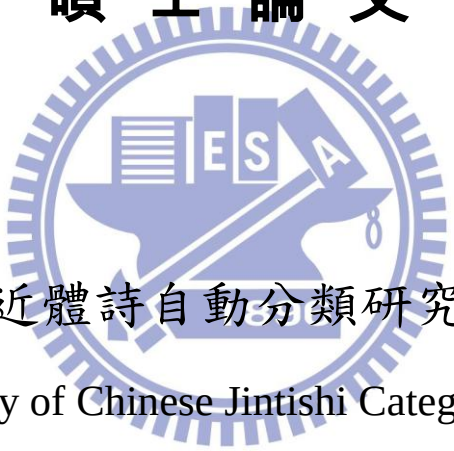
國立交通大學

資訊科學與工程研究所

碩 士 論 文



近體詩自動分類研究

The Study of Chinese Jintishi Categorization

研 究 生：劉博榮

指導教授：梁　婷　教授

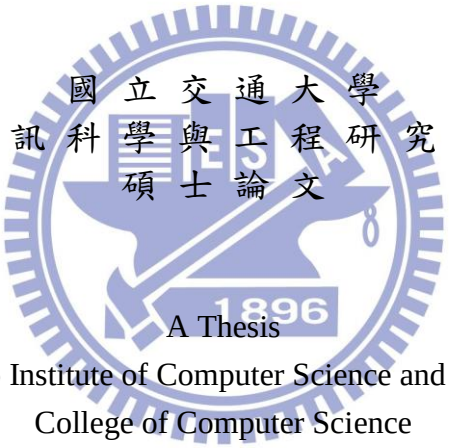
中 華 民 國　九 十 九 年 八 月

近體詩自動分類研究

The Study of Chinese Jintishi Categorization

研 究 生：劉博榮　　Student：Po-Jung Liu

指導教授：梁　婷　　Advisor：Dr. Tyne Liang



國 立 交 通 大 學
資 訊 科 學 與 工 程 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Computer Science and Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science

August 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年八月

近體詩自動分類研究

研究生：劉博榮　　指導教授：梁　婷　博士

國立交通大學

資訊科學與工程研究所

摘要

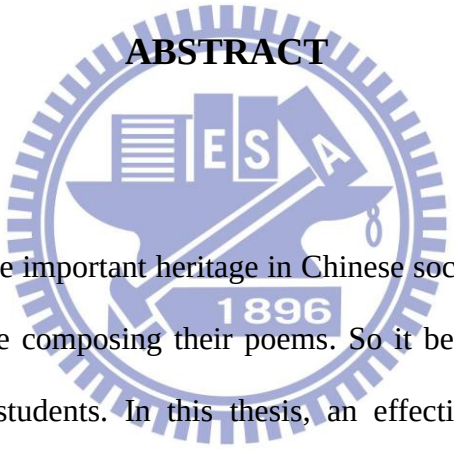
近體詩是華人社會中一項重要的文化資產，然而很多詩作中皆含有隱喻，使得近體詩對於學生而言不容易了解其中含義。在本論文中，我們提出幾個有效的方法來做近體詩的自動分類，藉以幫助學習者對於詩作的理解。我們利用法則式的方法搭配同義詞詞林來做語意標記，以及 SVM 的分類模型來做詩作分類。並從詩作的語料中探勘七種特徵來做為分類特徵，再利用 Forward Sequential Selection Algorithm 來做為選取特徵的演算法，而我們所提出的方法經過 217 首的五言絕句來做六個類別近體詩的詩作分類實驗，可達到 72.35%的正確率。

The Study of Chinese Jintishi Categorization

Student：Po-Jung Liu　　Advisor：Dr. Tyne Liang

Institute of Computer Science and Engineering

National Chiao Tung University



ABSTRACT

Chinese Jintishi is one important heritage in Chinese societies. Nevertheless, many poets use metaphors while composing their poems. So it becomes hard to understand Jintishi for high school students. In this thesis, an effective approach to automate Jintishi is presented with the aim to facilitate poem comprehension. We propose a method to tackle with semantic role labeling based on Tongyici Cilin and a SVM-based model to handle poem categorization. The categorization employs seven kinds of features mined from training corpus. Best set of features is selected by using forward sequential selection algorithm. The approach is justified in terms of 72.35% accuracy by categorizing 217 five-character quatrains into six types of Jintishi.

