國立臺灣大學管理學院資訊管理學系

碩士論文

Department of Information Management
College of Management
National Taiwan University

Master Thesis

應用倉頡編碼特徵於中文人名性別預測之研究

Predicting Genders of Chinese Names Using Sub-Character Features: An Experiment Using CangJie Codes

魏取向

Chu-Hsiang Wei

指導教授：盧信銘 博士

Advisor: Hsin-Min Lu, Ph.D.

中華民國 101 年 6 月

June 2012

致謝

這份論文的完成，對我幫助最大的絕對是我現在正在用的這台電腦—FUNclab，日夜跑資料，還要忍受我程式寫不好長時滿載、重跑之苦。沒有他，沒有今天的這份論文。不過之所以有 FUNclab 這個良伴，則一定要感謝我的指導教授盧信銘博士！當然，對盧老師的感謝絕對不僅止於提供設備，老師給我研究方向、研究建議，在論文的架構上、內容上，無不看到老師的影子，真的非常感謝老師花了很多時間協助我的研究。除了盧老師，也要感謝我的口試委員們：李瑞庭老師、魏志平老師、陳建錦老師，用心審視、無私指導，使我的研究能更有貢獻。所以說：「論文的肉體是 FUNclab 給的；而論文的精神是老師給的」一點也不誇張。

那我呢？我是肉體與精神連接的橋樑。這座橋梁也是在家人、朋友及許多同學的協助下才得以順利連接。我的家庭並不穩定，但是父親把生活大計一肩扛下，還花了許多寶貴的時間照顧年幼的妹妹，讓我能專心於學業，發展成就。女友妍皓，雖然健忘又嗜睡，卻非常講義氣，給我許多精神上的鼓勵，在我拚論文的最後一個晚上，陪我住實驗室，隔天正是基隆路大淹水難忘的日子…。資種蕙盈協助我英文摘要的翻譯，也很令人感動。另外還有實驗室的學長宇泰、學弟妹（崇瑋、如軒、久悌）及伙伴振和，在實驗上也給我許多寶貴的建議。最後，是幫我填寫問卷的長輩、同學及朋友們，我知道猜性別的問卷非常難填又枯燥，真的非常感謝您們還耐心地幫我填完。

我的學習歷程，是一段漫長又充實的旅途。時至今日，以浸淫了八年資管。這段日子裡，經歷了人生最大的成長，也一步步地朝理想邁進。承蒙輔大、政大與台大的照顧，讓我能以不同的角度看待我的人生。謝謝這裡的人們、環境和氣氛，今日取之於汝，他日必當使您們以我為榮！

中文摘要

日常生活中，對於素昧平生的人們，第一印象往往來自他的名字，我們常試著從名字中推敲他的性別、與其他人的關係（如是否與認識的人是兄弟）甚至樣貌。一般來說，性別是最顯而易見也最無爭議的。我們甚至可以推論，中文人名中本身就蘊含著性別資訊，而這些資訊往往能提供我們重要的人際線索。

本研究以倉頡碼對中文人名進行編碼，並配合性別資料藉由支援向量機學習中文字的性別特徵，進而達到以中文人名預測性別。在本研究中，我們比較了 K-最鄰近法與支援向量機的結果，並且對倉頡編碼採用不同的組合模式，企圖找出預測中文人名性別最精確的方法。

由於中文人名中存在著兩性皆可使用的名稱，所以性別預測難以達到 100%的準確率。在本實驗中發現以支援向量機搭配倉頡四連詞（4-grams）的準確率最高，達到最高可能預測結果的 93.59%。另外我們透過問卷比較人類判斷性別與系統判斷性別的差異，在統計檢定下為不顯著，代表系統處理中文人名的性別判斷與人類判斷無異。此外我們以模型對其他不同的資料集作測試，如臉書的好友名稱、英文譯名等，一樣展現出超過 85%的準確率。在本實驗的最後，我們將模型套用在台灣商家與台灣個股的名稱中，檢視不同類型的商店或類股是否會有不同的性別比例，從實驗結果中也發現的確存在這樣的差異。

本研究從中文人名的性別預測延伸到商家名稱等非人名的中文字，而發現以倉頡碼拆解中文字的確可以達到以字型表示文字某些特性，進而增加中文自然語言處理的可能性。除了利用本實驗的結果建立自動化大量人名性別判定的系統外，也可以在文件探勘時使用性別屬性而提供文章不同的特徵，可能可以提升文件分

類、分群或觀點分析的準確率。另外最重要的是，本實驗代表著可以以倉頡碼描述中文文字性別傾向，因而開啟後續研究以倉頡碼描述中文其他屬性的大門。

關鍵詞：文件探勘、中文人名、性別預測、支援向量機、中文字子結構、倉頡編碼

Abstract

In daily life, when we meet people we don't know, our first impressions usually come from their names: we often try to guess their gender, relationship with others (e.g. whether he is a brother of someone we know), or even appearance. Generally speaking, the gender characteristic in the name is the most obvious. We can even infer that a Chinese name contains gender information, and such information usually provides us with important clues concerning interpersonal relationships.

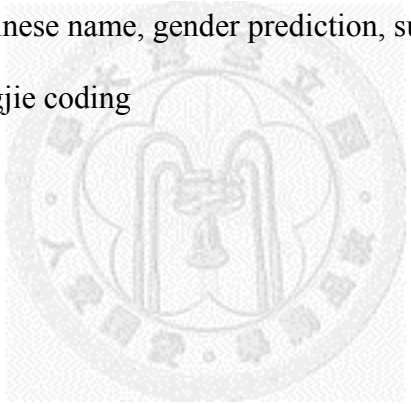
This paper uses CangJie code to represent Chinese names, and uses SVM (support vector machine) to learn the gender characteristics. In this paper, we compared the results of K-NN and adopted different combination modes to the CangJie coding in the SVM to find out the best method to predict of gender of a person through their Chinese name.

Because some Chinese names can be used in both genders, it is difficult to achieve the 100% accuracy when predicting the genders. We found that the highest accuracy of gender prediction is about 93.59% (by SVM with Cangjie 4-grams). On the other hand, we compare the gender prediction accuracy by humans and the systems through a questionnaire, and found that there is no significant statistical difference, which means there is no difference in the prediction of the gender of Chinese names between humans and our system. In addition, we applied the model to different data sets, such as Facebook friends' names, English names (translated in Chinese), and the accuracy also exceeds 85%. Finally, we applied the model to local shop names and stock names in Taiwan, finding the shop type or sector whether can have the different gender proportion, from the experimental result also found there indeed has such difference.

We found that the prediction of the gender of Chinese name can be extended to the

name of shops and the non-name Chinese characters, and found that the Cangjie code could possibly express the structure of the Chinese character, thus increasing the potential of Chinese natural language processing. The results of the experiment not only institutes the framework for a massive automatic name-sex prediction system, but can also be applied to text mining by provide more features of the articles and increase the accuracy of document classification, clustering, or viewpoint analysis. Moreover, the most importantly, Cangjie code can describe the gender characteristic of a Chinese character, thus opening the gates for future research on using Cangjie code to extract more attributes from Chinese characters.

Keywords: text mining, Chinese name, gender prediction, support vector machine, Chinese sub-character, Cangjie coding



目錄

圖目錄

表目錄

第一章　緒論

1.1 研究背景與動機

在歐洲語系的許多語言中，常在詞彙中帶有「性別」的觀念，例如法文、德文擁有「陽性」與「陰性」的詞。因此除了詞彙本身代表的意義，還蘊含著更多的資訊。但是這樣的特性在中文中是否存在呢？除了我們熟知的「女」為部首的字象徵著女性（陰性）外（朱寶安，2004；楊嘉敏，2006），是否具有其他的文字特徵暗示著性別？

若我們能有效掌握這些資訊，便能在中文的文件探勘中提供更多的分析線索，例如了解商標的命名是否為了取得特定性別族群的認同，而使用某種傾向的名稱；抑或公司行號名所代表的性別傾向是否會隨著產業性質而有不同的分佈…等等。這些研究除了可以擴展中文文件探勘領域的研究範疇，發展學術與商業價值外，更可以與語言學、社會學等其他學術領域整合，擴展人類對世界的認知。

然而，中文系統與歐洲語系不同，除了字元是由筆劃部件以二維方式組合；有別於歐洲語系是字母以一維橫向組合外，對性別的傾向也沒有歐洲語系來的嚴謹。本研究為了找到性別傾向的線索，我們透過中文人名男性與女性的差異作為出發點。我們觀察到中文人名具有意義性（Rossi, 1965; Adb-el-Jawad, 1986, Sung, 1981; and Watson, 1986），多蘊含對受名者的期望且男女有別，故假設中文人名在男性和女性的特徵差異可以一般化到中文的其他字彙。因此將問題以「將名字的性別自動分類」為開端，使用機器學習的方式找出中文人名的性別特徵，並試圖將學習的模型套用在其他中文字上。

值得注意的是，歐洲語系的相關研究（Bloomfield,1933; Cassidy, Kelly and Sharoni, 1999; Kilarski, 2007），多是以音節作為表示詞彙特徵的工具，但因為中文在造字時常透過字型傳達意義（許慎，121），因此我們認為若能妥善分析字型或筆劃元素（稱為字根或部件），找出部件中的規律，應可提供中文字性別表示更有效的線索。

總結而言，本研究有別於傳統中文文字探勘的斷詞研究，我們將焦點放在以部件拆解中文字並尋找性別屬性的意義規則，期望建立一套有效分類中文部件性別傾向的模型，對中文字彙提供從性別角度的分析線索。

1.2 研究目的

基於上述背景與動機，本研究中我們要找出有效表示中文部件筆劃的方法，將中文字轉化為序列化的特徵，並將此特徵套用機器學習的方式，找出蘊含在中文字中的性別傾向。我們希望透過實驗從中文人名建立中文性別預測的模型，並以此一模型套用生活中的名稱，了解中文詞彙的性別傾向。具體而言，本研究期望達到以下目的：

1. 使用機器學習找出中文名的命名規則，並建立適當的模型以區分人名的性別。
2. 發掘中文文字中部件的意義價值，尋找蘊含在筆劃形狀中的性別資訊。
3. 以實驗結果分析不同類型的名稱，了解真實世界中不同產業或店家在命名時性別傾向的考量。

1.3 研究架構

本研究目前可將架構分為七個部份，分述如下：

一、確定範圍：根據研究背景與動機，確認研究目的，以確定研究範圍。

二、相關研究探討：根據研究範圍蒐集與整理關於人名性別分析、人名使用研究、中文文字的部件組成、分類與預測工具等文獻資料，了解相關研究使用的方法。

三、性別預測系統設計：根據文獻的探討，設計基準線系統與主要的預測系統，並詳細討論中文特徵表達在時作上的細節。

四、資料處理與實驗結果：介紹如何蒐集實驗資料與清洗、轉換的步驟，並比較與呈現不同預測系統在中文人名判斷上的運作效能。

五、系統與人工判斷性別的比較：以問卷取得性別傾向明顯的人名、性別傾向模糊的人名、隨機挑選的人名與電腦依照性別規則產生的人名等四種狀況下人工判斷的結果，並以系統對同樣的資料進行預測，以評比人工判斷與系統判斷的異同。

六、真實名稱的性別傾向判別：我們使用臉書資料、英文譯名等接近但不同於人名的資料計算本系統正確性；另外針對網路服飾店、台灣本地商家與台灣個股名稱我們利用系統分辨在不同類型產業中的性別比例，以解釋性別比例在實際生活中的意義。

七、結論與建議：根據實驗結果確定本研究的貢獻性，並提出改進方式與未來研究方向供未來研究者參考。

研究架構如圖 1-1 所示：

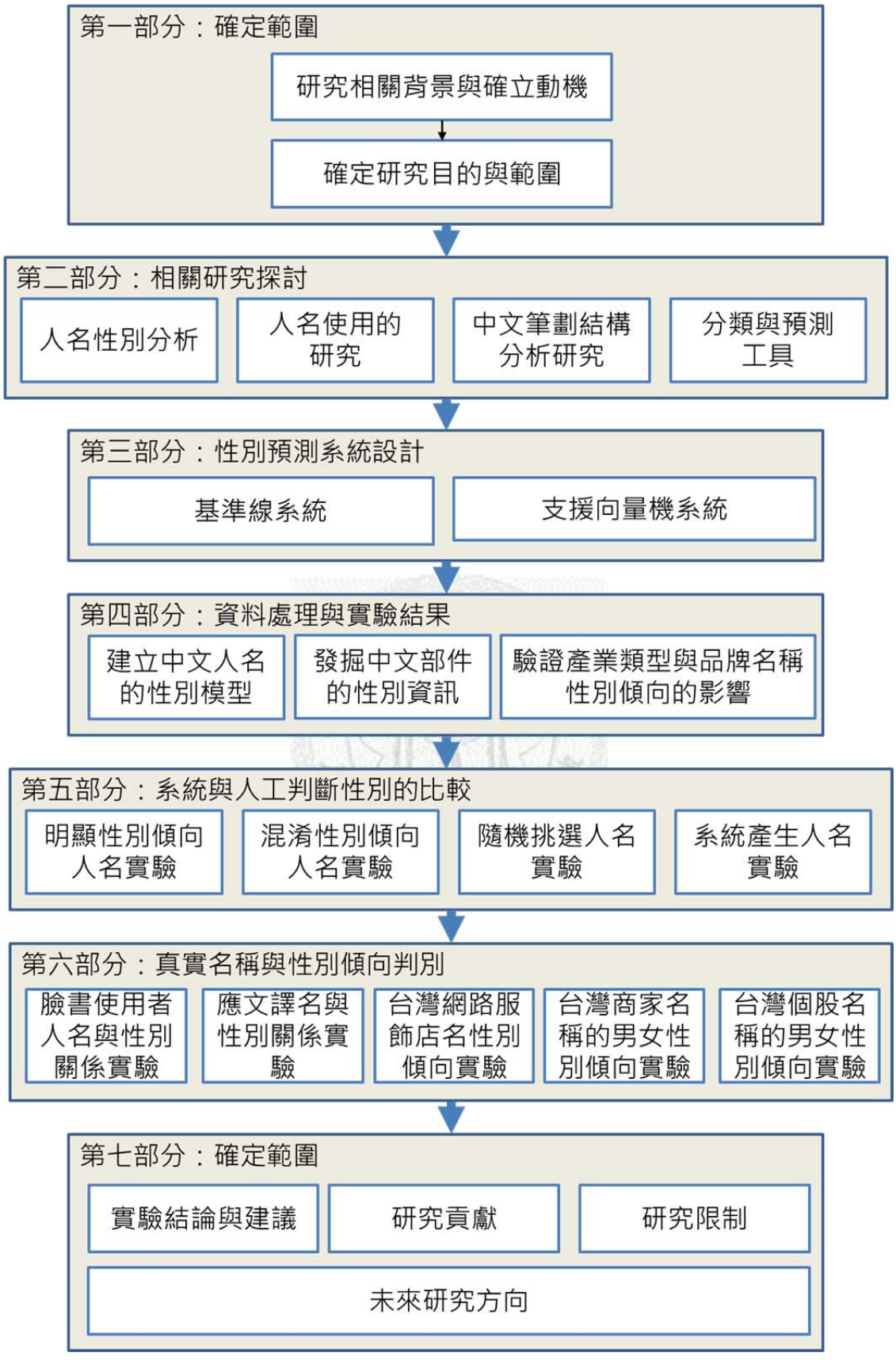

第一部分：確定範圍
研究相關背景與確立動機
確定研究目的與範圍
第二部分：相關研究探討
人名性別分析
人名使用的研究
中文筆劃結構分析研究
分類與預測工具
第三部分：性別預測系統設計
基準線系統
支援向量機系統
第四部分：資料處理與實驗結果
建立中文人名的性別模型
發掘中文部件的性別資訊
驗證產業類型與品牌名稱性別傾向的影響
第五部分：系統與人工判斷性別的比較
明顯性別傾向人名實驗
混淆性別傾向人名實驗
隨機挑選人名實驗
系統產生人名實驗
第六部分：真實名稱與性別傾向判別
臉書使用者人名與性別關係實驗
應文譯名與性別關係實驗
台灣網路服飾店名性別傾向實驗
台灣商家名稱的男女性別傾向實驗
台灣個股名稱的男女性別傾向實驗
第七部分：確定範圍
實驗結論與建議
研究貢獻
研究限制
未來研究方向

圖 1-1：研究架構圖

第二章　文獻探討

本研究企圖從中文人名裡找出規則，並使用系統識別這些名字分屬於男性或女性。因此本研究所涉及的範圍包括：人名的組成研究、人名中所隱含的特徵、特徵表示的方式與分類系統。以下各節介紹依序如下：2.1、人名最簡單的分類機制；2.2、人名使用的相關研究；2.3、中文的筆劃結構與型碼輸入法；2.4、分類與預測工具。

2.1 人名最簡單的分類機制

對於人名的分類，最直覺可見的方式是透過統計方法，計算出每個名字出現在男性或女性中的機率值，直接對名字加以分類（Gallagher and Chen, 2008）。例如從一組大量的嬰兒名字資料庫中計算英文名「Jack」為男生或女生的機率，計算結果發現為男性的機率大於女性，所以將該名字歸類為「男性名」。這樣做法存在兩個主要的問題：

2.1.1 只能針對訓練過的名字做分類

對於訓練過的名字，系統可以套用計算好的性別機率值進行預測；但是遇到沒見過的名字，系統沒辦法無中生有產生預測機率。這個缺點在英文名字中表現並不明顯，因為英文人名命名多從固定有限的名字庫「挑選」出適合孩子的名字，但中文名字創造性與組合複雜度都相對較高（Chang, 2003; Hsu, 1990），很容易碰到沒有訓練過的名字。在美國社會安全嬰兒資料庫（U.S. Social Security baby name database）中收錄了從 1880 年到 2010 年的英文人名資料，其中包含 28 億個嬰兒，但卻只包含了 6693 個人名，平均一個名字包含四千多人，分布相當集中。

2.1.2 許多名子男女皆可用，致使正確性無法提升

在以機率為基礎的模型中，系統將每個名字都歸到唯一的性別裡，雖然實際上存在許多兩性皆可使用的人名，因此此法的準確度決定在兩性皆有的人名數量與其中的性別比例。Gallagher 與 Chen（2008）指出，在英文人名中仍有極少數的名字存在雙性，而其中包含如 “Peyton”、“Finley”、“Kris”或 “Kerry”等兩性比例相近的名字。這樣的問題雖然也存在在中文人名中，但對於屬性完全相同卻存在不類別的物件，不管是何種分類器皆無法妥善處理。但在此提出的原因是，由於中文人名是由一到兩個字元所組成，每個字元又具有獨特的意義，因此可以掌握比英文名字更多的分類線索，使得分類時不需要考慮「整個」名字，而是對名字的每個「部分」中包含的性別傾向加以衡量。

綜合以上，以名字本身計算機率的方法雖然可以處理大部分的英文人名，但由於中文人名的獨創性與組成性質，此法未必能對中文人名的性別作出恰當的分類，且此法也忽略中文人名中明顯可用的屬性（字元）。

