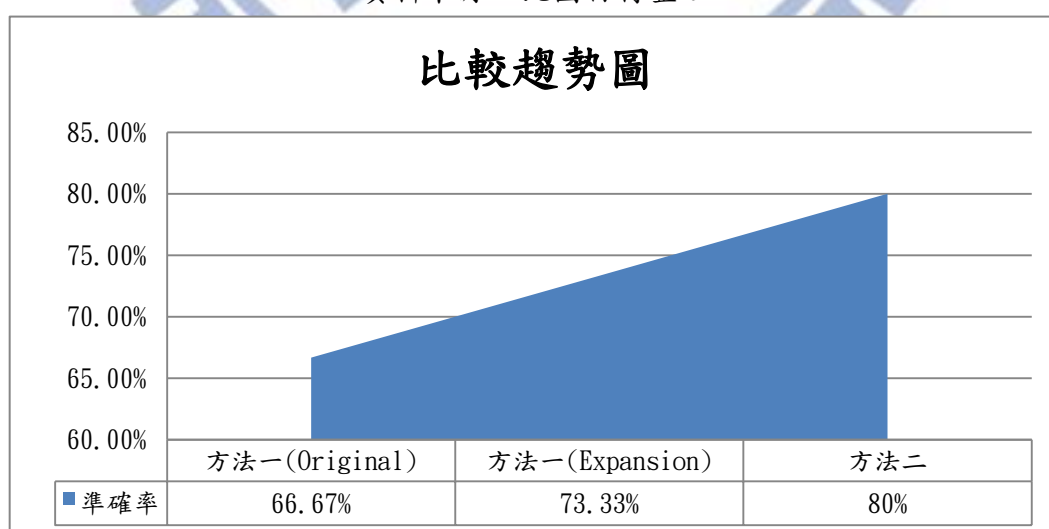


圖 51：擴充詞彙人格預測比較趨勢圖

資料來源：此圖自行整理

6.2 建議

1. 個資法的盛行，針對個人資料隱私是被重視的，而 Facebook 在的內容非經本人允許，並不合適給未來的雇主檢視，故此需經應徵者之同意及加入朋友才可截取資料及分析。
2. 若應徵者有心刻意刪除或透過權限設定隱藏顯示資料，反而可能也掩蓋本身實際的人格呈現，對雇主而言，可能會有所混淆應徵者的人格特質。所以有遇到類似此有心人的作法，若要防範可能需雇主與應徵者在面談時需清楚規範。
3. 本研究之二種方式透過中文貼文分析五大人格特質，未來亦可運用於其它平台之分析研究，可結合多平台的分析，更可加強研究的可信度。
4. 本研究的實際人格取得是透過 Google Doc 製作線上問卷的方式，以心理測驗回填問卷的方式作業，但由於回收率只有 73%，這中間隱含了有些 Facebook 朋友，他們對於 Facebook 資料被分析是不被允許的。即使本研究在問卷上已註明用途，也承諾不外流個人任何資料，亦有些朋友以為心理測試問卷為病毒或廣告，所以不加以回應問卷。這說明了網路使用者對於資安的考量是很敏感的，未來的研究者可針對此議題納入參考及研究。
5. 實際人格的取得，未來研究者可利用開發 APP 的方式取得其實際問卷結果。亦可增加取樣數，對研究分析及驗證更有幫助。
6. 企業可針對此二方法，利用完整的系統化程式，加以結合及開發成為一套即時線上五大人格分析系統，供招募即時使用。
7. 詞彙數的多寡影響預測的準確率，故建議亦可透過專家系統或其它知識庫擴充的方式同步進行。
8. 本研究未考慮貼文轉載的部份，故未來研究者可研究透過 Facebook 轉貼文的部份納入研究範圍內。

6.3 效益

本研究將為個人及企業帶來效益，而以企業所帶來的效益大於個人的效益，以下為個人及企業的效益

個人：

1. 讓人可更深層了解自己及朋友性格。
2. 可結交志同道合的朋友。
3. 可更緊密增進人與人之間的黏性。

企業：

1. 可減少不必要的時間填寫心理測驗試題。
2. 行銷工具，讓使用者 Lock In 某產品或活動。
3. 可搭配網路活動以提升銷售。
4. 可讓企業更了解用戶，以利推出更符合用戶需求的產品或活動。
5. 可提供企業用人之參考依據，可加快找到符合相關職務之工作人才。
6. 選用對的人才可為企業帶來價值的創造及效益的提昇，避免不必要資源浪費及成本的提升。
7. 低成本，無需耗費多餘的人力成本去檢視 Facebook 資料。
8. 節能節碳，達到無紙化的用人徵選。