誌謝

首先最感謝的是我的指導教授梁婷老師，感謝老師在這兩年的研究所生涯中啟發我對於論文研究的方向，並指導我對於論文寫作上的技巧，而且除了學術理論上的教導外，老師也經常將自己的人生經驗和我們分享，在老師的啟蒙之下著實讓我獲益良多。另外也要感謝口試委員張俊盛教授、楊武教授、葉慶隆教授對於論文的寶貴建議，讓本論文更加完善。

其次要感謝楊哲青學長，總是能將我零散的想法整理成完整的架構，每一次的討論都能將論文的內容向前推進一步，即使在老婆生子的期間仍撥出時間幫我校稿，真的十分的感謝!而典松學長告訴我們一些研究生活的處事之道，這些經驗談都相當有幫助。還有實驗室的冠熙老大，每次總是不辭辛勞的處理實驗室大小事務，解決我們各式各樣的疑難雜症，你的樂觀態度也總是讓我在低潮時給我力量，我們實驗室能有你這名大將真好。俊樺是我程式語言的啟蒙導師，每次程式有問題時問你總是能夠迎刃而解，那快速敲打鍵盤的背影實在太帥氣啦。家棋在實驗室時總讓實驗室充滿活力，不愧是有固定運動習慣的陽光男孩。笙權常幫我分擔系統的程式部分，而且總是有效率的完成，在此也感謝你的付出才得以完成論文中的系統。羿賢總是主動的幫大家訂便當和買飲料，解決大家的民生問題，這麼貼心的學弟哪裡找呢。泓達一整個就是情報王，和你聊天總是很開心，之後有機會再來交流一下吧。還有其他無法一一詳述的朋友們，感謝你們在我研究生涯中的一路相伴。

最後要感謝我的家人，感謝你們在我求學路上徬徨無助時適時的拉我一把，陪我分享著每一刻的快樂與憂傷，有了這個強力的精神支柱，讓我更有勇氣在人生道路上大步的向前邁進。

目錄

表目錄

第一章 緒論

1.1 研究目的與動機

近體詩是指唐代形成的格律詩體，以有別於前代不甚講究格律的古體。因其字數、句數、平仄和用韻都有一定的嚴密格式和規律，因此學習近體詩對一般中文語言使用者而言是困難的。有鑑於此，本論文提出一個近體詩的自動化處理程序，包括斷詞程序、語意標註和詩作分類，期使對於詩詞的推廣與電子化建檔有相當大的幫助。

由於近體詩運用了大量的意象，其詩作風格難以單純利用表面上的文字找出[王迺仁 ‘06]，故將詩作中的詞彙概念尋找出來便是詩作分類中相當重要的課題。因此在本論文中提出以同義詞詞林和語料庫的詞彙的統計訊息進行詩作語意處理。另外我們利用統計模組(支持向量機，Support Vector Machine) [Vapnik et al. ‘95]和特徵選取的方法做詩作的分類。此外，我們利用卡方檢定(Chi Square Test)[Liu and Setiono ‘95]和 Forward Sequential Selection Algorithm [Le et al. ‘04]進行最佳特徵組合選取。實驗證明，所提的方法是可行的。

1.2 問題定義

詩體流傳至今，各種詩體均累積了相當可觀的數量，為方便使用者的查詢和使用上的需求，編纂者會將各種詩體作不同的分類，讓使用者能夠快速的查詢到所需的資料。羅鳳珠[‘08]對於詩作分類可分為“依據標題”或“依據標題和內容”兩類分類方式，若採用依據標題的分類法，當標題的意義和詩文內容不符時，