2.2 人名特性的相關研究

為了要找出除了名字「整體字面上」的性別資訊，我們必須要對人名所蘊含的其他資訊加以分析。由於先前缺乏中文人名文字特徵的研究，因此本節以探討英文的人名蘊含的特徵為基礎，並且對中文人名中蘊含的意義與命名來源作進一步的剖析，希望能使讀者對中文人名有基礎的認識。

2.2.1 英文人名中的屬性

Fryer 與 Levitt（2004）指出，在美國白人所使用的名字與非裔美人、西班牙裔美人、亞裔美人中均存在差異。該研究發現隨著時代的推進，黑人所使用的名字與白人的名字差異逐漸拉大，文獻中以經濟學觀點解釋命名的變化，認為隨著黑人權利受到重視，黑人認為取「具有黑人特色的名字」能象徵出生地的文化，也較能被其他黑人認同，增加族群間的向心力。由此可知，人類的命名除了分辨身分外，其中多包涵了命名者對受名者的期待。

除了人名在不同種族中使用的分布不同外，在不同性別中亦有其差異。Nastase 與 Popescu（2009）的研究以語言學中的「詞」為基準，其中使用了詞的發音模式（pattern）及尾碼（suffixes or word endings）等資訊，分析了羅馬文與德文等雙性別語言（two gendered languages）中名詞的性別，發現雙性別語言中陰性的詞與陽性的詞在「聽起來的感覺上」的確存在有不同。該研究搭配了支援向量機（Support Vector Machine）進行名詞的分類，得到了 72.36%的德文及 78.83%的羅馬文正確率。對於詞中性別的研究推展到人名上，也得到不錯的自動分類成果，Cassidy, Kelly 與 Sharoni（1999）基於發音特性，使用名字中的重音（lexical stress）、音節數（syllable number）與結尾處的非重讀音的原音（word final schwa）等資訊，並搭配類神經模型進行學習，在分類人名上達到了 80%的準確率（男性 79%；女性 81%）。

2.2.2 中文人名的特性

前文提到，中文的命名較具獨創性。中文人名可以使用一到兩個字「組合」而成一個新的詞，不像英文常從既定的「詞庫」選用。廖恭鳳（1991）蒐集了 1950 到 1955、1970 到 1972 以及 1985 到 1987 四個時期共 1800 份的台灣地區人名資料，他指出中文人名的命名來源如表 2-1：

表 2- 1：台灣地區男女人名命名者分布比例（多選）(資料來源：廖恭鳳，1991)

命名來源	男性	女性
父親	62.04%	67.48%
母親	27.20%	30.56%
祖父	19.83%	10.51%
祖母	4.82%	4.65%
其他長輩	8.22%	11.49%
算命	24.08%	22.98%

由表 2-1 可以看出，中文人名命名者除了父母等長輩外，存在有一大部分的來源來自於算命，而算命的主要透過生辰八字與筆劃數等方法決定人的名稱。以下試對生辰八字與筆畫說明之：

生辰八字源自古代的中國（李鐵筆，2009），是一種利用天干、地支來記錄年、月、日、時的方式，八字包含年干支、月干支、日干支等八個字，如甲子年、乙丑月、丙寅日、丁卯時等。年、月、日、時的干支組合稱為「柱」，因有四組，又稱四柱。八字中的天干、地支接蘊藏陰陽五行的概念，存在相生相剋的關係，在我國常用以推測人的吉凶禍福。而使用生辰八字算命的原則便是透過八字了解受名者先天的命格，如：是否陰陽協調、五行均等，再藉著取名來調和、改善運勢。比如由八字得知受名者個性任性刁蠻，則命理師會在名字中用「理」、「德」、「修」或「維」等字來修飾個性。

而筆劃的命名常稱「三才五格」（李鐵筆，2009），五格是由姓名的筆劃中算得，從筆劃中可算出相應的五行，以了解五格及五行間相生相剋與吉凶關係；三才是指天格、人格、地格間的組合。舉例而言，姓名「魏取向」可以算出天格（「魏」的筆劃數+1）有 19 劃屬「陽水」、人格（「魏」加「取」的筆劃數）26 劃屬「陰土」、地格（「取」加「向」的筆劃數）14 劃屬「陰火」、外格（「向」的筆劃數+1）7 劃

屬「陽金」、總格（姓名總比劃）32 劃屬「陰木」。從筆劃數得出陰陽五行的關係後，再分辨五行的相生相剋關係，避免五行相剋並透過如《熊崎氏 81 數姓名學》等對照表找出五格筆劃數的吉凶意義，並以此為命名考量。

在實務上，台灣命名會使用多種算命的方式為新生兒命名，表 2-2 整理上述研究（廖恭鳳，1991）中台灣人名的命名決定方法的統計結果，由於採用多選，我們可以發現一般會考量不同的方式進行命名。坊間也常見將生辰八字與筆劃結合；市面上也有軟體工具先計算生辰八字，再要求使用者填入姓氏，由程式為使用者計算名字中「吉」的筆劃與五行，讓使用者在推薦字的清單中挑選合適的人名（圖 2-1）。

表 2- 2：台灣地區男女人名命名模式分布比例（多選）（資料來源：廖恭鳳，1991）

決定方法	**男性**	**女性**
選筆劃	39.09%	43.52%
依照生辰八字	41.36%	37.65%
依照家族輩分	9.63%	4.89%
出生順序	20.96%	18.58%
期待繼承人	3.97%	1.96%
好聽的聲音	22.38%	33.25%
容易寫	14.16%	16.14%
期望	22.66%	11.25%
典型的字	1.70%	1.71%
優雅的字	15.58%	26.41%
特殊的字	4.25%	8.31%
紀念特殊事件	1.42%	1.47%
宗教信仰	0.57%	1.22%
根據特定的人	0.85%	0.49%
職業相關	0.57%	0.24%
依照實體特徵	0.85%	0.49%
無特定原因	4.82%	9.29%
感謝神	0.57%	0.0%
其他原因	2.83%	2.20%

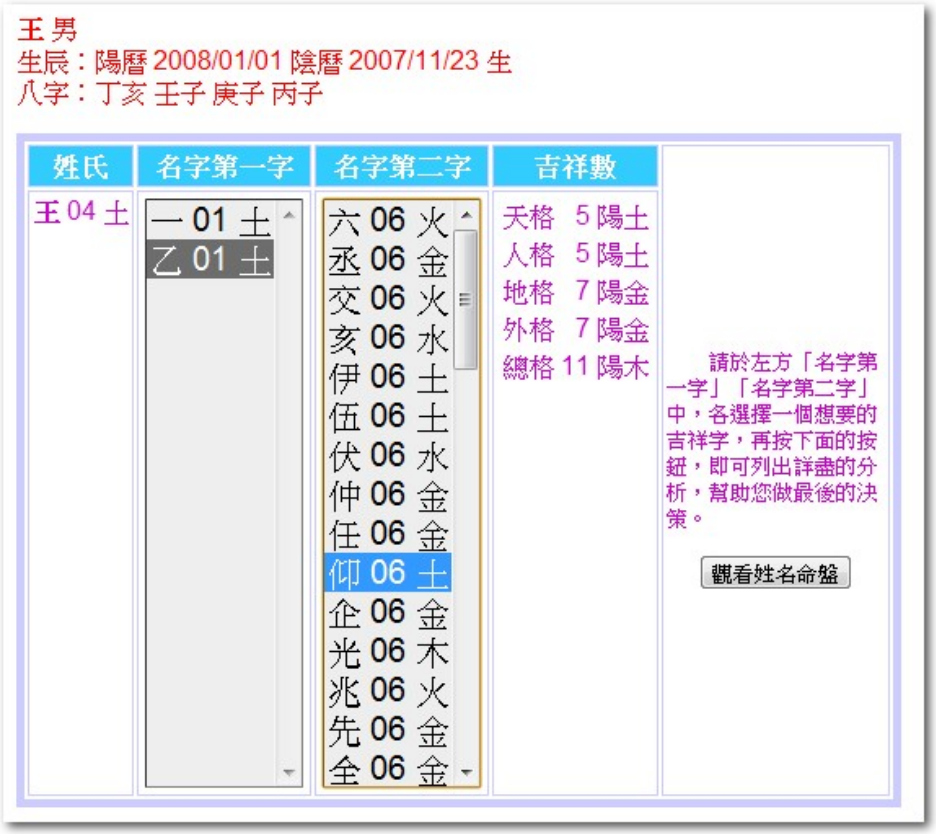


圖 2-1：嬰兒命名網站《兒女命名 e 點靈 plus [熊崎氏 81 數姓名學]》截圖

根據表 2-2 不論是男性或女性的名稱，都以筆劃數與生辰八字為大宗，而出生順序、好聽的發音與優雅的字亦佔相當大的比例。其中男性的人名比女性人名在統計上顯著有更多命名的原因來自期望，而女性名稱比男性名稱在統計上顯著有更多的命名並無特定原因。

表 2-3 是以問卷統計而成，該問卷提供如上述各種意義的選單，讓受試者勾選（多選）當時命名的期望。從表中我們可以發現台灣人名最多的期望都是「平安順利」，而「聰明好學」次之。在男女命名差異上，男性的名稱偏向於有成就、有領導能力、光宗耀祖、正直、誠實；而女性名稱偏向於快樂、生活安逸、英俊或美麗與整潔，另外沒有特別期望在女性人名上也遠多於男性。由此可知男女命名的期望上存在有一定的差異，而在選字上也應有所不同。

表 2- 3：台灣地區男女人名命名意義分布比例（多選）(資料來源：廖恭鳳，1991）

期望	男性	女性
平安順利	47.31%	46.94%
健康長壽	23.51%	24.69%
生活安逸	9.35%	13.69%
快樂	13.6%	23.47%
聰明好學	32.58%	29.58%
獨立，會分辨思考	14.16%	9.45%
有才能	15.3%	13.2%
有涵養	11.90%	7.36%
有成就	27.76%	9.78%
有領導能力	9.07%	3.42%
英俊或美麗	3.12%	10.02%
家庭幸福美滿	5.67%	9.29%
光宗耀祖	12.46%	1.47%
獲得友誼	0.28%	0.73%
造福社會	4.25%	2.20%
帶來子嗣或繼承子嗣	3.12%	1.96%
有宗族、豪族觀念	3.12%	1.22%
國家富強	0.85%	0.49%
國家民主	0.57%	0%
光復國土	0.28%	0%
紀念國家事件	0.28%	0.49%
富有寬裕	6.25%	3.67%
繼承祖業	0.28%	0.24%
有田宅	0.57%	0.24%
整潔	0.57%	3.18%
樸素	2.27%	4.40%
真誠	5.38%	5.62%
有禮	2.83%	1.47%
誠實	7.93%	3.67%
尊重他人	3.40%	1.22%
正直	9.63%	1.22%
信義	1.42%	0.49%
忠仁	3.12%	0.24%

孝順	7.93%	6.36%
勤勞	4.82%	2.69%
容忍	1.13%	2.93%
服從	0.28%	0.73%
謙遜	4.25%	2.69%
仁慈	5.38%	2.44%
沒有特別期望	7.93%	16.14%
其他	1.98%	1.22%

根據內政部（2010）的分析報告，表 2-4 整理我國人口姓名資料，其中對男女常見人名中常見的詞彙加以統計，我們可以發現男女常用字的前十名沒有重複的現象，在在說明了男女人名在用字遣詞上的不同之處。

表 2- 4：全國前十大常見的人名（資料來源：內政部）

名次	**男性**	**女性**
1	志明（14,265 人）	淑芬（33,562 人）
2	家豪（14,111 人）	淑惠（31,336 人）
3	俊傑（13,105 人）	美玲（28,104 人）
4	建宏（12,747 人）	雅婷（25,380 人）
5	俊宏（11,923 人）	美惠（24,593 人）
6	志偉（11,612 人）	麗華（24,154 人）
7	志豪（11,455 人）	淑娟（24,033 人）
8	文雄（11,255 人）	淑貞（23,981 人）
9	金龍（10,769 人）	怡君（22,341 人）
10	正雄（10,738 人）	淑華（20,644 人）

中文雖然每個字元皆有獨特的發音與意義，但相同發音的字元可能涵蓋許多不同的意義，與歐洲語系「以拼音為主」、「發音具有獨特的意義」等特性大相逕庭。例如中文「ㄐㄧㄝˊ」的發音，可能囊括的意義有：段落的「節」、出眾的「傑」、組構的「結」、乾淨的「潔」…等等，其中代表出眾（有成就）的「傑」字常出現在男性的名字；但乾淨（整潔）的「潔」卻常出現在女性的名字中，由此可知以拼音作為中文的特徵可能分類的效果沒辦法像英文一樣來的好。對此一問題，我

們必須尋求其他的解決辦法，找尋隱藏在文字型中的意義特徵。

2.3 中文的筆劃結構與型碼輸入法

介紹完中文人名後，本節側重於組成中文人名的中文字。如前所述，我們知道中文的人名具有意義性，而這些意義來自於中文字本身以及相互組合所產生。根據（許慎，121）中文字的構造是中文字意義的重要來源，因此本節試圖探究中文字的組成元素，期望對中文字能有進一步了解，而能在本研究中使用中文的「構造」對性別產生預測。

本段說明中文字的特性以及特性表示的方法，首先先就中文文字的組成加以探討，接著探討以中文文字結構為基礎的動態組字技術。

2.3.1 中文文字的組成與型碼輸入法

漢字最小的組成單位是「筆劃」，包括橫（一）、豎（丨）、撇（丿）、點（丶）、捺(㇏)、折（𠃌）等六種，再來可由筆劃構成一些小單位，稱為字根或部件。根據（倪耿，1972）若筆劃可單獨形成一個漢字，則稱為「字根」，如「日」、「月」、「木」等；若筆劃無法單獨成為一個漢字，則稱為「部件」，如「夂」、「乂」、「扌」。另外，漢字本身又可能組成其他漢字，例如「盟」是由「明」和「皿」所構成的。

因此，漢字、部件與字根的定義存在重疊，該研究在尋找字根時，為了找到適當的字根量，避免字根種類太少使得一個字要由過多的字根組成（例如以筆劃當字根，只有八種字根，則每個字的字根數就是它的筆劃數）；字根種類太多則失去字根的意義（每個字都形成一個字根，就無法使用字根化繁為簡），他們對《中文電腦基本用字表》中之 9,129 個字形，使用統計方法並搭配如（表 2-5）規則刪

減，逐字分解而得 496 個字根。他們統計每個字根或部件在字表中的出現頻率，如果些特定的筆畫集合在字表中出現的頻率高於某個門檻,則將該集合視為字根，如此可以避免字根過多或者是字根出現率過少的問題。另外他們也針對出現頻率較低的筆畫集合繼續往下拆解，檢視筆畫集合內部的子集合是否可以用來生成字根。

表 2- 5：機率選擇字根策略（資料來源：耿倪，1972）

出現頻率	處理方式
≥ 0.003758	當作字根
0.001879~0.003758	不可分解為 2 個以上的字根
0.001236~0.001879	不可分解為 3 個以上的字根
0.000939~0.001236	不可分解為 4 個以上的字根
其餘	可任意分解

該字根系統是漢字字根最早的研究。這些字根總計使用頻率超過字表中的 50%，依照使用頻率加權後的平均每個字的字根數僅 1.9。

使用字根技術在輸入法中實作稱為型碼輸入法，典型的如倉頡輸入法、鄭碼輸入法、五筆字型輸入法、大易輸入法等等。型碼輸入法將字根編碼，再以特定順序產生序列以表示特定文字。由於漢字的發音由聲母、韻母與音調組合，僅有 1,300 多種，有明顯一音多字的現象，因此基於聲音的音碼輸入法重碼率比型碼輸入法高上許多。下文中將會對不同的方法加以說明：

2.3.2 倉頡輸入法

倉頡輸入法原名「形意檢字法」，由朱邦復先生於 1976 年所創，用以解決電腦處理漢字的問題，包括漢字輸入、字形輸出、內碼儲存、漢字排序等。由於發明時正值三軍大學發展中文通訊系統，當時三軍大學校長蔣緯國以譽之為上古倉

頡造字精神，重新定名為「倉頡輸入法」。

倉頡輸入法基於倉頡碼所作，而倉頡碼目的是統一處理漢字的形、音、義、碼、序、辨等六大問題。透過字首及字身的概念，以字首做為分類，字身做為補充，讓電腦「理解」漢字，達到組字、字義理解，甚至與人溝通的功能。由於本以漢字檢索為目的，倉頡取碼依據視覺辨識原理，能反映漢字的細微特徵。
目前倉頡輸入法最新的公開版本是第五代，最常使用版本是第三代。

表 2-6：倉頡字母分類表

哲理類	筆劃類	人體類	字形類	特殊鍵
日（A）	竹（H）	人（O）	尸（S）	難（X）：用於
月（B）	戈（I）	心（P）	廿（T）	特殊字，如
金（C）	十（J）	手（Q）	山（U）	「卍」
木（D）	大（K）	口（R）	女（V）	重（Z）：用於
水（E）	中（L）		田（W）	輸入標點符號
火（F）	一（M）		卜（Y）	
土（G）	弓（N）			

倉頡碼並非使用筆順將漢字編碼，而是依照視覺順序分為字首及字身，字首為最左、最上、最外部份，剩餘部份為字身。若字身可以再分，則分為次字首和次字身。在取碼方面，字首最多取二碼，字身最多取三碼；無法明確分割為字首、字身者，則全取。一個漢字最少用一碼輸入，最長則為五碼（表 2-7）。

表 2-7：倉頡碼的幾個例子

漢字	分割	取碼	鍵盤按鍵
出	山、山	山山	UU
理	王、里	一土、田土	MGWG
菇	艹、女、古	廿、女、十口	TVJR
語	言、五、口	卜口、一一、口	YRMMR

2.3.3 鄭碼輸入法

鄭碼亦稱區位碼、字根通用碼，其正式名稱是「字根編碼輸入法及其設備」，屬字形輸入法。由中國文字學家、《英華大詞典》主編鄭易里先生所作，主要對象為簡體中文的使用者。

鄭碼定義「基本字根」作為輸入的基礎，透過將漢字拆解為字根序列達到輸入的目的。鄭碼的字根是以「區碼」跟「位碼」所組成的雙碼編號，如「羊」為 UC，U 為區碼，C 為位碼。在拆字方面，鄭碼以筆順作為拆字順序，並將一個漢字最多以四個碼表示。方法為計算一個漢字可拆分成幾個基本字根，並以此來套用規則以決定字根使用的是區碼，還是區碼及位碼。

2.3.4 五筆輸入法

五筆字型輸入法簡稱五筆，亦稱為王碼，為王永民先生在 1983 年所創。五筆名稱來自將漢字筆劃分為五個區：橫、豎、撇、捺（同點）、折（同提）。五筆字型完全依據筆劃和字形特徵對漢字進行編碼，主要用於使用簡體中文的中國大陸。