參考資料

- [1] Robert. J. Sternberg 著，普通心理學，陳億貞譯，雙葉書廊，台北，1頁，民國九十二年。
- [2] 黃富順，高齡學習，五南出版社，台北，160頁，民國九十七年。
- [3] 商業周刊-科技報橘，
<http://www.businessweekly.com.tw/blog/article.php?id=1140>
- [4] 陸洛，人格理論，台北市，民國九十五年。
- [5] 尹相志，SQL Server 2005 Data Mining 資料採礦，精誠資訊，台北，民國九十六年。
- [6] 游伯龍，HD:習慣領域，時報文化，台北市，民國九十八年。
- [7] Pang-Ning Tan·Michael Steinbach·Vipin Kumar 著，資料探勘，施雅月，賴錦慧譯，培生，台北，民國九十七年。
- [8] Jiawei Han and Micheline Kamber 著，資料探勘概念與方法，王派洲譯，滄海書局，台北，民國九十七年。
- [9] Richard Wiseman, 59 SECONDS:Think A Little, Change A Lot, Random House, New York, 2009
- [10] 尹相志，SQL Server 2008 Data Mining 資料採礦，悅知文化，台北，民國九十八年。
- [11] 李昇暉，決策支援與企業智慧系統，華泰文化，台北，民國一百年。
- [12] 王柏鈞，「部落格的異想世界－使用者人格特質與使用動機行為之關聯性初探」。
淡江大學，碩士論文，民國九十四年。
- [13] 黃木榮，鄭雅方，利鑑桓，「運用資料探勘技術在線上自我揭露行為之研究」，產業資訊管理學術暨新興科技實務研討會，1-10頁，民國九十九年。
- [14] Briggs, Assessing the Five Factor Model of personality Description, 1992.
- [15] Allport, G..W., Personality: A psychological interpretation, New York, 1937.
- [16] Ting-Ya Kuo & Yi-Fong Wang & Siou-Syong Hsieh, "An Exploration on Relationships among Personality, Social Network Sites and Leisure Activities" Journal of Commercial Modernization, Vol.1.6, No.2, pp.29-42, 2011.
- [17] 傅珮雯，「Facebook網站上口碑行為之研究」，國立中山大學，碩士論文，民國一百年。
- [18] 謝鎮宇，「意見探勘在中文評鑑語料之應用」，國立交通大學，碩士論文，民國九十九年。
- [19] 呂奇傑、李天行、周宗穎等，「應用資料探勘技術於人才甄選之研究」，第12屆科際整合管理研討會，450-465頁，民國九十七年。
- [20] 孫瑛澤、陳建良、劉峻杰等，「中文短句之情緒分類」，自然語言與語音處理研

討會，184-198頁，暨南大學，民國九十九年。

- [21] 蘇郁涵，「Facebook 之使用者人格特質、自戀行為暨涉入程度對使用者忠誠度影響之研究」。銘傳大學，碩士論文，民國一百零一年。
- [22] 池進通、李鴻文、陳芬儀，「五大人格特質與工作績效關係之研究」，經營管理論叢 Vol. 4, No. 2, 1-9 頁，民國九十七年。
- [23] 張瑋倫、林憲廷，「群體中個體性格相似性之研究」，圖書資訊學研究，83-113 頁，民國一百零一年。
- [24] Randall Wald and Taghi Khoshgoftaar & Chris Sumner, “Machine Prediction of Personality from Facebook Profiles”,IEEE IRI,Las Vegas,Nevada,USA, August 2011.
- [25] 中央研究院中英雙語知識本體詞網，<http://bow.sinica.edu.tw/intro/>
- [26] 中央研究院中文斷詞系統，<http://ckipsvr.iis.sinica.edu.tw/>



自 傳

我，張曉珍，N年前父母說若我生出來是男的，就要叫做張曉峰，但生出來卻是個女的，只好取這個名字了。也因為是個女的，一度長輩可能要把我被送給別人，但父母堅持留下自己養，所以才有現在的我。

我家中有四個兄弟姐妹，在家排行老么，雖是老么但不會特別受寵，在我們家是一視同仁的，父母對我們的教育也採放任制，不會太多限制，是個開明的家庭。父母說，有心念他們會就努力工作賺錢供我們念書，不想念的就去外面自己找工作。這也讓我們學習到凡事都自己管控自己的行為，自己要有自己的想法，且有能力自己去解決問題。做對的事，不要讓父母操心，這我們兄弟姐妹四個倒是都有做到，且大家都很孝順，其實這樣做父母的就很滿足了。雖然父母已不在了，但與父母相處的時光是很寶貴的，且是最幸福的時刻！