便會產生分類內容與詩作主題不和的情況，以蔣維翰的<古歌，二首之二>為例，其詩文內容為：“美人閉紅燭，獨坐裁新錦。頻放剪刀聲，夜寒知未寢。”，其詩文內容描述了婦人獨守空閨的心情，但詩題卻只寫明了詩作的形式，無法明確的表達詩文內容的意思，在做分類時只利用詩題即會出現錯誤。如能以詩作之內容作語意標記與分類，作為判斷詩作主題的依據，對於分類結果會較為準確，也對使用上的實用價值較有幫助，故我們在此研究中採用依據標題和內容的分類法。

在近體詩分類上，我們採用龔霽芃 [‘03]和朱我芯[‘07]採用的分類方法，這兩個方法皆採用依據標題和內容的分類方法，將兩個分類方法統整後，按題材內容分為六個類別如下：

本篇論文	朱我芯 [‘07]	龔霽芃 [‘03]
詠物述志	詠物述志	詠物抒懷
山水田園	山水田園	山水紀行
		田園隱逸
情愛閨怨	情愛閨怨	戀情閨怨
贈別思友	思鄉送別	羈旅鄉思
		舊雨新朋
		科場內外
邊塞征戰	邊塞征戰	邊塞軍旅
社會民生	社會民生	詠史懷古
		感時諷世

表 1：詩作分類類別對照

我們將本論文所使用的六個類別定義如下：

1. 詠物述志：借萬物寄託詩人自己的感情，詠物詩中的寄托往往跟詩人的經歷際遇、人生態度、生活作風、價值取向等有關係，以表現手法上說是借物來抒發志向。
2. 山水田園：描寫清新的自然景色，山水草木，都富含詩人獨特的審美情趣，或歌詠閑適恬淡的田園生活、田間勞作為題材的詩歌。
3. 情愛閨怨：描寫男女愛慕之情和愛情生活，或抒發離別相思之苦，大多是用第一人稱來直敘自己的愛情，也有些是以第三人稱觀點來寫。
4. 贈別思友：表現朋友之間的摯愛深情、離情別緒，一般為即景抒情，詩的開頭是敘事，或寫景，然後是抒情表意。
5. 邊塞征戰：描寫邊塞風光、反映邊疆將士生活為基本內容，抒發報效國家、渴望建功立業的豪情，或狀寫將士的鄉愁、邊塞征戰的殘酷、描寫塞上絕域的奇異風光等。
6. 社會民生：利用嘲諷或勸喻手法，揭露社會黑暗、世態炎涼。或以憑弔古跡、歷史故事、古人事跡為題材，借此抒發情懷，諷刺時事。也有懷才不遇時，詩人抒發情感，或是感嘆年華老去仍無所做為。

此六個類別的範例詩作如表 2：

類別	詩題	詩文
詠物述志	梨花	豔靜如籠月，香寒未逐風。 桃花徒照地，終被笑妖紅。
山水田園	登鸛雀樓	白日依山盡，黃河入海流。 欲窮千里目，更上一層樓。
情愛閨怨	浣紗石上女	玉面耶溪女，青娥紅粉妝。 一雙金齒屐，兩足白如霜。
贈別思友	送別	山中相送罷，日暮掩柴扉。 春草明年綠，王孫歸不歸。
邊塞征戰	哥舒歌	北斗七星高，哥舒夜帶刀。 至今窺牧馬，不敢過臨洮。
社會民生	傷秋	歲去人頭白，秋來樹葉黃。 搔頭向黃葉，與爾共悲傷。

表 2：詩作類別範例

1.3 論文架構

在本論文中，我們提出如何從詩作中分析概念的方法，從詩作中選取具代表性的概念，並搭配其他的詩作特徵，然後利用分類模型來自動分類的工作，並分析實驗結果挑選出最佳的特徵組合。

本論文結構如下：第二章為論文的相關研究部分；第三章和第四章探討研究方法的部分，首先於第三章中介紹從詩作中找尋其概念的方法，並解決概念標記的岐義問題，然後在第四章中，對於詩作中所存在的特徵加以擷取並分析，之後詳述我們於自動分類中所使用的分類模型，並對分類器使用的特徵尋找適當的特徵組合，第五章敘述結論和未來的發展方向。