五筆輸入法使用字根組字，取碼時最少使用一碼，最多四碼。鍵盤上有多組字根，每一個鍵上有一個最常用的字根稱為「鍵名字」，另外亦有幾個次要字根稱為「成字字根」。若一個字是由多個鍵名字或成字字根組成，則依照是否可分解成四個字根判斷，若可，則以字根輸入；否則需要加上識別字。

另外，五筆輸入法有收入一些片語，如「自」（THD）+「相」（SHG）+「矛」（CBTR）+「盾」（RFHD）可以取每個字的一個字跟（TSCR）輸入；在常用字方

面，具備「簡碼字」，如「的」可直接打 R 輸出。

2.3.5 大易輸入法

由王贊傑先生於 1988 年所創，使用 46 個字碼，每個字碼具有多個字根以供拆字。其字根多借用漢字傳統造字原則，含有大量的部首，使熟悉書寫漢字的使用者容易拆解單字成為不同的字根。在拆字方面，使用筆劃順序，一個單字最少取一碼，最多取四碼。若超過四個字碼則按照筆順取最先出現的前三個字碼以及最末字碼來輸入。稱為「取首三尾一」。

2.3.6 嘸蝦米輸入法

嘸蝦米輸入法為劉重次先生於 1989 年發明，嘸蝦米一詞，為閩南語「沒啥物」之國語音譯，意思為「沒什麼」，外文則使用「Boshiamy」一名。其將漢字拆解成數個字根結構，再以字根的形、音、義與英文字母加以聯想，拼出漢字（表 2-8）。嘸蝦米最大優點是以 26 個英文字母為字根，可同時練習中、英打。

表 2- 8：嘸蝦米字母分類表

分類	舉例
形狀	「命」，由「A」、「O」、「P」組成
發音	「西」是 C 「爾」是 R 「平」是 P
意義（聯想）	「車」是 C（Car） 「女」是 G（Girl） 「手」是 H（Hand） 「水」是 W（Water） 「火」是 F（Fire）
其他（必須記憶）	「土」是 Y

	「鬼」是 V

取碼規則每字取一碼至四碼，拆字採用視覺順序，不完全與筆順符合。例如「彎」字上部由左至右拆成 S（糸）、I（言）、S（糸），而非按筆順先取 I（言）。

2.3.7 型碼輸入法總整理

在比較輸入法時，通常考慮輸入碼的重碼率。重碼率指的是相同的輸入鍵所對應到中文字的數量所換得之機率，其可拆解為「重碼字根」與「重碼字」。重碼字根指的是一組輸入的按鍵能對應多個漢字的情況；而重碼字就是共享同一組輸入的漢字。本研究以 Big-5 的 13,063 個漢字，在不考慮出現頻率的環境下計算而成，結果整理如表 2-9。

表 2- 9：型碼輸入法總整理

輸入法	倉頡	鄭碼	五筆	大易	嘸蝦米
發明時間	1976 年	1980 年代	1983 年	1988 年	1989 年
發明者	朱邦復	鄭易里	王永民	王贊傑	劉重次
字根	使用鍵盤 25 個字母，分為哲理類、筆劃類、人體類、字形類與特殊鍵。	分為區碼及位碼，每個字根用兩個鍵盤符號表示。	使用鍵盤 25 個字母，每個鍵再分為鍵名字與成字字根多組字根。	使用鍵盤 46 個鍵，字根包含大量部首	使用鍵盤 26 個字母，分為形狀、發音、意義與其他。
取碼數	1~5 碼	1~4 碼	1~4 碼	1~4 碼	1~4 碼
拆字順序	視覺順序，分為字首、字身	筆順	筆順	筆順	視覺順序
重碼字根	508 個	1,149 個	1,134 個	593 個	3,007 個
重碼字	1054 個	2,490 個	2,426 個	1,284 個	6,946 個
台灣地區使用比例	9.9%	無資料	無資料	2.6%	10.7%

2.4 分類與預測工具

本研究最終目的是建立一套分類機制，區分男性和女性的人名，因此接續中文人名的表示方法後，本文以分類器的角度，介紹目前較為常見的分類技術、分類器及其主要應用，期能對分類技術作全盤的了解。

2.4.1 簡單貝氏分類（Naïve Bayes Classifier）

簡單貝氏分類是一個基於貝氏定理（Bayesian Theorem）的機率分類方法，假設各屬性彼此間是互相獨立的（conditional independence），透過監督式學習訓練樣本，紀錄分類根據所使用屬性的關係，產生這些訓練樣本的中心概念，再用學習後的中心概念對未歸類的資料進行類別預測，以得到受測試資料物件的目標值。

每筆訓練樣本，一般含有分類相關連屬性的值，及分類結果；一般而言，屬性可能出現兩種以上不同的值，而目標值則多半為二元的相對狀態，如「是／否」、「增／減」…等等。

2.4.2 決策樹（Decision tree）

決策樹透過不同的屬性來分類，屬於屬性導向（attribute-based）的分類方法。樹中的節點（Node）表示特定屬性，都安排一個適當的測試條件，利用該測試結果來決定資料將分類在哪一棵子樹發展，分叉路徑則代表屬性中的值；葉結點（Leaf node）則代表案例所對應的類別。

建立一顆決策樹步驟如下：

1. 依據屬性選擇指標，從現有的屬性中，挑出分類能力最好的屬性做為樹的內部節點。
2. 將內部節點的所有值，產生出對應的分支針對每一個新產生的分支，將訓練資料重新排列，以進行下一個內部節點的產生。
3. 重複上面的過程（Recursive），直到滿足終止條件。

常用的選擇指標有使用資訊獲利（Information Gain）的 ID3 演算法、使用獲利比例（Gain Ratio）的 C4.5 演算法以及使用吉尼係數（Gini Index）的 CART 演算法等。

2.4.3 K-最鄰近分類（K-Nearest Neighbor, K-NN）

K-最鄰近分類法透過向量空間模型來分類，主要概念是相同類別的案例，彼此的相似度高。因此藉由計算與已知類別案例之相似度，就能推論未知案例可能的類別。它和一般的機器學習方法的主要不同在於它不是處理所有訓練資料，其尋找相似度最高的 K 份文件，並檢查是否大於門檻值且以二元法表示，即結果為 1 或 0，最後將這些值加總作為分類和測試文件的強度，決定待分類之文件類別。

如圖 2-2 所示，中心綠色的點為未知分類的資料，而紅色三角型與綠色圓形是已分類好的案例。若將 K 設為 3，則會比較最鄰近的三個案例，發現紅色有兩個；藍色有一個，因為紅色比較多，因此將未知的資料分類為紅色三角形。若將 K 設為 5，則藍色方形較多，故將未知資料歸為藍色方形，以此類推。

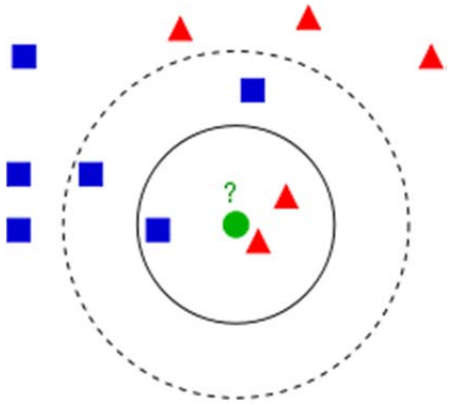


圖 2- 2：對於未知的點（綠色），以不同的 K 值算出不同的結果（取自網頁：

http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/jthomas/knn.html）

K-最鄰近分類的分類步驟如下，假設有 j 個以之集合的案例 D1、D2…、Dj，Item 是待分類的項目：

1. 使用相似度公式 Sim 分別計算 Item 與不同集合之相似度，得到 Sim(Item, D1)、Sim(Item, D2)…、Sim(Item, Dj)。
2. 將 Sim(Item, D1)、Sim(Item, D2)…、Sim(Item, Dj)排序，若相似度超過預先設定的門檻，則將 Item 放入鄰居案例集合最鄰近鄰居。
3. 自鄰居案例集合最鄰近鄰居中取出前 K 名，依多數決，得到 Item 可能類別。

K-最鄰近分類最令人詬病的缺點是需要大量的計算，因為每個未知樣本在計算距離時必須要跟全部已知類別比較後，才能找出離特定未知樣本最近的已知樣本。此外，若類別比例不平均，則很可能都預測成比例高的類別。

2.4.4 類神經網路（Neural Network）

類神經網路是一種基於腦與神經系統研究，所啟發的資訊處理技術，因類似人腦神經的結構而得名。類神經網路具有人腦基本功能，如記憶和歸納。與迴歸分析不同，類神經網路沒有任何假設的機率分佈，而是透過模式識別和誤差最小

化的過程，在每一次經驗中提取和學習資訊。

類神經網路主要結構是由神經元（neuron）、層（layer）和網路（network）三個部份所組成。神經元透過權重（weight）相互連接，而根據接觸面不同，可分為輸入層、輸出層和隱藏層，每一層的每個神經元和前一層、後一層的神經元連接，整體成為網路的形式。輸入層只從外部環境接收資訊，該層的每個神經元相當於自變數，不完成任何計算，只為下一層傳遞資訊；輸出層生成最終結果，為網路送給外部系統的結果值；隱藏層介於輸入層和輸出層之間，這些層完全用於分析，其函數聯繫輸入層變數和輸出層變數，使其更配適資料。

類神經網路必須透過訓練（training）才能正常運作，以確保每個輸入都能正確對應到所需要的輸出。簡言之，訓練就是使用樣本不斷調整權重的過程，目前大多透過誤差度量來調整權重。先前的研究中，（Cassidy, Kelly and Sharoni, 1999）就是透過神經網路學習名字的拼音規則，用以預測英文人名的性別傾向。在該實驗中，使用三層的前饋網路（feed-forward network），輸入層包含 44 個單元（神經元）、隱藏層有 15 個單元，而輸出結果為性別，因此只有兩個單元（男性和女性）。

2.4.5 支援向量機（Support Vector Machine, SVM）

支援向量機是一種監督是學習的方法，以統計理論為基礎，且有廣泛的應用能力，在處理分類問題中廣泛被使用。其主要概念是在多維空間的資料中，利用一個多項式或是三角函數組成的方程式形成超平面（hyper-plane）區分資料的所代表的座標。這個超平面能夠達到將兩類點分的最開，利用找出來的核心函式（kernel function）將資料的座標輸入，即可知道資料屬於哪個類別，即可達成分類的目的。

Nastase 與 Popescu（2009）的研究中以發音的方式使用支援向量機對德文與

羅馬文的詞彙性別作分類，得到了不錯的分類結果；Bergsma, Lin 與 Goebel（2009）使用半人工的方法選出特徵，再搭配支援向量機對英文的專有名詞性別作分類，得到 95.5%的正確率。此外根據（Joachims, 1998）比較簡單貝氏分類（Naïve Bayes Classifier）、K-最鄰近分類（K-Nearest Neighbors）、類神經網路（Neural Networks）、支持向量機（Support Vector Machines）和線性最小平方適配（Linear Least Square Fit）等五種分類演算法在文件分類之正確率，其發現以 SVM 正確最高，故本研究以 SVM 作為主要的分類方法。

2.5 小結

本章從人名中性別分類開始，介紹人名的內容及其特性。我們發現人名具有其意義性，但中文人名與英文人名不同。除了以往的研究多在語言學中進行，缺乏計算機背景外，中文在結構與命名方式也與英文有很大的不同。

接著我們從中文拆解的角度，模擬以發音特性拆解英文單字，找到有效表達中文文字結構的方法。我們比較不同的字形分解理論，發現倉頡在構字上所有字皆以字型表達，不向其他輸入法會因字的出現頻率而改變表達方式，且倉頡的重碼率低，符合我們表達文字的特性。

最後我們比較現有的分類演算法，企圖對中文人名的性別提出最好的分類機制。不同於先前研究，本研究透過既有拆解中文字的研究，找出中文字特徵分類的依據，應用於尚未在中文文字探勘領域中的性別分析上。

第三章 性別預測系統之設計

基於前述的研究，本章將詳細說明在本實驗中的分類器架構與內容。我們先以 K-最鄰近法作為實驗的基準線（Baseline），介紹 K-最鄰近法在人名性別預測中的實作架構，進而介紹基於支援向量機中以字元為基礎（Character）及以倉頡編碼（CangJie Input Method）將文字拆開表示等兩種建立特徵的方式。由於支援向量機不同系統的差別只有特徵的不同，因此我們設計一套模組化的預測架構可以動態的選擇訓練與測試資料，並且能抽換與組合不同特徵表示方法。詳細說明如下：

3.1 基準線系統設計

根據先前的研究，K-最鄰近分類設計簡單，且不需要經過訓練的階段，可以直接根據資料的屬性對資料進行分類，適合做為實驗初步的判定，因此我們將此法作為實驗的基準線（Baseline），期望本實驗的結果能優於此分類器。本法在以下實驗過程中以 Baseline 簡稱，以下說明兩個核心的環節—設定相似度公式與訂定 K 值：

3.1.1 設定相似度公式

K-最鄰近分類的核心就在相似度公式，如前文所述，一般而言使用餘弦相似度（Cosine Similarity）相似度或歐幾里得距離（Euclidean distance）作為相似度的計算。但因為在本實驗中我們的對象是每一個中文字，若以「字」作為人名的維度，在計算餘弦相似度或歐幾里得距離時，遇到不同的字若直接以字的順序作為座標則毫距離的意義。因此我們先將字轉換為倉頡碼，再使用較為簡單的「最短編輯距離（Shortest Path Edit Distance）」作為相似度的計算。編輯距離越大，則代

表越不相似。

最短編輯距離是計算從字串 A 變為字串 B 最少需要幾個插入、刪除與取代的動作數量，例如若字串 A 是 eeba、字串 B 是 abac，則從字串 A 到字串 B 的最短編輯步驟為：

1. 將字串 A 的第一個 e 取代為 a
2. 刪除字串 A 的第二個 e
3. 將 c 接(插入)在字串 A 之後

由於最少需要 3 個步驟，故最短編輯距離為 3。本實驗中我們將每個中文字轉換為倉頡碼，再透過最短編輯距離的方法計算字與字的編輯距離，以得到相似度。值得一提的是，由於計算最短編輯距離需要不斷的遞迴運算，本程式為了提升效率，採用動態規劃的方式實作，儲存之前算好的結果，以降低重複計算所花費的時間。根據估計，如果沒有使用動態規劃，此法中一萬筆的人名資料以 Java 程式處理，需要長達約 3 個月的時間。

3.1.2 K 值的設定

K 值代表的是選出與目標最接近的 K 個鄰居，藉由鄰居的類別以「投票」的方式決定目標的類別。為了讓投票能產生唯一的結果，我們在選擇 K 時一定是以奇數作為考量。由於在不同的領域中 K 的數值是不同的，因此我們必須要找到在此「人名-性別」問題中最佳的 K，以進行性別的判定。本實驗使用十折交叉驗證（10-fold cross-validation）將 k 的值從 1 帶到 27，試圖找出最適當的 K。

3.2 以倉頡表達中文特徵

在介紹應用支援向量機進行性別預測前，我們先針對其中的特徵表達（feature representation）作詳細的說明。在特徵表達上，我們考慮兩種基本的模式 — 「每個字元視為一個維度」與「將一個字元以倉頡編碼拆成多個維度」。由於每個字元視為一個維度已經相當直觀，就不再贅述。本節將詳細介紹以倉頡碼表達中文字的方法。

3.2.1 字碼對照表

不管是「每個字元視為一個維度」或「將一個字元以倉頡編碼拆成多個維度」都需要對中文字作維度的轉換，在此我們透過對照表完成。在「每個字元視為一個維度」的情況時，我們建立字典檔作為對照表。字典檔是統計人名資料庫中出現過的中文字，並加以編號的檔案，以代表該文字的維度編號（如圖 3-1）。欄位 A 為中文字元，欄位 B 為該字元的維度編號。

	A	B
1	羙	0
2	玉	1
3	秀	2
4	文	3
5	明	4
6	淑	5
7	惠	6
8	麗	7
9	志	8
10	華	9

圖 3- 1：本研究建立之字典檔

在「將一個字元以倉頡編碼拆成多個維度」中，我們要將中文字轉換為倉頡碼，因此我們建立了倉頡對照表，如（圖 3-2）。我們採用《第五代倉頡通用版原始碼表》，共收錄 21,541 個繁簡體的漢字。欄位 A 為中文字元，欄位 B 則是對應的倉頡碼。以第一個字「覃」為例，用倉頡輸入法需要鍵入「一田日十」，對應鍵盤按鍵則為 mwaj。因此，我們將中文字「覃」以 mwaj 來表示。

	A	B
1	覃	mwaj
2	覈	mwhio
3	覓	mwv
4	覀	mllw
5	覇	xmbtj
6	覆	mwhoe
7	覇	mvphh
8	覗	mwoik
9	見	buhu
10	覊	mwtjf

圖 3-2：倉頡碼對照表

3.2.2 中文字以倉頡編碼（Uni-gram）

在「將一個字元以倉頡編碼拆成多個維度」中，每個中文字可以對應一組鍵盤輸入，例如「覃」可以對應 mwaj。因此我們可以將每個字用 25 個維度的向量所表示（倉頡共使用 a 到 y 等 25 個字母對應中文 25 個筆劃結構）。若「覃」字以向量表示，我們可以將使用一次的筆劃設為 1；沒被使用到的筆劃設為 0（如圖 3-3）。在我們的實驗設計中，若某個字使用 n 個相同的字母，則該維度的數值為 n。

覃 =(1, 0, 0, 0, 0, 0, 0, 0, 0, 0 ,1 ,0 ,0 , 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0)

圖 3-3：以向量維度表示倉頡碼的「覃」字

通常一個人名由兩個字所組成，但我們只會用一組向量所表示，因此在使用本方法（uni-gram）的時候，為了跟以字元為基礎的方式有所區隔，我們把多個字的維度疊加在一起，以表示同一組人名。

3.2.3 中文字以倉頡編碼（bi-gram 以上）

考慮將文字以 bi-gram 或以上的筆劃組合當作維度使用，乃因倉頡碼的排列含有順序性。將中文字以倉頡編碼時會依照書寫的習慣，由上而下、由左而右拆解文字的筆劃，因此在倉頡表示時，「代」是「人戈心」而「怜」是「心人戈戈」，雖然都有「人」、「戈」和「心」，但若以單連詞（uni-gram）的向量表示，兩字只差在維度「戈」的長度，可是實際上這兩個字的形體差異頗大。

鑒於此一問題，我們以雙連詞（bi-grams）等方式擴充可使用的維度，再以「代」為例，我們將之拆解為使用「人戈」、「戈心」兩個維度，如此可以保留部分結構的順序性，使向量能更正確的區分文字。如前所述，中文的人名多由兩個字組成，以「偉強」搭配倉頡雙連詞為例，「偉」可拆解為「人木一土」、「強」可拆解為「弓戈中戈」，則「偉強」使用到的維度是「人木」、「木一」、「一土」、「弓戈」、「戈中」與「中戈」等六個。值得注意的是，為了保持每個字的獨立性，字與字之間的筆劃不會以雙連詞連結，因此不存在「土弓」的維度。