出生於苗栗這個鄉下地方，小時候跟著左鄰右舍的小朋友一同到處玩耍，老是搞得全身髒兮兮，或是跌倒受傷，這都是常有的事。這段時間是最多回憶的，也可說是最無憂無慮的日子，現在的小孩已無此童年了，每天補習上課，已無什麼童年可言，比較起來，那時，真得比他們現在幸福多了。

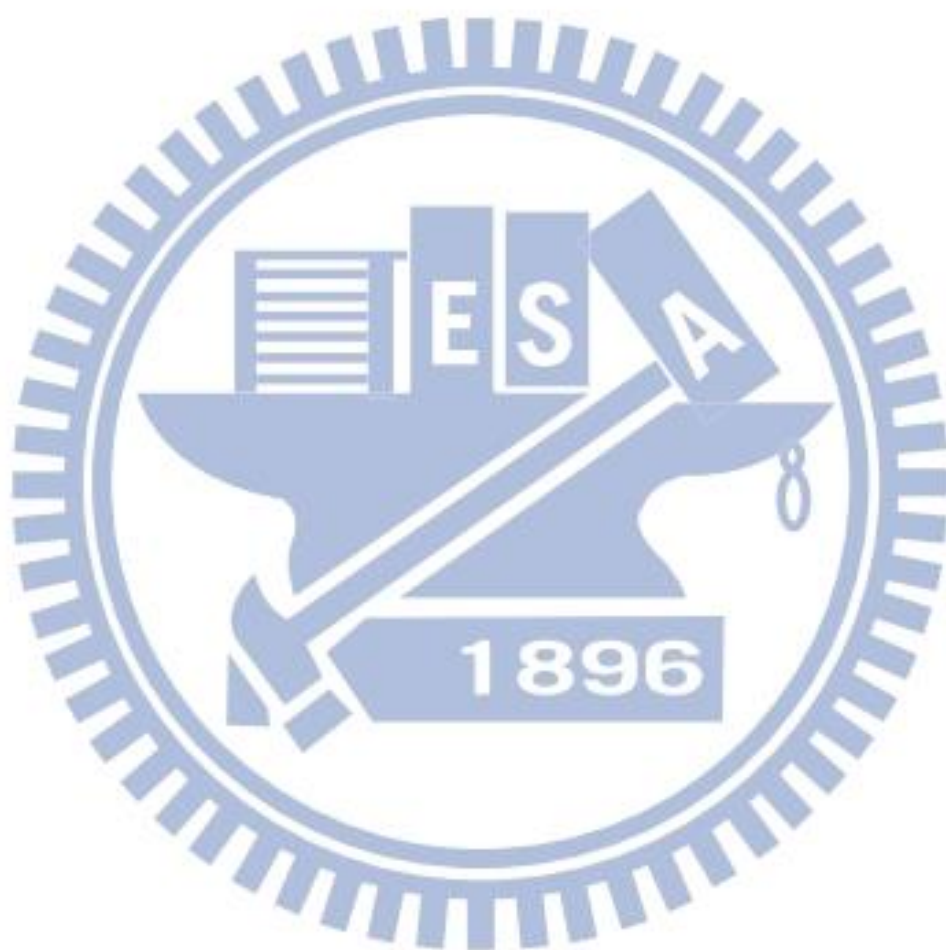
高中選擇技職學校就讀，當時就讀新竹高商會統科，在校期間學到很多技能，這是一般普通高中學生所無法學習到的，學校又很重視五育並重，那段求學過程真得多采多姿，收獲良多。接下來求學過程經歷二專、插大，資訊背景在二專開始展開，一路下來就跑不了這個領域了。

之後就開始工作生涯，忙忙碌碌。這幾年回想起來，沒什麼回憶且時間過的好快。隨著時代變遷，也不想安於現狀，要多替自己未來規劃，於是讓我想起繼續念書的念頭。考慮了一陣子後，決定報考交通大學資管所，第一年筆試通過，但口試沒過。決心隔年一定要一試就過，且直接錄取，要雪恥一下，很幸運的第二年如我所願考上了交大資管所，也讓我完成了當初大學畢業沒完成的夢想。