3.2.4 結合 uni-gram、bi-grams 與 tri-grams 的方法

除了前文介紹的單連詞、雙連詞等方法外，本研究企圖在一組向量中結合不同的連詞（grams）組合，以降低切割不正確所帶來的影響。舉例而言中文字「琪」可能為女性名字中常見的字，但其他相關的「棋」、「淇」或「祺」也可能出現在女性名字中。考慮表 3-2：

表 3-1：用倉頡碼表示「其」的相關字

中文字	倉頡碼	uni-gram	bi-grams	tri-grams
琪	mgtmc	m, g, t, m, c	mg, gt, tm, mc	mgt, gtm, tmc
棋	dtmc	d, t, m, c	dt, tm, mc	dtm, tmc
淇	etmc	e, t, m, c	et, tm, mc	etm, tmc
祺	iftmc	i, f, t, m, c	if, ft, tm, mc	ift, ftm, tmc

以人類目視可以發現以上四個字都共享「其」(tmc）這個部件，但是這個部件要到三連詞（tri-grams）才能正確被表達。以雙連詞而言，雖然存在 tm 與 mc 兩個維度，但是包含這兩個維度的字為數眾多，有可能被其他的字干擾，而失去「其這個部件有女性特徵」的特性。我們使用混合單連詞、雙連詞與三連詞的方法，在支援向量機中完整呈現不同的組合方式，讓支援向量機引擎自行強化具有判別力的字根組合，以達到更有彈性的拆解字根方式，期望能產生較佳的分類結果。

3.3 支援向量機系統設計

如前所述，本研究以分析名稱與性別為對象，使用支援向量機建立預測模型。在系統設計時，為了能彈性的套用各種特徵表示機制，我們將特徵編號與向量編碼拆開成獨立的模組，包藏在訓練與測試兩個類別中，並使用物件導向的繼承動態替換所使用的再搭配資料庫中的人名性別資料。在本實驗中，我們考慮以支援向量機搭配下的特徵表達方式：

1. 字元為基礎（Character-Based）的特徵表達，以下簡稱 SVM-CB；
2. 倉頡碼表示中文字元（每一個鍵視為一個維度，逛 25 個維度），以下簡稱 SVM-CJ1；
3. 倉頡碼表示中文字元，並使用雙連詞連結相鄰鍵，以下簡稱 SVM-CJ2；
4. 倉頡碼表示中文字元，並使用三連詞連結相鄰鍵，以下簡稱 SVM-CJ3；

5. 倉頡碼表示中文字元，並使用四連詞（4-grams）連結相鄰鍵，以下簡稱 SVM-CJ4；
6. 結合倉頡單連詞、雙連詞與三連詞，以下簡稱 SVM-CJ123；
7. 結合倉頡單連詞、雙連詞、三連詞與四連詞，以下簡稱 SVM-CJ1234。

基於上述方式間差異小、組合方法複雜，並同時考量系統要動態讀取人名資料集，因此將系統架構設計如下：

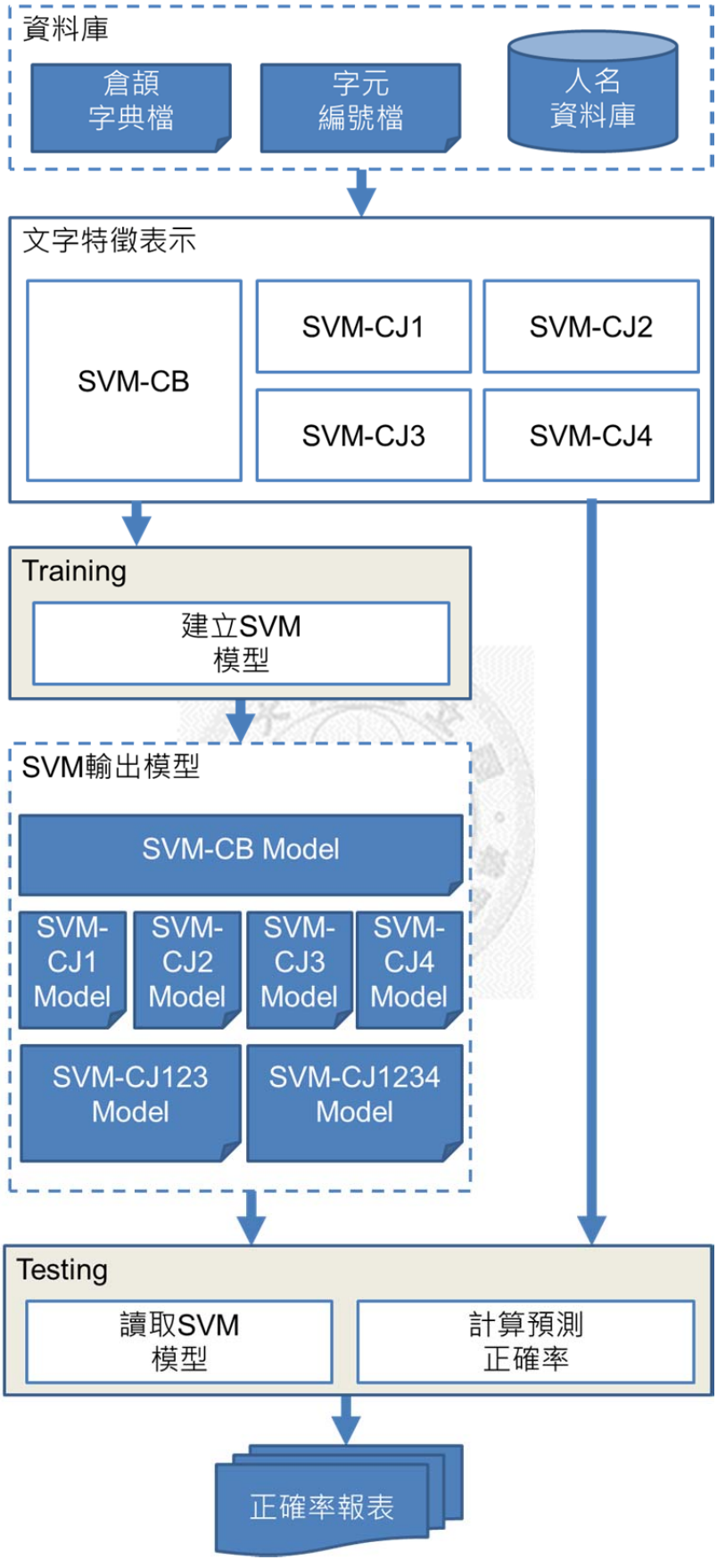

資料庫
倉頡
字典檔
字元
編號檔
人名
資料庫
文字特徵表示
SVM-CB
SVM-CJ1
SVM-CJ2
SVM-CJ3
SVM-CJ4
Training
建立SVM
模型
SVM輸出模型
SVM-CB Model
SVM-
CJ1
Model
SVM-
CJ2
Model
SVM-
CJ3
Model
SVM-
CJ4
Model
SVM-CJ123
Model
SVM-CJ1234
Model
Testing
讀取SVM
模型
計算預測
正確率
正確率報表

圖 3-4：系統架構圖

如圖 3-4 所述，完整的系統可以被分為訓練（Training）與測試（Testing）兩種操作，而訓練階段的主要目的就是建立不同特徵方法所產生的模型；測試階段則是透過模型對輸入的人名進行判別，並統計整體預測的準確度。

第四章 資料處理與實驗結果

本章描述實驗中測試與訓練資料的特性，並說明實驗的過程與系統效能。第一節介紹人名資料的蒐集與處理工作，並先就資料的內容提出觀察結果，協助讀者更了解往後實驗結果的特性；第二節說明人名預測的實驗結果，並比較不同方法的效能。

4.1 資料蒐集

為了讓系統充分的對中文人名與性別關係進行學習，本研究援引台灣某醫院歷年來的病歷資料。其中資料包含四個部分：編號、姓名、性別與出生年。以下分成兩個部份說明：

4.1.1 資料清洗

由於資料蒐集時間長且來源繁多，資料中含有許多重複、無法使用的名字，因此必須對資料進行清洗。清洗的步驟與詳細內容如下：

1. 移除重複的資料：我們假設在資料表中同年同月同日生且同名同姓的資料即表示為同一人，因此本步驟先透過微軟 Office Excel 的功能將生日與姓名欄相同的重複資料刪除。
2. 移除缺乏性別及名字中的空格：我們使用 Java 程式去除名字中的空格，並直接跳過缺乏性別欄位的資料。因為若缺乏性別欄位，空有名字是無法讓系統進行訓練與測試的。另一方面，也必須對姓名欄位中的空格移掉，避免單名中姓與名中間有空格而造成系統的誤判。

3. 移除名字中不含中文字或長度不符合規定者：經過觀察，最顯而易見不符合「正常」人名的特徵是名字的長度不正確，因此我們一樣透過程式保留長度在 2 到 4 個字之前的姓名，兩個字的名字是單名，此舉可以移除如「廖億菁之子」等姓名；四個字的名字是複姓加雙名或冠夫姓的婦女名。另外我們發現部分的名字中參雜非中文的字元，我們透過 unicode 的中文編碼的範圍作為限制，僅保留名字中全為中文人名者，故可移除如「顧晏＊」等姓名。
4. 保留姓氏正確者，並將姓氏與人名分開：由於系統只針對人名加以研究，必須將姓氏與人名切開，才能確保姓氏不影響我們觀察到的人名結果。在此為了簡化資料清洗的作業與提升資料正確性，我們借助（內政部，2010）中全台常見姓氏，取出列表中的前 200 名作為參考標準。我們設定若姓名中的前兩個字（考慮複姓與冠夫姓）不存在前 200 名姓氏中，則該姓名則為垃圾資料，此舉可以去除如「瑪妮妲」等姓名。

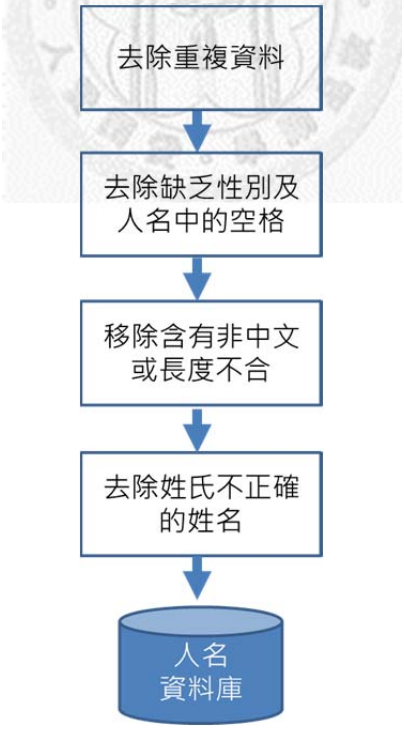


圖 4- 1：人名資料清洗的流程

圖 4-1 描述清洗的流程，經上述流程，共過濾掉約原本 7.37%的資料，其錯誤

資料分布如表 4-1：

表 4-1：錯誤比例分部表

類型	百分比
包含非中文	4.45%
沒有中文字	2.87%
沒有性別	1.32%
姓名共二字，姓氏規則錯誤	8.85%
姓名共三字，姓氏規則錯誤	16.88%
姓名共四字，姓氏規則錯誤	4.88%
長度不正確	60.74%

4.1.2 資料儲存

資料清洗後，我們使用剩下的 271,092 筆資料，由於資料整理時已被排序，因此我們用亂數將每一筆資料編號，並儲存於 PostgreSQL 資料庫中。儲存在資料庫的目的是方便未來動態選取訓練與測試資料，因為我們可以用很點單的 SQL 語法限制資料的數量、性別與內容。在資料庫的設計中我們一共使用四個欄位（表 4-2）：

表 4-2：儲存人名資料的資料表

欄位	中文名稱	型態	說明
id	編號	integer	由一開始的漸增編號
name	人名	character varying(30)	儲存 1~2 個字的中文人名
sex	性別	character(1)	以字元 M 代表男性；F 代表女性
rand_key	亂數編號	integer	協助以亂數抽取人名；也可以透過對亂數求於數過濾資料。

4.1.3 資料概觀

對資料內容進行統計，詳細結果如表 4-3。在 27 萬個人名中，男女比例為 103 比 100，兩性接近相等，而總人名數有 93,106 個，代表平均一個人名可以包含 2.91 個人，但實際上的分布極不平均，我們稍後再作討論。另外人名與性別的關係，我們可以發現只出現在男性的人名有 56,022 筆，而只出現在女性的人名卻只有 33,372 筆，差距相當大。由於人口比例均等，所以我們可以推斷女性人名的重複度高於男性，而男性人名的歧異性較高，與先前對於美國人名的研究（Fryer 與 Levitt, 2004）相反，應該是中文人名的一大特性。

表 4-3：人名資料概觀

項目	人名數	所佔比例
總人數	**271,092**	
男性	137,542	總人數之 50.74%
女性	133,550	總人數之 49.26%
總人名數	**93,106**	**總人數之 34.34%**
只有在男性出現的人名	56,022	總人名數之 60.17%
只有在女性出現的人名	33,372	總人名數之 35.84%
兩性皆會使用的人名	3,712	總人名數之 3.99%
男性比例較高的人名	57,302	總人名數之 61.54%
女性比例較高的人名	34,611	總人名數之 37.17%
兩性比例完全相等的人名	1,193	總人名數之 1.28%

圖 4-2 橫軸中我們將人名依照出現數量排名，出現最多次的人名為第一名，次多為第二名，以此類推；而縱軸為涵蓋人數的百分比。我們發現在男性中前 100 名占了 7,730 個人，約為總男姓名的 5.62%；而女性的前 100 更佔了 18,236 人，約為全部女性的 13.65%。男性的前 1,000 名佔了 20.33%，女性則為 38.2%，而到了前一萬名時，男性佔 51.33%，女性佔 66.13%，都超過總人數的一半。另外從圖中我們可以發現，兩性在超過 10,000 名後，名次與人數的關係開始由指數轉為線性

成長，此現象可以推測出兩性中都有許多人名是只對應到一個人或少數人的，且這種問題在男性中尤為明顯，也在在表示了男性人名歧異度高於女性的現象。

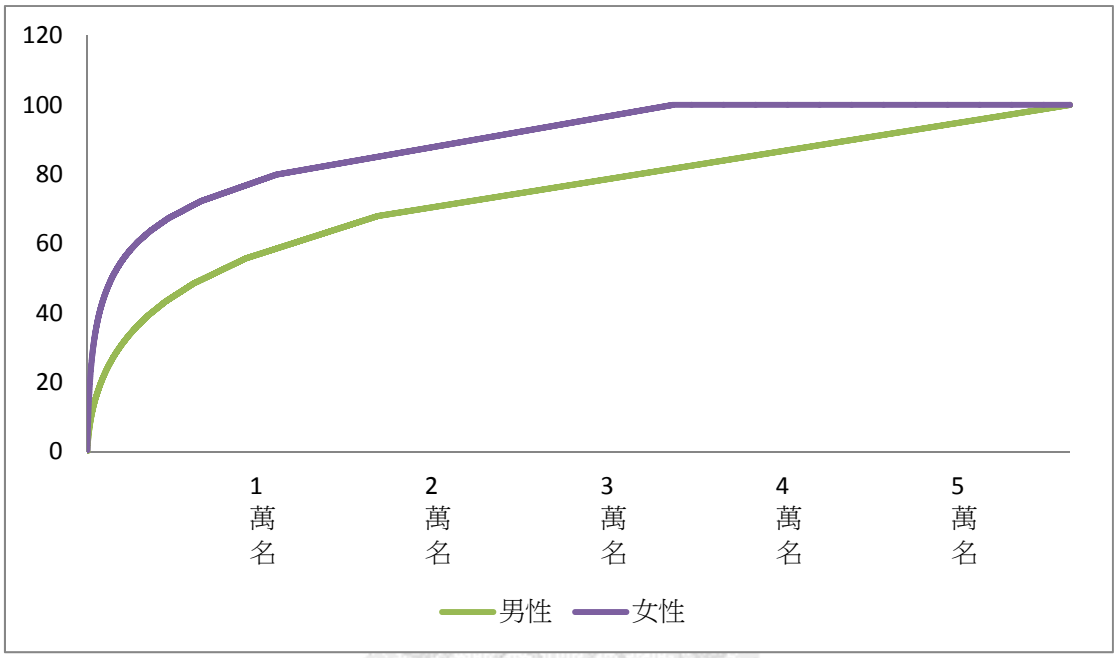


圖 4-2：兩性人名出現數與所包含的人數關係

我們將出現最多的男性名與女性名整理於表 4-4 中，並且顯示其佔總人數的十萬分筆數。從資料中我們可以明顯發現女行人名的出現次數遠大於男性，在女性人名的第 10 名仍比男性的第 1 名比例高出許多。另外就直觀而言，我們會發現這些人名的性別不難辨認，而且許多名字可能都出現在我們生活中，由此可知此資料與我們認知的生活環境類似。

表 4-4：本實驗資料中兩性出現最多的人名及其比例（十萬分比）

名次	男性名	出現比例	女性名	出現比例	中性名	出現比例
1	志明	71.56	淑芬	172.64	水金	7.38
2	志強	65.29	美玲	157.88	世英	5.90
3	文雄	63.08	淑惠	149.76	群	5.16
4	家豪	61.60	秀琴	142.02	明秋	3.69
5	俊宏	57.18	秀蘭	131.32	維華	3.69

6	俊傑	56.44	美惠	127.63	喜	2.95
7	正雄	53.86	麗華	119.52	奇	2.95
8	志豪	52.75	淑貞	117.67	根	2.95
9	金龍	52.01	雅婷	116.57	懿軒	2.95
10	文龍	51.27	秀鳳	115.60	明瑩	2.21

另外也存在兩性皆有使用的人名，如果我們以出現比例計算，例如一個人名出現在男性次數比出現在女性多時，則此人名歸類為男性，則計算得傾向男性的總人名數有 57,302，女性有 34,611，基本上與純性別（只出現在男性或女性）的人名比例接近，一樣男性人名數量高出許多。但仍然有一部分比例的人名是兩性出現數量皆相同的，圖 4-4 最後一欄「中性名」中顯示出現次數前 10 的兩性比例一樣多之人名。

由於兩性比例相同,就會造成分類器無論如何學習都有可能出現錯誤的狀況，此為本實驗分類的準確度上限。根據統計，這些名字佔全部出現名字的 1.28%，因此若在名字不重複的情況下（不管是否常見，都只出現一次），得到的系統最高的分類正確率為 98.72%。我們將之定義成為本實驗的最佳線（Topline）。

4.2 中文人名的性別預測

本節詳細比較我們所使用的中文人名性別預測工具，並分析其準確率，以找出後續實驗最有效的方法。

4.2.1 基準線實驗

在本實驗中，我們取前九萬筆資料作十折交叉驗證（10-fold cross-validation）找出最適當的 K 值，每回的準確率如表 4-5。

表 4- 5：K-最鄰近分類十折交叉驗證的結果

回數	準確度	K 值
0	91.07%	9
1	91.38%	5
2	91.38%	5
3	91.34%	3
4	91.19%	27
5	91.21%	11
6	91.61%	11
7	91.08%	11
8	91.33%	9
9	91.73%	9

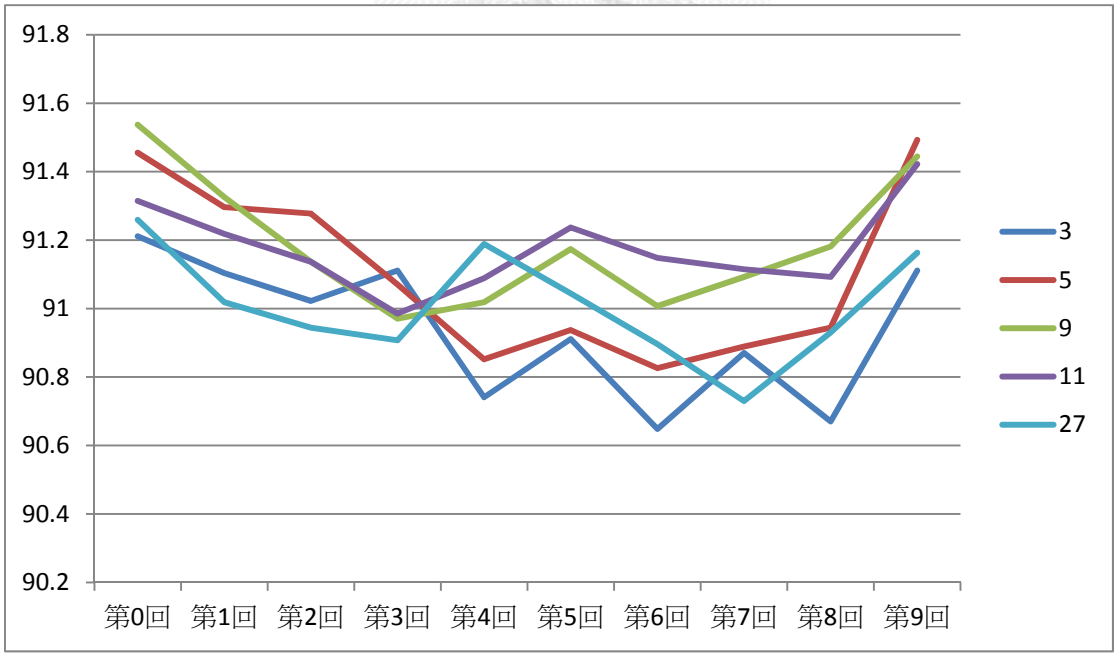


圖 4- 3：比較每回中 K 為 3, 5, 9, 11 與 27 實準確率的變化

由表 4-5 可以看出，K 的值變動很大，從 3 到 27 皆有，但是以 9 和 11 出現次數最多，平均為 10，標準差達 6.28。而準確度都在 91 到 92%之間。因此我們將 K 設為 9，用全部資料的 9/10（293,983 筆）作為參考的座標；1/10 做為測試資料，另外位的對精確度更嚴格的要求，我們將重複的人名去除，故實際上測試資料為

17,180 筆。經計算後得到的準確率為 89.6%，我們將之定義為基準線（Baseline）。

4.2.2 使用支援向量機對人名進行預測

本研究針對前文提到的支援向量機預測方法進行實驗，比較比較不同特徵表示法的特性。我們一樣以前九萬筆資料做十字交叉驗證，找出不同預測方法中支援向量機的 C 值，再以效能最佳的 C 值作為模型訓練的依據。表 4-6 比較各種方法在 C 值變化下所產生的結果：

表 4-6：以十折交叉驗證所得各種支援向量機不同 C 值的結果

回數		0	1	2	3	4	5	6	7	8	9	最佳 C
SVM-CB	最佳 C	0.75	0.25	0.25	0.35	0.25	0.25	0.4	0.35	0.35	0.75	0.395
	正確度	94.31	93.8	93.98	94.33	93.89	94.16	93.99	94.06	93.83	94.43	94.08
SVM-CJ1	最佳 C	3.0	3.5	3.5	3.5	3.0	2.75	3.0	3.0	3.0	3.0	3.125
	正確度	76.63	75.19	77.27	76.62	75.74	76.81	75.54	75.79	76.29	77.07	76.30
SVM-CJ2	最佳 C	1.25	0.5	1.0	0.5	0.4	0.75	0.25	0.5	1.5	0.75	0.74
	正確度	90.9	90.44	90.97	91.17	90.2	90.81	90.12	90.85	90.08	91.28	90.69
SVM-CJ3	最佳 C	1.25	1.25	1.25	1.25	1.25	1.5	1.25	1.25	1.0	1.5	1.275
	正確度	92.17	92.58	92.29	92.71	92.04	92.19	92.13	92.08	92.04	92.59	92.28
SVM-CJ4	最佳 C	1.25	1.5	1.25	1.25	1.0	1.0	1.25	1.0	1.0	1.25	1.175
	正確度	92.49	92.22	92.54	92.78	92.12	92.71	92.5	92.17	92.27	92.74	92.46
SVM-CJ123	最佳 C	0.1	0.15	0.1	0.15	0.15	0.25	0.15	0.15	0.15	0.1	0.145
	正確度	91.89	92.02	92.18	92.25	91.77	92.03	92.05	91.9	91.88	91.98	92.0
SVM-CJ1234	最佳 C	0.1	0.15	0.1	0.15	0.15	0.15	0.15	0.1	0.1	0.1	0.125
	正確度	91.88	92.09	92.08	92.11	91.62	92.26	91.9	91.92	91.87	92.22	92.0

根據表 4-6 的結果，我們以最佳的 C 值帶回全部的資料進行測試，方法同前，各種方法在 17,180 筆資料中，得到的準確率如表 4-3：

表 4- 7：支援向量機不同方法的準確度比較

方法	準確率	與最佳線比率
SVM-CB	92.52%	93.72%
SVM-CJ1	65.13%	65.97%
SVM-CJ2	88.44%	89.59%
SVM-CJ3	92.20%	93.4%
SVM-CJ4	**92.39%**	**93.59%**
SVM-CJ123	92.20%	93.4%
SVM-CJ1234	92.38%	93.58%

由表 4-7 我們可以發現，除了以 SVM-CB 外，準確率最高的為 SVM-CJ4 的系統，高達 92.39%，且達到最佳線的 93.59%。結果與之前推論的 SVM-CJ123、SVM-CJ1234 效果最好不同，原因可能是上述兩種方法字母可能的組合太多（維度過多），遠超過訓練資料的量，使得很多特徵處在未訓練的狀態，因此在目前資料中無法完全進行訓練，準確率無法提升。

表我們隨機選出 10 筆系統判斷性別錯誤的資料，我們可以發現這些人名即使是人工判斷都很難分辨其性別，且往往與我們的認知相反（我們的認知常常比較接近系統判斷結果），因此判定錯誤的人名應屬性別比較容易混淆或少用的。分析表格的男性比例與出現比例可以發現，僅有女性名「瑋庭」與「皓云」是出現較多次且男性比例較低的卻被系統預測成男性的，不過這些人名也包含了兩性皆可使用的字，人工判斷也同樣有其困難性。其餘如「瑤」的男性比例低，但剛好選出的資料是男性，因此這些資料在統計上應判斷為男性較合理。最後，以目前資料 27 萬筆而言，如果出現比例為十萬分之 3.69，代表該人名在整個資料集中僅出現一次，因此出現識別困難是情有可原的。

表 4- 8：隨機選取預測錯誤的人名，比例以十萬分比計算

	出現比例	男性比例	正確性別	系統判斷結果
毓璿	3.69	100%	M	F
瑋庭	73.78	25%	F	M
瑤	36.89	10%	M	F
艷川	3.69	100%	M	F
瀚儀	14.76	0%	F	M
猜	40.56	18.18%	M	F
世秦	3.69	50%	F	M
逢滿	3.69	0%	F	M
耀尹	3.69	0%	F	M
皓云	14.76	25%	F	M

我們將 SVM-CJ4 與 Baseline 系統比較，由於所預測的資料集相同（為同樣的 17,180 筆資料），因此我們採用 McNemar 檢定，辨別兩種方法的差異是否顯著，以確定我們使用的 SVM-CJ4 是否真的比 Baseline 好。在此檢定中，我們求得 P 值小於 0.0001，且 X^2（Chi squared）值為 75.956，在 95%信賴區間下為統計顯著（表 4-9），因此除了準確率數據比 Baseline 較佳之外，在統計上也有明顯的不同。

表 4- 9：以 McNemar 檢定 SVM-CJ4 與 Baseline 的預測結果

	SVM-CJ4			
K-最鄰近分類		男性	女性	總和
	男性	9290	635	9925
	女性	987	6268	7255
	總和	10277	6903	17180
P 值	小於 0.0001			
X^2 值	75.956			

第五章 系統與人工判斷性別的比較

本章透過問卷比較人類判斷性別時與電腦判斷的差異。問卷內容分為四個部分：第一部分為比較判斷常見且性別特徵明顯的名字；第二部分為比較兩性都有可能出現等較難正確判斷性別的名字；第三部分我們以隨機的方式選出人名進行比較；第四部份，我們選出非常用於人名的中文字，讓受試者對這些字組作性別標記，以檢視系統的性別模型套用在非人名中文字的可能性。

5.1 問卷設計與實驗準備

5.1.1 問卷內容

本實驗的問卷分為前述四種類型的題目：明顯性別人名、混淆性別人名、隨機挑選人名與機器產生人名，在後面的章節中將會詳細介紹這四種類型人名的產生方法。每一類別有兩個性別各 30 個名字，四組即為 240 個。另外為了評估問卷的有效性，我們選出 6 個常見且性別傾向明確的人名（志強、文雄、志豪、淑芬、美玲與淑惠）在問卷中重複出現，故整份問卷共包含 246 個人名。另外為了避免使用者作答時因為有預設立場而干擾問卷的準確度，我們以電腦將四種類型的題目亂數打散。

受試者只須對題目中的人名勾選男性或女性即可，電腦系統則以準確率最高的 SVM-CJ4 進行預測。為了讓答案方便與電腦比較，我們採用類似「投票」的方式。基於網路問卷的預設值，我們以 1 代表男性；2 代表女性。接著計算所對同一個人名所有使用者答案的平均，若小於 1.5，我們將受試者的答案視為男性，反之則為女性。詳細問卷如附件。

5.1.2 問卷發放

顧及方便性及時效性，採網路問卷與實體問卷並行。除了作答方法與發放方法不同外，描述、內容與編排順序皆相同。發放時間為民國一〇一年五月，共回收實體問卷 40 份及網路問卷 181 份。剔除作答不完全、母語非中文及重複的人名答案不一致的問卷，共得有效問卷 196 份。

5.2 問卷資料產生與實驗目的

本結分別介紹問卷四個部分人名產生的方法以及實驗目的。

5.2.1 明顯性別傾向人名實驗

本實驗要了解我們的預測系統與人工預測對明顯性別傾向的人名在預測上的差異。

我們將資料庫中所有的人名依照性別、出現次數分類，分為男性人名、女性人名與中性人名三類。男性人名指的是這個名字出現在男性中次數較多；女性人名則反之；中性人名是出現在男女次數均等。

而男性人名與女性名人中，除了人名出現次數外，我們亦另外計算性別比例。例如若某名字在資料庫中出現 3 次，其中有兩次為男性，一次為女性，則算得男性率為 66.67%。我們分別從男性與女性人名中，找出在兩性中出現次數最多且性別比率為 100%的前 30 個人名（共 60 個）進行實驗。

5.2.2 混淆性別傾向人名實驗

本實驗要了解我們的預測系統與人工預測對混淆性別的人名在預測上的差異。由於混淆性別的人名在男女性上皆會出現，因此在預測上有一定的難度。透過本實驗，我們可以了解系統預測錯誤與人工預測錯誤的內容，藉以說明系統與人類所採用的預測邏輯差異。

為了讓人名具有性別傾向，我們並非選擇「中性人名」作為題幹，而仍是從「男性人名」與「女性人名」中做選擇。我們先設定每個人名出現在資料庫中的門檻為 20 次，接著我們分別在兩個性別中找出性別比率最低的 30 個。設定出現門檻是為了避免人名太少見造成受試者的困擾，而失去了解混淆性別的人名在判斷上差異的用意。另外在篩選過後，男性人名的男性率最高為 66%；女性人名的女性率最高為 64%，均構成混淆的條件。混淆的人名與出現次數及其性別比率如表 5-1 所示。

表 5- 1：性別混淆的人名

男性混淆人名	總出現數	男性率	女性混淆人名	總出現數	女性率
春貴	63	50.79%	嘉華	60	51.67%
瑋	95	51.58%	群	48	52.08%
金連	69	52.17%	森	109	53.21%
日春	40	52.5%	定	37	54.05%
靖	67	53.73%	宇庭	37	54.05%
秀明	46	54.35%	庭安	57	54.39%
霖	42	54.76%	佳霖	91	54.95%
子嘉	36	55.56%	子儀	58	55.17%
平	106	55.66%	宥均	38	55.26%
順	58	56.9%	阿屘	66	56.06%
冠穎	68	57.35%	明芳	125	56.8%
冠樺	46	58.7%	清秀	70	57.14%

正芳	42	59.52%	英華	57	57.89%
鈞	43	60.47%	阿貴	53	58.49%
家華	89	60.67%	靖恩	39	58.97%
捷	64	60.94%	拉	37	59.46%
榮	83	61.45%	秋金	87	59.77%
水金	73	61.64%	惠群	33	60.61%
輝	45	62.22%	國英	33	60.61%
和	32	62.5%	恩	72	61.11%
吉	70	62.86%	玉清	67	61.19%
謙	38	63.16%	春華	114	61.4%
世英	41	63.41%	乃文	52	61.54%
宇	39	64.1%	文	155	61.94%
義	42	64.29%	郁文	82	62.2%
秋榮	59	64.41%	瑞芳	114	62.28%
欽	34	64.71%	佳儒	49	63.27%
庭瑋	108	64.81%	寶華	60	63.33%
嘉文	124	65.32%	子欣	33	63.64%
清	96	65.63%	錦華	114	64.04%

5.2.3 隨機挑選人名實驗

本實驗要了解我們的預測系統與人工預測對真實情況在預測性別上的差異。所謂真實情況是隨機輸入人名讓系統進行判定。本實驗除了能夠了解更真實的系統效能外，也同樣可以從預測錯誤的內容，了解系統與人類所採用的預測邏輯。

如前所述，我們的人名是使用亂數在資料庫中挑選的，男女各 30 個。由於一個人名在系統中可能重複多次，因此我們在 SVM-CJ4 訓練時會考慮到跳過這些名字，因此去除了 4,269 筆人名再作訓練。另外值得一提的是，我們為了模擬更真實的環境，在隨機挑選人名的時候只看該人名單筆資料所標上的性別。舉例來說，「秀琪」這個人名在資料庫中出現 33 次，其中有 2 次為男性，31 次為女性。若依照機率該人名應歸屬於女性，可是在我們這次實驗中，我們不考慮他出現的機率，直

接就單筆資料的性別作標準答案，而剛好選出的為男性，則將「秀琪」設為男性。這樣性別比率與答案不符的資料在本實驗中有四筆，在表 5-2 中以星號加底色標明。

表 5-2：隨機選擇的人名

男性混淆人名	總出現數	男性率	女性混淆人名	出現在女性的次數	女性率
俊男	191	100%	阿雪	86	100%
朝熙	22	100%	春美	496	100%
源峰	3	100%	津芳	1	100%
爾嘉	2	100%	玲玉	49	100%
政智	11	100%	淑媚	54	100%
發	33	84.85%	凱媛	5	100%
銘鴻	60	100%	雅如	108	100%
裕淞	2	100%	育束	1	100%
榮光	41	100%	美妹	142	100%
三郎	73	100%	淑英	219	100%
功強	1	100%	惠婷	189	100%
成茂	4	100%	秀珠	580	99.88%
志榮	164	99.39%	靜姍	3	100%
宗欽	19	100%	月美	187	100%
明清	59	91.53%	嘉茵	9	100%
金湖	4	100%	添妹	22	100%
建賢	28	100%	**永敏***	**4**	**25%**
柏驊	2	100%	千瑜	22	95.45%
國泰	69	100%	丹	62	90.32%
勝弘	16	100%	玉珍	459	99.35%
世燦	2	100%	**章***	**13**	**23.08%**
志翔	79	100%	羽倩	1	100%
俊充	1	100%	雯	91	100%
士傑	96	100%	**芳銘***	**18**	**5.56%**
秀琪*	**33**	**6.06%**	明妹	8	100%
昌廷	6	100%	杏芳	10	100%
守龍	1	100%	修夢	1	100%
安郎	2	100%	浣羽	1	100%

安倫	7	100%	淑卿	329	100%
玟德	3	100%	甚	65	98.46%

5.2.4 系統產生人名實驗

本實驗要了解系統的模型是否能套用在其他中文字上，以區分非人名中文字的性別。若系統模型套用在人類未見過的名字且人類能正確進行性別判定，則代表系統所建構之模型具有通用性。

為了讓所產生的文字不要太過於艱澀少用，我們先計算 2011 年上半年 Yahoo! 奇摩新聞中每個字的字頻，並以系統對每個單字進行性別預測，透過支援向量機求值方式預測，算得中文字的性別值。依照支援向量機的預設，求出數值越大代表越具男性象徵，反之則越具女性象徵。

我們計算每個字的頻率與性別象徵後，設定性別象徵的門檻，男性字的預測值必須大於 0.9；女性字則必須小於-1.0001。依照門檻由出現率高到低兩個性別各取 60 個字，在以亂數將同個性別中的字兩兩組和成雙名。基於此一門檻，可以使每個字在半年內至少出現 260 次，以避開冷門詞彙。表 5-3 列出兩個性別原始選取的各 60 個字及其支援向量機預測值；表 5-4 呈現隨機組合成的新人名及其支援向量機的預測值（其中「原始分數」欄位在後文說明）。

表 5- 3：從奇摩新聞中選出的具性別傾向的常見字

男性字						**女性字**					
字	**出現數**	**預測**	**字**	**出現數**	**預測**	**字**	**出現數**	**預測**	**字**	**出現數**	**預測**
局	179690	1	粽	4506	1	了	354893	-1	鴨	3389	-3
即	66788	1	劫	3928	1	者	329369	-2.18	茨	3103	-1
稱	45402	1	縱	3578	1	去	167767	-1	鍰	2893	-2.18
勞	41475	1	弗	3301	1	點	157729	-2.18	侶	2866	-2.18