工作、上課、家庭，當這三種事情同時發生時，多工真得不好當，我腦袋的CPU快燒了，每天時間都不夠用。但發現原來我可以當個鐵人、當個打不死的蟑螂，這種生活很充實、但卻很累，這種酸甜苦辣回想起來真得是回憶無窮，重點是當初的辛苦是很值得的。

從工作再走回校園，讓自己換個心情學到不同東西，再運用到工作上，這才是發揮學以致用的效用。但人際關係更重要，不可忽略，不管在工作、在學校，遇到

的同學、同事、朋友，都是如此。有緣才會聚在一起，值得珍惜，多吸取別人的好處，改善自己的缺失，這才是邁向成功的累積，相信做起任何事來會更順遂，就如同老師所說的要站巨人的肩膀上，我想這樣才能做得更好，看得更遠。



附錄一

中研院平衡語料庫詞類標記集

精簡詞類	簡化標記	對應的CKIP詞類標記 ¹	
A	A	A	/*非謂形容詞*/
C	Caa	Caa	/*對等連接詞，如：和、跟*/
POST	Cab	Cab	/*連接詞，如：等等*/
POST	Cba	Cbab	/*連接詞，如：的話*/
C	Cbb	Cbaa, Cbba, Cbbb, Cbca, Cbcb	/*關聯連接詞*/
ADV	Da	<i>Daa</i>	/*數量副詞*/
ADV	Dfa	Dfa	/*動詞前程度副詞*/
ADV	Dfb	Dfb	/*動詞後程度副詞*/
ASP	Di	Di	/*時態標記*/
ADV	Dk	Dk	/*句副詞*/
ADV	D	<i>Dab</i> , Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	/*副詞*/
N	Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
N	Nb	Nba, Nbc	/*專有名稱*/
N	Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
N	Ncd	Ncda, Ncdb	/*位置詞*/
N	Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
DET	Neu	<i>Neu</i>	/*數詞定詞*/
DET	Nes	<i>Nes</i>	/*特指定詞*/
DET	Nep	<i>Nep</i>	/*指代定詞*/
DET	Neqa	<i>Neqa</i>	/*數量定詞*/
POST	Neqb	<i>Neqb</i>	/*後置數量定詞*/
M	Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	/*量詞*/
POST	Ng	Ng	/*後置詞*/
N	Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	/*代名詞*/
Nv	Nv	Nv1, Nv2, Nv3, Nv4	/*名物化動詞*/

¹ 斜體詞類，表示在技術報告#93-05中沒有定義，即後來增列的。

T	I	I	/*感嘆詞*/
P	P	P*	/*介詞*/
T	T	Ta, Tb, Tc, Td	/*語助詞*/
Vi	VA	VA11,12,13,VA3,VA4	/*動作不及物動詞*/
Vt	VAC	VA2	/*動作使動動詞*/
Vi	VB	VB11,12,VB2	/*動作類及物動詞*/
Vt	VC	VC2, VC31,32,33	/*動作及物動詞*/
Vt	VCL	VC1	/*動作接地方賓語動詞*/
Vt	VD	VD1, VD2	/*雙賓動詞*/
Vt	VE	VE11, VE12, VE2	/*動作句賓動詞*/
Vt	VF	VF1, VF2	/*動作謂賓動詞*/
Vt	VG	VG1, VG2	/*分類動詞*/
Vi	VH	VH11,12,13,14,15,17,VH21	/*狀態不及物動詞*/
Vt	VHC	VH16, VH22	/*狀態使動動詞*/
Vi	VI	VI1,2,3	/*狀態類及物動詞*/
Vt	VJ	VJ1,2,3	/*狀態及物動詞*/
Vt	VK	VK1,2	/*狀態句賓動詞*/
Vt	VL	VL1,2,3,4	/*狀態謂賓動詞*/
Vt	V_2	V_2	/*有*/
T	DE	/*的，之，得，地*/	
Vt	SHI	/*是*/	
FW	FW	/*外文標記*/	
COLONCATEGORY			/* 冒號 */
COMMACATEGORY			/* 逗號 */
DASHCATEGORY			/* 破折號 */
ETCCATEGORY			/* 刪節號 */
EXCLANATIONCATEGORY			/* 驚嘆號 */
PARENTHESISCATEGORY			/* 括弧 */
PAUSECATEGORY			/* 頓號 */
PERIODCATEGORY			/* 句號 */
QUESTIONCATEGORY			/* 問號 */
SEMICOLONCATEGORY			/* 分號 */
SPCHANGECATEGORY			/* 雙直線 */