聽	31916	1	釣	3125	1	此	134616	-1.12	凸	2794	-2.18
絕	26423	1	牲	2804	1	隊	105624	-1	撒	2273	-2.18
輻	24965	1	沃	2739	1	熱	55203	-1	匪	2140	-1
跑	22306	1	狼	2469	1	稅	48486	-1	糊	2087	-1
講	21077	1	猴	2368	1	射	41140	-1	艙	1840	-1.88
靠	20096	1	謠	2280	1	藥	38594	-1	蚵	1702	-3
輸	18738	1	坎	2279	1	額	38460	-1	漠	1669	-1
答	15052	1	闢	1812	1	售	36811	-2.18	汐	1537	-2.18
革	14892	1	梭	1491	1	陰	28517	-1	緋	1474	-1
霸	11610	1	瀾	1302	1	汽	18979	-1	匹	1391	-1
賠	11201	1	譚	1146	1	迷	18611	-2.18	糯	1313	-2
概	10432	1	呆	1101	1	跳	17260	-1.97	遴	1079	-2.18
趨	9976	1	敖	827	1	抵	16128	-2	穫	1044	-1
颿	9346	1	鷺	767	1	塊	14609	-1.67	檸	1031	-1
蹤	8802	1	燉	698	1	嚇	12344	-2.18	磋	999	-1.97
跡	8165	1	繆	663	1	奪	12198	-1	寂	939	-2.18
托	7693	1	豁	637	1	搬	6551	-2.18	帖	922	-3
趁	7530	1	趾	595	2	尖	6212	-1	臀	848	-2.18
牆	7121	1.35	篷	548	1	盃	5843	-1	壟	784	-1
橫	6096	1	幟	547	1.04	睛	5384	-2.18	啞	627	-1.97
緝	6058	1	唅	411	1.17	龐	4780	-1	棘	625	-1
砲	5857	1	鵡	384	1	眠	4379	-2.18	誣	415	-1.4
徑	5489	1	鷗	364	1	繞	4103	-1	噶	359	-1.72
鹿	5357	1	貳	313	1	誇	3976	-1	瀝	343	-1
玻	5288	1	鏢	309	1	斑	3937	-1	粘	310	-2.18
礦	4680	1	崐	261	1.35	蹈	3886	-2	拎	304	-2

表 5-4：系統產生的人名及其預測值

系統產生的男名	**SVM 預測值**	**原始分數**	**系統產生的女名**	**SVM 預測值**	**原始分數**
霸崐	3.3501	1.01	誣藥	-1.3993	1.12
沃唅	3.1745	1.06	陰壟	-1.0006	1.20
弗豁	3.0005	1.05	盃磋	-1.9667	1.03
即趨	2.9996	1.02	緋糊	-1.0004	1.36
革譚	3.0002	1.03	此誇	-1.119	1.04
牲礦	3.0005	1.05	糯去	-1.9999	1.06

謠瀾	2.9991	1.18	侶跳	-3.1416	1.11
粽跡	3.0002	1.02	匹穫	-1.0005	1.03
幟稱	3.0371	1.05	凸者	-3.3501	1.02
趾橫	3.9992	1.03	蹈寂	-3.1751	1.25
牆敖	3.3499	1.02	蚵額	-2.9995	1.24
闢狼	3.0004	1.01	射匪	-1.0007	1.01
燉貳	2.9996	1.07	**鍰塊**	**-2.8413**	**1.24**
颿鏢	2.9999	1.02	售汐	-3.3501	1.12
趁劫	3	1.02	**睛抵**	**-3.1746**	**1.61**
砲聽	2.9998	1.01	**了茨**	-1.0011	**1.62**
玻概	2.9999	1.12	粘遴	-3.3501	1.11
局賠	3.0001	1.05	搬稅	-2.1754	1.06
靠跑	2.9997	1.00	斑繞	-1.0004	1.19
蹤徑	2.9991	1.02	帖漠	-2.999	1.05
釣鷺	**2.9995**	**1.46**	熱棘	-1.0003	1.15
勞鹿	2.9992	1.11	臀嚇	-3.3501	1.09
篷托	2.9999	1.02	瀝噶	-1.7158	1.05
縱坎	2.9994	1.06	奪點	-2.1755	1.04
鵡繆	2.9996	1.21	**汽拎**	**-1.9996**	**1.20**
絕猴	2.9996	1.01	艙尖	-1.8783	1.02
呆答	2.9997	1.11	**龐檸**	**-1.0004**	**1.27**
緝輸	2.9995	1.08	撒鴨	-4.1743	1.05
梭輻	2.9995	1.01	隊迷	-2.1752	1.06
鷗講	2.9993	1.01	啞眠	-3.1416	1.19

（註：為了避免數值過度接近，本資料取到小數點下四位）

5.3 實驗結果與討論

本結討論實驗結果。

5.3.1 明顯性別傾向人名實驗結果

根據表 5-5 我們可以發現系統與人類的答案完全相同。剖析問卷內容，會發現

在這組實驗中的某些受試者對於部分的人名仍有少數的人會誤判，不過所有男性的平均值都低於 1.03，而女性的平均值均大於 1.96。對於性別傾向明顯的人名，人工判斷的準確率極高，但系統也表現極為良好，由此可知人名中倉頡碼必然有存在性別的規則，且人與系統是極有共識的。

表 5- 5：明顯性別傾向人名在不同方法下的準確率

	男性填答者	**女性填答者**	**所有填答者**	**倉頡 4-grams**
男姓	100%	100%	100%	100%
女性	100%	100%	100%	100%
平均	100%	100%	100%	100%

5.3.2 混淆性別傾向人名實驗結果

根據表 5-6 有個有趣的發現，男性在猜測男性人名時，正確性高於女性；而相對的女性也比男性更容易猜出女性的人名。不過若從另一個角度解釋，也可以推測為當人們在遇到難以判定性別的名字時，會先以自己的性別作為優先考量。另外還有一個很特別的現象是當把所有使用者的答案一起投票時，女性的準確率 46.67%低於兩性猜測準確率的平均。此一現象可能是女性在猜測女性人名時有較高的歧異性，使得投票平均很接近 1.5 的門檻，因此容易被男性受試者的投票結果影響，而偏向於錯誤的答案。另外在系統預測上，會發現雖然預測男性的準確率不如男性受試者及全部受試者高，但平均而言卻表現較好。

表 5- 6：性別混淆的人名實驗在不同方法下的準確率

	男性填答者	**女性填答者**	**所有填答者**	**倉頡 4-grams**
男姓	**90%**	50%	80%	76.67%
女性	43.33%	**80%**	46.67%	**80%**
平均	66.67%	66.67%	63.33%	**68.33%**

在顯著程度方面，我們借助 McNemar 檢定，比較系統預測與人工預測的答案，計算所得 P 值為 0.4042，X^2（Chi squared）值為 0.696，在 95%信賴區間下為統計不顯著（表 4-5）。由此可知，系統是以「人類的方法」進行預測，因此與人類判讀在統計上效能是相同的。

表 5- 7：混淆人名以 McNemar 檢定系統與人工預測的結果

	系統預測			
人工預測		男性	女性	總和
	男性	26	14	40
	女性	9	11	20
	總和	35	25	60
P 值	0.4042			
X^2 值	0.696			

5.3.3 隨機挑選人名實驗結果

由表 5-8 中我們可以發現若就平均正確性而言，本系統的表現比人工預測還為準確。其中人工預測以男性正確率最高，女性與所有受試者平均答案相同。討論資料內容，可以發現男性人名都只錯了一筆資料—秀琪，就是男性率最低的男性名。而以人工平均而言，女性錯了「育東」、「丹」、「章」、「芳銘」、「修夢」與「甚」等六筆，其中「章」與「芳銘」同屬女性率較低者，而「育東」與「修夢」在本系統中僅出現一次，故為較少見的人名，對人工區分而言可能較為困難。另外「丹」與「甚」都出現 60 多次，女性率也都在 90%以上，可是人工判斷時出現錯誤，就直觀而言這兩個人名的確比較難判斷，原因可能是單名對國人而言在生活中還是相對少見的，若遇到不常見的字，較容易與古人聯想，而史書中多以男性為主角，因此古人比較容易聯想到男性而作男性人名的判斷。

表 5- 8：隨機選取的人名實驗在不同方法下的準確率

	男性填答者	女性填答者	所有填答者	倉頡 4-grams
男姓	96.67%	96.67%	96.67%	96.67%
女性	83.33%	80%	80%	**86.67%**
平均	90%	88.33%	88.33%	**91.67%**

在顯著程度方面，我們一樣以 McNemar 進行檢定，計算所得 P 值為 0.6171，X^2 值為 0.250，在 95%信賴區間下為統計不顯著（表 5-9）。

表 5- 9：混淆人名以 McNemar 檢定系統與人工預測的結果

	系統預測			
人工預測		男性	女性	總和
	男性	32	3	35
	女性	1	24	25
	總和	33	27	60
P 值	0.6171			
X^2 值	0.250			

本實驗的結果與混淆人名相似，系統的正確性比人工預測準確，且經過檢定發現不顯著，代表與人工的邏輯類似。實際上，這次系統與人工的相似度比模糊性別的人名來得更高，在女性的部分除了「永敏」外，其他系統預測錯誤的人名—「章」、「芳銘」與「修夢」都是人工也預測錯誤的。若以「男性率」與「女性率」為標準答案，則系統在男性的準確率可以提升為 100%，女性也可以達到 96.67%。

5.3.4 系統產生人名實驗結果

由於我們是先以 SVM-CJ4 的預測值作為挑選常見字的門檻，因此這組實驗在 SVM-CJ4 下預測準確率必然為 100%。不過我們在此是想藉由本實驗了解人供判別時是否會因為某些字的特徵而真的能分辨出男女性別的用字。

在人工預測方面，我們從表 5-10 可以發現男性受試者、女性受試者與所有受試者其實存在著高度的一致性。細部討論會發現，男性的受試者在猜測男性人名時錯了一題—「釣鷺」，但女性受試者卻能正確的將該名判斷為男性；而所有猜測女性人名的結果均慘不忍睹，僅對了「睛抵」、「了茨」兩個，其餘人名均被判別為男性。這是一個有趣的發現，我們可以推斷當人們在遇到未見過的人名，會先以「男性」作為推論，這也許跟第三節中描述男性的人名數量遠多於女性，且僅出現一次的人名也以男性居多有關。可能是基於這樣的經驗，使得人們在遇到沒見過的人名時不會考慮字的結構，而是以經驗法則先判斷為男性有關。

表 5- 10：系統產生的人名實驗在不同方法下的準確率

	男性填答者	女性填答者	所有填答者	SVM-CJ4
男姓	96.67%	100%	100%	100%
女性	6.67%	6.67%	6.67%	100%
平均	51.67%	53.33%	53.33%	100%

對於非用於人名的中文字，我們很難區分他們所代表的性別含意。即使一個字具備有某些特定性別的特徵。例如以「艸」為部首的字常出現在女性人名中，但是「藥」卻難以與女性聯想。不過我們仍會發現，若以投票的平均分數計算，所算得的男性人名分數為 1.06，而女性人名算得的分數為 1.15，的確會有所不同，若將性別門檻降低到 1.15，會發現女性人名納進了「龐檸（包含「寧」的部件）」、「汽拎（包含「令」的部件）」、「斑繞（包含「玉、糸」等偏旁）」、「鍰塊（包含「爰」的部件）」等具有女性名常見特徵的字，且整體正確力提升到 65%，表 4-10「原始分數」欄位即為受試者投票的平均結果。由此可知，字型特徵仍具有性別傾向的作用。未來進行此實驗時，可以考慮取出出現次數較少的人名讓受試者判讀，以避免受試者主觀認為這些名字不會出現在人名中而出現判斷的困難。

第六章 以真實名稱判定性別傾向

本章將人名與性別的實驗擴大到不同的領域中，首先先以相似的資料集—臉書中的人名信行測試。接著我們試圖了解中文譯名中在選字上是否仍具有性別傾向，透過系統判讀英文譯名的性別。最後的三個實驗已跨出「人名」的範圍，將問題推展到商家與股票的名稱，分析不同產品、產業別所具備的性別傾向。

在實驗中我們仍以 SCM-CJ4 為主，並搭配效能最差的 SCM-CJ1 及字元為基礎的 SCM-CB，試圖了解不同方法的預測差異。由於倉頡編碼具有拆解中文部件的特性，因此若遇到沒有訓練過的字，仍然能進行預測；但以字元為基礎的方法無法對沒訓練過的字進行預測，這也是本節要討論的第二個重點。

6.1 資料蒐集與實驗目的

本結說明實驗目的，並分別介紹臉書使用者、英文譯名人名產生的方法以及本地商家、個股名稱的資料來源。

6.1.1 臉書使用者的人名與性別關係實驗

在本實驗中，我們對臉書中的好友名稱進行預測，以了解在不同資料集中人名預測的結果。

我們透過臉書官方所提供的Facebook Graphic API對研究者的臉書朋友列表抓取資料。在欄位的選擇上，我們只抓取性別與名稱。由於臉書的設計性別非必填欄位，且許多使用者以英文命名，因此在篩選過後共得 322 筆資料，其中男女比

例為 149.61 比 100。

6.1.2 英文譯名與性別關係實驗

英文的人名具有明確的性別傾向，而在台灣，英文譯名也大都有固定的翻譯方法。本實驗想了解若以目前的文字性別規則，是否能正確的分辨出英文譯名的性別傾向，藉以作為判讀譯名時的參考依據。

我們先從國際知名的愛斯英語學習網（http://www.24en.com/）中英文取名及意義的網頁抓取英文人名的中文翻譯，共得 889 筆資料，男女比例為 128.2 比 100。

6.1.3 網路拍買男女服裝與店家名稱性別傾向實驗

網路服飾商店通常都以「男性」、「女性」作為商品類型最大區別，其下再分為衣服、褲子、鞋子、配件…等許多子類別。本實驗收集 Yahoo!奇摩超級商城（http://tw.mall.yahoo.com/）與 PCHome 商店街（http://www.pcstore.com.tw/）中服裝店名稱，根據報導（蘇文彬，2010）指出，這兩個網路商城為目前台灣最大的商家匯流處，Yahoo!總共有 1,200 家以上，而 PCHome 則共有 8,600 家。

我們選擇服飾類別，依照男裝女裝分類，並以熱門度由高到低排序，各取出前 60 家熱門的店，去除名稱全為英文者，共得 105 家，男女比例為 81 比 100。

6.1.4 台灣商家名稱的男女性別傾向實驗

本實驗使用台灣商家的名稱作為測試資料，試圖了解中文性別傾向是否會因為行業別而有不同的比例。根據先前研究（Cassidy, Kelly & Sharoni, 1999）指出，

在運動的雜誌中所提到的品牌多偏向男性；時裝雜誌品牌偏向女性，因此如果一個能夠區分性別傾向的模型，應該可以在不同類型的行業中發掘潛在的命名特性。

為了取得台灣本地的商家資料，我們整合奇摩生活+（http://tw.ipeen.lifestyle.yahoo.net/）與愛評網（http://www.ipeen.com.tw/）的店家名稱資料與評價人數，並以電話號碼將重複的店家整合。另外我們參考奇摩生活+的分類，將本地商家為六個類別：美食餐廳、飯店住宿、百貨購物、娛樂運動設施、交通相關與專業服務（表 6-1）。

店家選擇的部分，一樣限制店家名稱必須完全為中文，另外我們手動將名稱中的地名、商品名、行業名及商家慣用字去除，如去除九份、燒烤、餐廳、館…等等。最後我們根據評價數量判定店的熱門度，評價越多的店我們視為越熱門，找出每個類別最熱門的前 30 家店。

表 6-1：台灣商家的六種類別

類別	**說明**
美食餐廳	所有以「吃」為主的店家，如餐廳、地方美食、異國料理或小吃…等等
飯店住宿	提供「住」為主的服務，如飯店、民宿、賓館與汽車旅館…等等
百貨購物	販賣「實物」以賺取利潤的店家，如服裝店、便利商店、超級市場、電子用品…等等
娛樂運動設施	提供「娛樂」或「運動」以賺取利潤之店家，如主題樂園、遊樂場、撞球間、網咖或游泳池、健身中心…等等
交通相關	提供「行」為主的服務，如客運、計程車隊、租車公司、火車站…等等
專業服務	是以「服務」為主的商家，隨然上述類別中有不少也提供服務，但這裡可以解讀為「其他的服務業」，如診所、補習班、房仲業…等等

6.1.5 台灣個股股名稱的男女性別傾向實驗

本實驗測搜集台灣目前上市公司的名單，並依照台灣證券交易所設定之類別分類，試圖利用既有的中文性別規則，預測不同公司名稱的性別，並計算每隻類股公司名稱的男女比例，發掘不同類股的性別傾向。本實驗與前一實驗概念相似，不過測驗對象為企業集團的名稱。

由於坊間類股分類的方法無衷一是，故本研究採用較具權威性的台灣證券交易所《大盤、各產業類股及上市股票本益比、殖利率及股價淨值比月報》所訂之分類作為類股分類的標準。另外為了避免資料過少的偏頗情形，我們僅針對各類股內包含超過 10 家公司者，並根據公司治理資料取得類股中公司的全名，再從全名中去除與公司名稱無關的辭彙，如「水泥」、「電纜」…等等，減少干擾的情形。另外我們也會將名稱中的地名去除，如「台灣」、「台南」等，故類股內包含的公司數量會比原始資料來的要少。

6.2 實驗結果與討論

本結討論實驗結果。

6.2.1 臉書使用者的人名與性別關係實驗結果

我們以效能最佳的 SVM-CJ4 作為本系統的預測。由表 6-2 可知，如同先前的實驗，即使男性的資料數量較大，但我們在預測男性時還是可以得到較高的正確率，且平均正確率也在 93%左右，與先前訓練階段所使用 SVM-CJ4 的測試資料正確率相近。在另一方面，我們比較不同支援向量機的方法(表 4-13)，發現 SVM-CJ4 與以 SVM-CB 的方式表現最理想，再來是 SVM-CJ3 與 SVM-CJ1234，而效果最差

的一樣是 SVM-CJ1。

表 6- 2：以倉頡 4-grams 預測臉書好友性別

	SVM-CJ4 結果				
性別答案		男性	女性	總和	正確率
	男性	186	7	193	96.37%
	女性	15	114	129	88.37%
	總和	201	121	322	93.17%

表 6- 3：以不同方法對臉書人名進行預測的準確率

方法	**正確率**
SVM-CJ4、SVM-CB	93.17%
SVM-CJ3、SVM-CJ1234	92.55%
SVM-CJ123	91.61%
SVM-CJ2	85.71%
SVM-CJ1	64.29%

我們從結果中選出幾組比較有代表性預測錯誤的資料，比較 SVM-CB、SVM-CJ4 與 SVM-CJ1 的預測差異（表 6-4）。從下表中我們可以發現，以 SVM-CB 與 SVM-CJ4 判斷性別時錯誤的地方幾乎一樣，而仔細檢視這些錯誤，會發現其實以人工判定也很難正確分辨。此外臉書的資料除了一般的真實人名外，還會包含許多的暱稱，如：男性中使用的「大皓」、「小宇」、「燕少」…；女性中使用的「哆哆」、「杯子」、「豬仔」、「太陽」、「妞妞」…等，都會影響判斷的正確性。值得一提的是在暱稱的使用上，此樣本中女性使用的比例遠高於男性，男比女為 12 比 22，因此女性名稱與訓練資料集理應差異較大，解釋了無法正確預測的原因。

表 6- 4：比較不同方法中預測錯誤的資料內容

人名	**真實性別**	**SVM-CB**	**SVM-CJ1**	**SVM-CJ4**
		93.17%	64.28%	93.17%
岳穎	F	M	M	M

雯澤	M	F	M	F
小隻	M	F	M	F
紫蘇	M	F	F	F
太陽	F	M	M	M
小麥	M	F	M	F
孟芳	F	M	F	M
語辰	F	M	F	M
銘祥	M	M	M	M
大關	F	M	F	M

6.2.2 英文譯名與性別的關係實驗結果

我們一樣以效能最佳的 SVM-CJ4 作為本系統的預測結果，如表 6-5。實驗結果如同預期，雖然預測能力不如先前的實驗，不過都能在男性人名上有較好的表現，而平均正確率也達到 82.9%，已有相當的鑑別能力。

表 6- 5：以 SVM-CJ4 預測英文譯名的性別

	SVM-CJ4				
性別答案		男性	女性	總和	正確率
	男性	437	62	499	87.57%
	女性	300	90	390	76.92%
	總和	737	152	889	82.9%

比較不同支援向量機的準確率（表 6-6），其中以 SVM-CJ4，高於 SVM-CB。我們認為可能是在英文譯名中有較多不易出現在中文人名中的用字，使得 SVM-CB 的準確率無法發揮。這也是本實驗一大價值所在，透過倉頡部件選擇的特性，遇到即使是未訓練過的中文字，一樣可以從文字的結構推斷字的性別。另外本實驗中我們發現與之前結果不同的是 SVM-CJ3 與 SVM-CJ2 準確率降低的更快，且與 SVM-CJ1 的差異越來越小。可能的原因是翻譯會固定使用某些特定的字，而且這些字的筆劃較簡單，所以在連詞（grams）重複數低的時候，會得到許多相

同的向量，而較不具性別傾向的鑑別性。

表 6- 6：不同支援向量機特徵表示對英文譯名正確性的比較

方法	**正確率**
SVM-CJ4	82.9%
SVM-CB、SVM-CJ1234	81.78%
SVM-CJ123	81.21%
SVM-CJ3	74.13%
SVM-CJ2	73.45%
SVM-CJ1	66.93%

我們一樣從結果中選出幾組比較有代表性的資料，比較以 SVM-CB、SVM-CJ4 與 SVM-CJ1 的預測差異（表 6-7），我們將其他方法預測錯誤，唯獨某個方法預測正確的欄位標上底色。從下表中我們可以發現，若就單字的角度，許多英文人名的用字並非會依照中文人名的規則。這可能跟專有名詞的翻譯規則有關，例如母音 a 常翻譯成「安」或「阿」、子音 b 常翻譯成「伯」或「布」、子音 d 常翻譯為「德」、「達」或「大」、子音 t 常翻譯為「塔」、「坦」或「特」…其中「伯」、「德」或「達」在中文人名中男性意味較濃；「安」女性意味較濃，但在譯名中是男女通用的；而「塔」、「布」等字在訓練資料中也較少見，均會降低判斷正確性。

表 6- 7：比較不同支援向量機方法英文譯名的預測結果

譯名	性別	SVM-CB	SVM-CJ1	SVM-CJ4
		81.78%	66.93%	82.9%
艾倫(Allen)	M	F	**M**	F
班(Ben)	M	F	F	**M**
布萊恩(Brian)	M	F	**M**	F
凱撒(Caesar)	M	**M**	F	F
愛德華(Edward)	M	F	F	F
安琪拉(Angela)	F	M	**F**	M
卡蘿(Carol)	F	M	M	M
多拉(Dora)	F	M	M	M

弗羅拉(Flora)	F	M	M	M
珍尼佛(Jennifer)	F	M	**F**	M

另一方面，先前研究指出，英語系人名女性比男性變化高。但此測試資料中，女性資料筆數反而較少，無法驗證變化高的部分，且所得結果與先前中文人名的性別預測相近，男性的正確率較女性高。

6.2.3 網路拍買男女服裝與店家名稱性別傾向實驗結果

我們以 SVM-CJ4 作為本系統的預測結果，如表 6-8。實驗結果也可以發現，雖然比先前人名相關的實驗準確率更低，但仍然「男女有別」，平均正確率達 70.48%，可表示不同性別為主的服飾店名稱會帶有不同的性別傾向。此外，男女性名稱的正確率差異不大，可以帶看出系統在店家名稱的穩定度不錯。

表 6- 8：以 SVM-CJ4 預測網拍男女服裝店家名稱的性別

	SVM-CJ4				
性別答案		男性	女性	總和	正確率
	男性	34	13	47	72.34%
	女性	18	40	58	68.97%
	總和	52	53	105	70.48%

比較不同支援向量機的準確率（表 6-9），與先前實驗結果差異最大的是這次的第一名是 SVM-CJ1234，而且以往表現較佳的 SVM-CB 也輸給 SVM-CJ3，最重要的是 SVM-CJ1 以往與前項差異都非常大，但這次卻有較接近的正確率。在此可能是店家名稱與人名的差異更大，且因為店名的長度普遍較人名長，因此就算用正確度較差的 SVM-CJ1 也因為可以得到比較多特徵而可能產生正確性較高的結果。

表 6- 9：不同支援向量機特徵表示對英文譯名正確性的比較

方法	**正確率**
SVM-CJ1234	71.42%
SVM-CJ4	70.48%
SVM-CJ3	69.52%
SVM-CB、SVM-CJ123	65.71%
SVM-CJ2	63.8%
SVM-CJ1	61.9%

為了對資料能進一步的了解，我們特別選出 SVM-CJ4 分類錯誤的店家名稱並與 SVM-CB 及 SVM-CJ1 做比較。從欄位左側店家名稱之處可以發現，店家名稱也使用與平常熟知的人名差異較大的詞彙，而且也有長度較長的趨勢。在選出來的店名中，剛好 SVM-CB 與 SVM-CJ4 預測錯的地方是相同的，而 SVM-CJ1 則又出現與先前實驗相似不一樣的猜測方向。

表 6- 10：比較不同支援向量機方法店家名稱的預測結果

店家名稱	性別	SVM-CB	SVM-CJ1	SVM-CJ4
鎖螺絲	M	F	**F**	F
酷樂	M	F	**F**	F
極品	M	F	M	F
体匠	M	F	M	F
衣酷	M	F	**F**	F
天后	F	M	M	M
四葉幸運草	F	M	**F**	M
白奇	F	M	M	M
伊勢	F	M	**F**	M
名模衣櫃	F	M	**F**	M

6.2.4 台灣商家名稱的男女性別傾向實驗結果

我們以 SVM-CJ4 的方法對店的名稱作預測，在此由於要計算的是店家類型的

性別傾向，故我們先將所有的答案稱設為「女性」，再計算有多少的店能夠準確預測（被預測成女性），結果如圖 5-1：

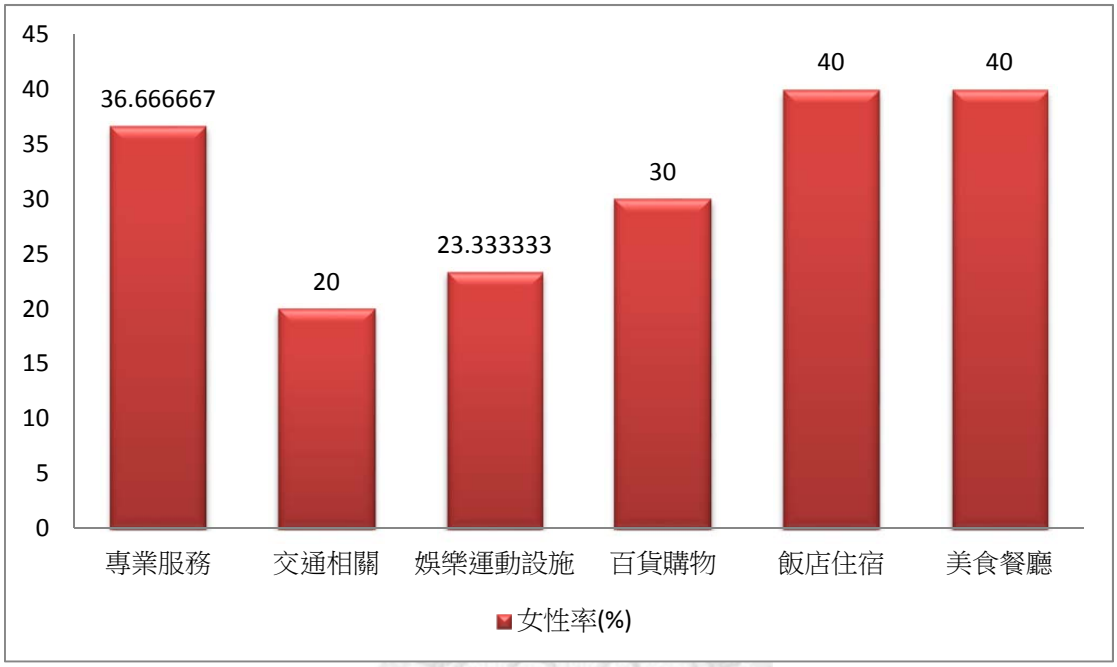


圖 6- 1：台灣本地商家不同行業別的性別比例。

根據上圖可以發現所有行業的性別比例都低於 50%，因此都偏向男性。不過其中的美食餐廳與飯店住宿類別的女性比例較高；而交通相關的男性比例最高，後文針對這三種類型進行分析。

6.2.3.1 美食餐廳結果分析

美食餐廳選用偏向女向的詞彙可能是因為女性字可以會讓人有「可口」的感覺，也會比較柔和，使用餐的感覺較為舒緩浪漫。表 6-8 中我們比較 SVM-CB、SVM-CJ1 與 SVM-CJ4 三種方法在美食餐廳類別中的預測值，並挑選幾個比較具有代表性的店名作為比較。我們可以發現在 SVM-CB 的女性率較高，可能是因為該法只要遇到沒看過的字，支援向量機皆會預測出-0.470519 的數值，使答案結果為女性，因此也有可能是美食餐廳存在較多未訓練過的字。而大部分情況而言，使用 SVM-CJ4 可以得到與 SVM-CB 類似的結果。值得一提的是，由於台灣受外來文

化影響，在餐廳的種類也比較多元，許多異國料理的店家會使用類似譯名的方式命名，也有可能降低猜中女性的機率。

表 6- 11：以不同方法比較美食餐廳類別名稱的差異

名稱	**SVM-CB**	**SVM-CJ1**	**SVM-CJ4**
女性率	**43.33%**	**43.33%**	**40%**
吃飽無罪	F	M	F
依蕾特	F	F	F
油庫口	F	M	M
米朗琪	M	F	M
樂子	F	F	F
蜜糖	F	M	F
花月嵐	F	F	F

6.2.3.2 飯店住宿結果分析

飯店住宿類別中，由於包含汽車旅館，且汽車旅館偏好較浪漫柔和的名稱，故多女性。我們將不同方法預測的幾個代表的名字整理於表 6-9。由表中我們可以發現雖然「老爺」在語意上我們認知偏向男性，但此兩字在訓練資料中並無出現。單測「爺」字，只有在 SVM-CJ1 時會判定為男性，其他的方法皆認定為女性，即便在 SVM-CJ4 中也是，故得出出乎預期的結果。另外「春天」在語意認知上偏向女性，但原因可能是人們會將注意力集中在偏向女性的「春」而自動忽略了男性的「天」，但電腦卻是將兩個字視為同等重要，經兩個字平衡後得到男性較高的分數。

表 6- 12：以不同方法比較飯店住宿類別名稱的差異

名稱	**以字元為基礎**	**倉頡 uni-gram**	**倉頡 4-grams**
女性率	**36.67%**	**63.33%**	**40%**
薇閣	F	F	F
八方美學	F	F	F
夏天	M	F	M

老爺	F	M	F
悅豪	M	M	F
春天	M	M	M
鶴雅	F	F	F

6.2.3.3 交通相關結果分析

交通相關類別男性比例最高，我們將不同方法預測的幾個代表的名稱整理於表 6-10 中。根據分析，交通相關類別中以租車公司與計程車業行佔大多數，實際搜尋發現這些車行的經營者多為老一輩台灣人，在命名上本土氣息濃厚；另外除了使用類似人名的名稱外，也常將發財的觀念帶入，如「好客來」、「一路發」等。這些發財的概念在在先前的研究（廖恭鳳，1991）也指出多出現在男性中。而另外我們也發現，較新成立的公司有較高的比例使用偏向女性帶有「關懷」意味的名稱，因此可以推斷產業的性別比例會隨著時代有所調整。

表 6- 13：以不同方法比較交通相關類別名稱的差異

名稱	**SVM-CB**	**SVM-CJ1**	**SVM-CJ4**
女性率	**23.33%**	**40%**	**20%**
隆韻(1999 年創)	F	M	F
中美(2005 年創)	F	F	F
好客來(1978 年創)	F	F	M
一路發(1989 年創)	M	F	M
和欣(1999 年創)	M	M	M
國美(20 年歷史)	F	M	F
泛亞(2003 年創)	F	M	F

6.2.5 台灣各類股名稱的男女性別傾向實驗結果

從圖 6-1 中可以看出，觀光類股的女性率為全部類股中最高者，與本地店家不同的是，除了女性率偏低以外（超過 20%的只有一隻，而本地店家最低卻為 20%），有幾隻類股女性率還是 0%（金融、塑膠、電子通路、電機），即是該類股內所有

的公司名稱都被預測為男性；而類股中女性率最高的是觀光類股，再來是光電與電腦。

細部而言，「觀光」的女性率為 27.27%，是全類股中最高的最高，另外較高的是「光電」的 16.92%、「電腦」的 13.46%與「鋼鐵」的 12%。而化學、半導體、生技、食品都在 10%左右。而如前所述，「金融」、「塑膠」、「電子通路」與「電機」的女性率為 0%。根據表中的數據交叉比對，我們可以發現類股中個股數與女性率間沒有明顯的關係，例如觀光類有 11 隻個股，女性率第二高的光電卻有 65 隻。不過一個有趣的觀察是在女性率為 0%的類股，除了電機類外，其餘的個股數都約在 20 隻左右。

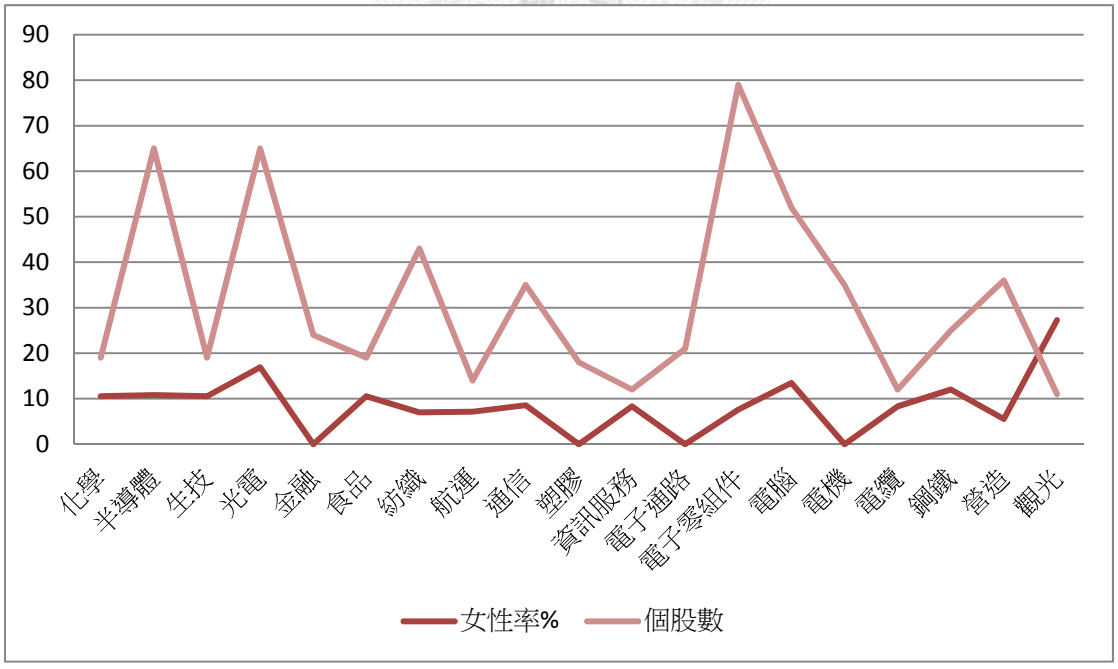


圖 6-2：台灣公司名稱在各類股中的性別比例

為了更了解各類股的細部特性，我們挑選女性率最高的觀光類與光電類；另外也找到女性率為 0%且各股數最多的電機類進行分析。

6.2.4.1 觀光類股分析

觀光類個股數為 11 隻（表 6-11），在其他類股中相對較少。我們將預測為女性的三隻個股以粗體標示，從表中可以發現，偏向女性的名稱與人名的型態較為接近，而偏向男性的名稱有部分的字是非訓練資料的，因此較有可能出現男性的預測結果。

表 6- 14：以不同方法比較觀光類股名稱的差異

名稱	SVM-CB	SVM-CJ1	SVM-CJ4
女性率	**27.27%**	**54.55%**	**27.27%**
萬企	M	M	M
華園	**F**	**F**	**F**
國賓	M	M	M
六福	M	M	M
第一華僑	M	F	M
晶華	**F**	**F**	**F**
夏都	M	M	M
美食達人	M	F	M
王品	M	M	M
鳳凰	**F**	**F**	**F**
新天地	M	F	M

6.2.4.2 光電類股分析

光電類個股有 65 隻，我們在表 2-25 中挑選較具有代表性的 10 隻個股進行分析。根據實際觀察，我們發現光電類股的個股名稱與人名較為相似，而且可能與產業內容有關，名稱中常帶有「晶」字，造成相似於女性人名。以 SVM-CB 與 SVM-CJ4 的差異不大，且從主觀的角度認為，SVM-CJ4 將「璨圓」判斷為男性也具有合理性，另外 SVM-CJ1 的方法作出的判斷與人類認知上存在有較大的差異。

表 6- 15：以不同方法比較光電類股名稱的差異

名稱	SVM-CB	SVM-CJ1	SVM-CJ4

女性率	**15.38%**	**54.55%**	**16.92%**
華晶	F	M	F
璨圓	F	M	M
昱晶	F	M	F
晶彩	F	M	F
奇美	F	F	F
友達	M	M	M
銨德	M	M	M
明基	M	F	M
昇陽	M	F	M
佳能	M	F	M

6.2.4.2 電機類股分析

電機類股為目前女性率最低且各股數最多的類股，我們一樣以表 6-13 呈現電機類股中 10 隻代表性的個股名稱。我們特地挑選了幾隻名稱與人名差異較大的個股，以粗體標示。我們可以發現與人名差異小的個股在用字上也較偏向於男性，而與人名差異較大的個股則在判斷上也多被歸為男性。