表 12：中研院平衡語料庫詞類標記集
資料來源：中研院中文斷詞系統—詞類標記

附錄二

心理測驗格式

＊ ＊ 以下 10 題測驗問題：請衡量你目前的概況，點選較符合的程度 ＊ ＊
以「我覺得我是...」回答以下問題

註：測驗結果僅供學術分析統計使用，不外流個人測驗結果供其它用途，
亦不提供個人測驗結果於學術文件中，僅以全數統計結果顯示。

*必填

1. 派對上的靈魂人物 *

- ☐ 極不同意 ☐ 不同意 ☐ 有點不同意
☐ 沒意見 ☐ 有點同意 ☐ 同意 ☐ 非常同意

2. 對他人不太關心 *

- ☐ 極不同意 ☐ 不同意 ☐ 有點不同意
☐ 沒意見 ☐ 有點同意 ☐ 同意 ☐ 非常同意

3. 總是做好準備 *

- ☐ 極不同意 ☐ 不同意 ☐ 有點不同意
☐ 沒意見 ☐ 有點同意 ☐ 同意 ☐ 非常同意

4. 容易感到壓力大 *

- ☐ 極不同意 ☐ 不同意 ☐ 有點不同意
☐ 沒意見 ☐ 有點同意 ☐ 同意 ☐ 非常同意

5. 有絕佳點子 *

- ☐ 極不同意 ☐ 不同意 ☐ 有點不同意
☐ 沒意見 ☐ 有點同意 ☐ 同意 ☐ 非常同意

6. 話不多 *

☐ 極不同意 ☐ 不同意 ☐ 有點不同意

☐ 沒意見 ☐ 有點同意 ☐ 同意 ☐ 非常同意

7. 對人感興趣 *

☐ 極不同意 ☐ 不同意 ☐ 有點不同意

☐ 沒意見 ☐ 有點同意 ☐ 同意 ☐ 非常同意

8. 常忘了把東西放回原位 *

☐ 極不同意 ☐ 不同意 ☐ 有點不同意

☐ 沒意見 ☐ 有點同意 ☐ 同意 ☐ 非常同意

9. 大多時間都很放鬆 *

☐ 極不同意 ☐ 不同意 ☐ 有點不同意

☐ 沒意見 ☐ 有點同意 ☐ 同意 ☐ 非常同意

10. 難以理解抽象概念 *

☐ 極不同意 ☐ 不同意 ☐ 有點不同意

☐ 沒意見 ☐ 有點同意 ☐ 同意 ☐ 非常同意

資料來源：(Richard Wiseman, 2009) [9]

附錄三

心理測驗衡量評分表

	極 不 同 意	不 同 意	有 點 不 同 意	沒 意 見	有 點 同 意	同 意	非 常 同 意
1. 派對上的靈魂人物	1	2	3	4	5	6	7
2. 對他人不太關心	7	6	5	4	3	2	1
3. 總是做好準備	1	2	3	4	5	6	7
4. 容易感到壓力大	1	2	3	4	5	6	7
5. 有絕佳的點子	1	2	3	4	5	6	7
6. 話不多	7	6	5	4	3	2	1
7. 對人感興趣	1	2	3	4	5	6	7
8. 常忘了把東西歸回原位	7	6	5	4	3	2	1
9. 大多時候都很放鬆	7	6	5	4	3	2	1
10. 難以理解抽象概念	7	6	5	4	3	2	1

表 13：心理測驗衡量評分表

資料來源：(Richard Wiseman, 2009) [9]

得分計算

以第 2 項敘述為例「對他人不太關心」與第 7 項敘述「對人感興趣」。把那兩個數字加總的分數就是「隨和型」的得分，如果分數在 10 分以下為低分者，而如果分數在 10 分以上為高分者，本研究取高分者為抽樣者之實際人格特質。

開放型

第 5 與第 10 項敘述的總分

☐ 低分（10 分與以下） ☐ 高分（10 分以上）

嚴謹型

第 3 與第 8 項敘述的總分

☐ 低分（11 分與以下） ☐ 高分（11 分以上）

外向型

第 1 與第 6 項敘述的總分

☐ 低分（9 分與以下） ☐ 高分（9 分以上）

隨和型

第 2 與第 7 項敘述的總分

☐ 低分（10 分與以下） ☐ 高分（10 分以上）

神經質型

第 4 與第 9 項敘述的總分

☐ 低分（9 分與以下） ☐ 高分（9 分以上）

資料來源：(Richard Wiseman, 2009) [9]