表 6- 16：以不同方法比較電機類股名稱的差異

名稱	**SVM-CB**	**SVM-CJ1**	**SVM-CJ4**
女性率	**0%**	**25.71%**	**0%**
士林	M	M	M
東元	M	M	M
中興	M	M	M
堤維西	**M**	**F**	**M**
高林	M	M	M
車王	**M**	**M**	**M**
上銀	**M**	**M**	**M**
為升	**M**	**F**	**M**
帝寶	**M**	**M**	**M**
羅昇	M	F	M

經由本實驗，我們可以發現在不同類股中性別的比例的確存在差異，而個股

的名稱比起本地商家更接近人名。另外我們發現女性比例高的類股除了光電業喜歡用「晶」字外，帶有「華」字的也多被判斷為女性，可能跟產業的型態有密切的關係。

第七章 結論與建議

7.1 實驗結論與建議

從上文中諸多實驗中可以發現,中文字的部件的確具有代表性別特性的意義，因此透過充足的學習可以讓電腦系統從中找出規則，進而達成性別的預測。此外，在比較以人工方式預測與系統預測的結果後，也可以發現系統的預測是以「人」的方式進行，系統判斷錯誤之處人工通常也存在錯誤，而且從數據顯示在人名領域的性別判定能表現得比大多數的人更好。但另一個有趣的發現是，人們在從人名中判別性別時，字的結構未必會是第一考量。優先考量的通常是「這名字有沒有看過」或是「這名字有沒有部分的字看過」,實際上比較接近「以字元為基礎的判別」或 K-最鄰近法的方式，必須要在看過的名字時才會有比較好的判斷，而字的形狀對人而言雖然存在影響力，但不是最直覺的判斷條件。此外，在中文人名中存在的特性是男性的名字歧異度較高，因此不管是人工預測或系統的預測，在遇到結構陌生的文字時判斷成男性的機率均較大，此一現象也滿足現實的文字分布。

在以現有模型對非人名的字進行判定時，的確可以發現不同產業性別有不同的傾向。但如前所述，許多商家的名字中慢慢參雜類似翻譯的詞彙，女性的用字也隨著時代年代推進越來越普及，這些不只會改變產業中名稱的性別比例，更會使得單從「人名判斷性別」所訓練出的模型效能降低，因為翻譯有固定的詞彙，很多時候是不分男女的，但是在電腦系統中每個字的權重相等，無法如同人類進行選擇性注意而找到隱含的性別關鍵字。而在解釋後可以發現，若一個商家或產業以必須透過服務完成交易時，則命名成偏向女性的比例則越大。

在建議部分，可以考慮以下方式改進實驗結果：

7.1.1 考量選擇性注意問題

可能透過 TF-IDF 等方式對訓練資料中的人名計算字的權重，以找出較具影響力的字，進而在搭配倉頡結構判斷性別。此法其實與「以為字為基礎」的方式相似，不過若就特定少數字的權重進行調整，例如「春天」的「春」字權重增加，也許會得到更合乎人類邏輯的預測。

7.1.2 考慮字的位置

目前本研究的作法是將人名的兩個字打散在同一個向量空間中，一起進行支援向量機的計算，但我們知道，中文人名中第一個字常帶有家族輩分，因此與性別較無關係。若我們在訓練模型時對文字的位置也加以考量，例如只訓練最後一個字或加上位置的標注，也許在測試階段可以得到較高的準確率，套用到其他領域也能得到較接近真實的答案。

7.1.3 考慮以倉頡詳碼對文字編碼

由於倉頡碼設計時為了輸入方便，字碼上限為五碼，致使相同構造的部件可能因為和在筆畫複雜的字而產生簡化，使得系統無法正確判讀出所有的部件。如果能先將文字轉換為倉頡詳碼在進行學習，系統必能學習到更多的部件特徵，提升準確率與解釋能力。

7.1.4 結合不同的預測方法提升準確度

目前我們以「SVM-CB」及「SVM-CJ4」在效能展現上最由優異，但不同的方法各有所長，我們也發現「SVM-CJ1」的預測結果與其他方法差異甚大，因此若我們能將以上的方法結合起來，用類似「投票」或搭配回歸等方法訂定每種方法的權重，再來決定性別，則可能會出現比目前更高的準確率。

7.2 研究貢獻

從中文字部件判斷其中所蘊含的性別特徵是本研究最重要的發現，且根據此方式以支援向量機學習後的預測系統，可以達到近似或超越真人的預測能力，可以協助人們自動解決性別判定的問題，也將會是未來中文自然語言處理重要的一環。另外，本研究也試圖用既有模型處理非人名中文字的問題。我們透過系統選出具有性別特性的文字，在人工判斷下，也與部分受試者所作的推測相同，進而能印證模型的確具有產生具特定性別傾向字型的特性。再者，我們也發現即使是英文的譯名,翻譯者在翻譯時還是會依照原文的性別傾向選用具有性別傾向的詞。最後，我們將模型套用到不同的商品、商家類型與類股中，發現在不同產業中的確存在著不同的性別傾向。基於以上的研究成果，本研究可以提供四種方向的應用貢獻：

7.2.1 建立全自動的性別判斷系統

對於性別判斷，最直觀的貢獻是建立從人名判斷性別的系統。由於本實驗在正確性上超過人類的辨識，且電腦系統有不會疲累的特性，因此可以輕易地兼具質與量的對人名進行性別辨識。在性別資訊不足或獲取成本較高時，本系統便能發揮很好的價值，以下舉出幾種可能的情境：

1. 設計先進的自動化線上客服系統，自動從使用者電子信箱的署名辨別使用者性別，除了可以配合客戶分析的工作外，也可以給予客戶擁有適當尊稱的回應。
2. 進行性別比例分析，如若無法取得大學聯考考生性別，一樣可以透過本系統概略估算出每個系的男女比例，也可以針對過往資料進行比例趨勢分析。
3. 協助問卷修正性別缺漏的資料，也可以應用於核對資料以避免錯誤。
4. 另外也可以成為華語教學的一個環節，協助外國人士辨認中文人名的性別，提升中文教育的廣度與實用性。

7.2.2 建立競爭智慧系統

競爭智慧（Competitive Intelligence）是指蒐集外在環境，包括社會結構、經濟變遷、科技與法律及競爭對手相關資訊的技術，常用以協助商業智慧系統發展。由於本實驗中可以發現，在不同性別傾向的產品命名上，會使用不同的名稱，因此，我們可以使用此一特性，加上系統自動預測的能力，對競爭對手尚未發布的產品做目標族群分析，例如可以知道對手預計將此產品應用在哪個性別，也協助我們及早擬定商業競爭對策。

7.2.3 協助文件探勘的後續研究

由於我們能透過性別比例的標注提供文件中額外的資訊，而這些資訊能夠對文件進行先前無法得知的特徵標記，將有助於目前文件探勘研究的推進。舉例而言，我們可以計算對文件中每個字或句子的性別傾向，進而對文件或句子進行分群或分類。我們可以訂定不同性別傾向的門檻來拆解文章，或將類似的文章聚成叢集。

另一方面，我們也可以透過性別標注輔助在文件中辨別專有名詞或人名的工作，配合其他機器學習的演算法，可能性別傾向可以變成文字的重要特徵。此外我們也可以透過人名加上性別的預測，了解文章中的主角性別，作更深入的社會學分析工作。最後，我們也可以使用女性具有較溫柔、較美麗的特性，透過性別傾向進行文章的觀點分析，或了解語氣的強弱等研究。

7.2.4 提供其他對於文字結構的研究參考

基於本實驗的高精準度及高度與人工預測的共識，我們可以知道中文字的部件的確具有某些特殊的意義，系統也可以藉由中文部件對文字的意義進行判讀，使得電腦從字形中找出中文字的特殊規則。若我們能掌握更多大量的資料，可以讓電腦系統學習中文的其他特性，進入機器情緒感知的領域。例如讓電腦從字的部件對未知詞的概念有一定認知，進而對文句意義進行推測。這裡說的情緒感知與前文中觀點分析不同的是，前文的觀點分析一樣使用「男性」、「女性」的概念對文章進行判讀；而此處是一樣使用倉頡拆字得到中文部件的方式，配合其他的測試資料，得到不同於「男性」、「女性」的分類，例如「喜」、「怒」、「哀」與「懼」等。

7.3 研究限制

本研究透過支援向量機搭配倉頡碼的轉換對人名進行性別的學習，因此有以下限制：

1. 對於非用於人名的中文字無法提供準確的性別預測，例如暱稱、譯名等資料就容易出現錯誤；
2. 基於倉頡碼最多 5 碼的限制，相同構造的部件可能因為和在筆畫複雜的字而

產生簡化，使得系統無法正確判讀出所有的部件，因而降低學習的準確度；

3. 基於資料的有限，無法對 SVM-CJ123 與 SVM-CJ1234 進行足夠的訓練，使得可能得到較高準確率的方法無發發揮。

7.4 未來研究方向

在了解中文的部件具有性別傾向的特性後，在未來的研究可以朝三個方向進行。第一個方向是繼續提升中文性別判斷的正確性，例如透過前述方法來優化支援向量機的結果，或將目前的方法整合成新的判斷機制；第二個方向是將目前現有的性別標注應用在其他文字探勘領域，藉由性別標注創造更多的文件屬性，提升目前文字探勘工作效能；第三個方向是利用中文字部件的觀念，配合其他的訓練資料找出部件中所蘊含的其他意義，進而讓電腦系統更能了解文字本身的意義，創造出更貼心的科技發明。

參考文獻

1. Bergsma, Shane, Lin, Dekang and Goebel, Randy, "Glen, Glenda or Glendale: Unsupervised and Semi-supervised Learning of English Noun Gender", *CoNLL*, 2009.
2. Bloomfield, Leonard, " Language", *Holt, Reinhart & Winston, New York*, 1933.
3. Cassidy, Kimberly Wright; Kelly, Michael H.; Sharoni, Lee'at J., "Inferring Gender From Name Phonology", *Journal of Experimental Psychology: General Vol. 128. No.3. 362-381*, 1999.
4. Chang, G.-M., "A social analysis of person naming in Taiwan for the past century", *Master thesis. Taipei: Tam-Kang University*, 2003.
5. Fryer, Roland G. Jr. and Levitt, Steven D., "The Causes and Consequences of Distinctively Black Names", *Quarterly Journal of Economics Volume119, Issue3*, Pp. 767-805, 2004.
6. Gallagher, A.C., Chen, Tsuhan, "Estimating Age, Gender, and Identity using First Name Priors" *Computer Vision and Pattern Recognition. CVPR 2008. IEEE Conference*, 2008.
7. Hassan, Adb-el-Jawad, "A Linguistic and Sociocultural Study of Personal names in Jordan.", *Anthropological Linguistics 28*:80-92, 1986.
8. Hsu, Y.-S., "A sociolinguistic study on the cultural values reflected in Chinese men's and women's given names in Taiwan", *Master thesis, Taipei: Fu-Jen Catholic University*, 1990.
9. Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", *Proceedings of the European Conference on Machine Learning Springer*, 1998.

10. Kilarski, Marcin, “On grammatical gender as an arbitrary and redundant category”, In Douglas Kilbee, editor, *History of Linguistics 2005: Selected papers from the 10th International Conference on the History of Language Sciences (ICHOLS X)*, pages 24–36. John Benjamins, Amsterdam, 2007.
11. Rossi, A. S., “Naming Children in Middle-Class Families.” , *American Sociological Review 30*: 499-513, 1965.
12. Sung, Margaret, M. Y., “Chinese Personal Naming.”, *Journal of the Chinese Language Teachers Association 16(2)*:67-90, 1981.
13. Nastase, Vivi and Popescu, Marius, “What’s in a name? In some languages, grammatical gender”, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1368–1377, 2009.
14. 五筆教學研究組，《五筆・拼音速查字典：部首檢字版》，機械工業出版社，出版日期：2011 年 1 月 1 日。
15. 中華民國內政部，《全國姓名分析》，內政部出版，出版日期：2010 年 10 月
16. 朱保安，《说文·女部：字所反映的女性社会地位的演变》，殷都学刊 2004 年第 3 期，2004。
17. 李鐵筆，《命名一書通》，*益群出版社*，2009 年 12 月 1 日。
18. 倪耿，《中國文字之結構模式及其分析》，交通大學碩士論文，1972。
19. 張泰昌，《圖解當代漢字與鄭碼輸入法》知識產權出版社，出版日期：2006 年 02 月 01 日。
20. 楊嘉敏，《部首輕鬆學：第六單元：認識女部》，人間福報，發行日期：2008 年 6 月 3 日。
21. 太易資訊，《大易輸入法》，參考網址：www.dayi.com/DAYI_AREA/Default.htm，查考日期：2011 年 12 月。
22. 行易有限公司，《嘸蝦米輸入法》，參考網址：boshiamy.com/，查考日期：2011 年 12 月。

23. 朱邦復，《倉頡輸入法與中文字形產生器》，參考網址：http://cbflabs.com/book/gif_cg/gif_cg/，查考日期：2011 年 12 月。

24. 馬來西亞倉頡之友，《第五代倉頡通用版原始碼表(UTF-8)》，參考網址：http://www.chinesecj.com/newsoftware/download.php?download=http://www.chinesecj.com/download/cj5-21000.zip，最後更新日期：2006 年 10 月 19 日；查考日期：2011 年 12 月。

25. 蘇文彬，《PChome 商店街：年底店家數成長至 1 萬家》，參考網址：http://www.ithome.com.tw/itadm/article.php?c=63838，發表日期：2010 年 10 月 11 日；查考日期：2011 年 6 月

國立台灣大學管理學院資訊管理學系碩士班問卷

指導老師：盧信銘博士

研究學生：魏取向 (r99725042@ntu.edu.tw)

說明：

- 這份問卷的目的是要了解人們對於**中文人名裡的性別涵義**，分辨人名中男性或女性的象徵，以作為後續研究使用。
- 我們會**妥善保管**您在本問卷中的個人資料，絕不作為其他用途，在研究結束後也會依規定程序將問卷銷毀，因此請您不用擔心個資洩漏的問題。

為了要確切掌握填寫者的型態，請配合填寫（或勾選）以下資料：

- 您的名字（不必填姓氏）：_____
- 性別： ◯男 ◯女
- 年齡： ◯18 歲以下 ◯19~22 歲 ◯23~35 歲 ◯35 歲以上
- 教育程度： ◯高中以下 ◯大學 ◯研究所以上
- 與主要為中文名字的人相處時間：
  ◯一年以下 ◯1~5 年 ◯6~12 年 ◯超過 12 年
- 母語：
  ◯國語 ◯閩南語 ◯客家語 ◯粵語 ◯原住民族語
  ◯英文 其他_____
- 最習慣的文字： ◯正體中文 ◯簡體中文 ◯英文 其他_____

提示：

接下來我們會請您判斷表格中的人名，每個人名只有一個選項，不是男性就是女性。請依照**第一印象判斷**即可，不需要過度揣測。

範例：

下表中的人名「信銘」我認為他應該是男生的名字，我就在男生的格子中打個勾：

人名	男姓	女性
信銘	✓	

請翻頁後開始填寫：

人名	男姓	女性
秀珠		
勞鹿		
美雲		
惠美		
燉貳		
月美		
俊宏		
文德		
即趨		
雅玲		
羽倩		
闢狼		
射匪		
明輝		
錦華		
明清		
志宏		
阿雪		

人名	男姓	女性
玻概		
清		
龐樽		
雯		
奪點		
淑卿		
津芳		
建成		
欽		
秀玲		
士傑		
添妹		
志成		
聰明		
永敏		
宗欽		
蚵額		
凱媛		

辛苦了！已經寫了 1/6

人名	男姓	女性
釣鷺		
玲玉		
建賢		
順		
美玲		
森		
素貞		
世燦		
秋榮		
定		
佳霖		
革譚		
蹤徑		
粘遴		
帖漠		
國英		
文祥		
麗珠		

試試把名字冠上常見的姓念念看，會比較好分喔

人名	男姓	女性
家華		
陰壟		
冠樺		
吉		
霸崐		
售汐		
發		
秋金		
怡君		
裕淞		
芳銘		
惠群		
牆敖		
昌廷		
柏驊		
撒鴨		
趾橫		

局賠		

寫到這邊已經完成 1/3 了！

人名	男姓	女性
阿貴		
建宏		
志翔		
文傑		
靜姍		
冠穎		
鷁繆		
阿屘		
庭瑋		
雅雯		
雅惠		
平		
志鴻		
惠玲		
捷		
勝弘		
嘉華		
凸者		

人名	男姓	女性
安倫		
志榮		
明妹		
睛抵		
甚		
靖		
淑英		
春華		
淑美		
秀琪		
功強		
瑞芳		
志雄		
武雄		
爾嘉		
文雄		

淑玲		
淑芬		

人名	男姓	女性
恩		
呆答		
麗華		
守龍		
啞眠		
飆鏢		
秀美		
鈞		
寶華		
隊迷		
匹穫		
糯去		
緋糊		
家豪		
文龍		
盃磋		
鷗講		
安郎		

想像一下擁有這名字的是帥哥（或正妹），這樣比較有趣

人名	男姓	女性
淑惠		
麗卿		
丹		
冠宇		
庭安		
明芳		
美妹		
搬稅		
璋		
玉清		
淑娟		
雅如		
群		
俊男		
瀝噶		

俊賢		
國泰		
朝熙		

還差 1/3 就完成了

人名	男姓	女性
和		
秀梅		
育束		
子儀		
杏芳		
宇		
了茨		
正義		
嘉文		
篷托		
文章		
章		
梭輻		
正雄		
拉		
謙		
秀明		
源峰		

人名	男姓	女性
謠瀾		
沃唅		
趁劫		
美珠		
英華		
牲礦		
汽拎		
三郎		
砲聽		
郁文		
熱棘		
成茂		
艙尖		
志強		

世英		
子欣		
宥均		
榮光		

還差 1/6 唷！

人名	男姓	女性
志豪		
玉珍		
秀琴		
日春		
水金		
志銘		
春貴		
侶跳		
宇庭		
靖恩		
千瑜		
榮		
金連		
美惠		
俊充		
此誇		
正芳		
月娥		

人名	男姓	女性
秀珍		
鍰塊		
佳儒		
志忠		
宗翰		
斑繞		
玉梅		
淑媚		
浣羽		
金龍		
文忠		
雅婷		
修夢		

誣藥		
弗豁		

春美		
銘鴻		

絕猴		

再幫我填最後一條

人名	男姓	女性
玟德		
幟稱		
靠跑		
粽跡		
義		
淑貞		
輝		
清秀		
緝輸		
文		
惠婷		
霖		
秀蘭		
政智		
俊傑		
臀嚇		
縱坎		
麗娟		
乃文		
金湖		
子嘉		
俊雄		
嘉茵		
瑤寂		

完成了，

非常非常感謝你！