第二章 相關研究

2.1 同義詞詞林

同義詞詞林這個辭典[梅家駒 ‘06]，它主要收錄為現代的漢語詞彙，也收錄部份常見的古語詞，和其相似的語料庫有 E-Hownet [Chen et al. ‘05]、SUMO[Niles and Pease ‘03]…等，這些語料將概念視為節點，概念間會有互相引用的關係，詞彙與概念間的對應較為複雜，相對而言同義詞詞林對詞義的分類架構較為明確，從大類到小類之間有明確的階層關係，故我們主要利用同義詞詞林來進行詩作詞彙的概念標記。

同義詞詞林共分成 12 個大類，94 個中類，1428 個小類，小類之下再劃分成 3925 個詞群。其中前四大類(A、B、C、D)多屬名詞，第五大類(E)屬形容詞，第六至第十大類(F、G、H、I、J)多屬動詞，第十一大類(K)屬虛詞，而第十二大類(L)為敬語，同義詞詞林共收錄詞彙約七萬個，我們以表 3 說明同義詞詞林的“階層代號”：

詞彙	階層代號	階層概念
鴛鴦	Ah08	夫、妻、夫妻

表 3：同義詞詞林詞彙範例

其中大寫字母所標示的為階層的大類概念(Level 1，編號為 A、B、C…)，此大類概念底下的為中類概念，其用小寫字母標示(Level 2，編號為 a、b、c…)，在中類概念底下的為小類概念，其用數字表示(Level 3，編號為 01、02、03、04…)，

我們在此研究中，所用到的分類特徵之一即為小類概念，故將其特別標記於“階層概念”中，且小類概念下對應了若干的名詞詞彙(Level 4)。我們將詩作斷詞過後的詞彙對應到同義詞詞林中查詢，可以得到其在同義詞詞林的階層編碼，做為斷詞其概念階層中的位置編碼，以表 3 的“鴛鴦”這個詞彙為例。其大類的概念編號為(A)，對應的大類概念為“人”。其中類的概念編號為(h)，對應的中類概念為“親人、眷屬”。其小類的概念編號為(08)，對應的小類概念為“夫、妻、夫妻”。

2.2 詞義處理

詞義處理主要是解決自然語言中一詞多義的問題，目前常用於解決詞義消歧的方法有以字典比對的詞義消歧技術和以語料為主的統計方法，後者又可分為監督式詞義消歧及非監督式詞義消歧，以下將根據這幾個重要的方法分別做介紹。

在字典為基礎的詞義消歧技術中，字典提供了字彙以及相關的定義，對於一個要被標記的詞彙而言，其所有的意思皆先會被找出，再去比對字典內的定義和分類的內容來當作其特徵值，找出與原需被標示詞彙最相近的意思，在有大量可使用的語料庫和字典時，可應用此方法。

在監督式方法的詞義消歧中，會先收集已標記的語料當作系統的訓練語料，系統再根據這些訓練語料當作詞義分類的特徵，搭配系統內設定的演算法，來標記其他未標記的測試語料。

Navigli [‘06]使用 Semantic network 和 Semantic interconnection pattern 來解決詞義消歧(Word Sense Disambiguation)的問題，作者並設計了一個圖形化的界面 Valido 來結合人工標示和機器標示的部份，並採取多數決的方式來決定語意，利用人工

與機器並行的方式來提高詞義消歧的正確率，但明顯的需要使用人工標示時仍需耗費心力，對大量的語料(Corpus)而言此方法並不實用。

Chan and Ng [‘07]利用 Naive bayes 來解決 Active learning 在一般語料(Brown Corpus)和財經語料(Wall Street Journal)間詞義消歧的問題，以 Brown Corpus 作為訓練的語料庫去標記測試的語料庫 Wall Street Journal，並使用 Count-merging 對於不同領域的語料對字義給予權重並預測其出現機率，再利用 EM(Expectation-Maximization)-based algorithm 來對字義作優先順序的排序。

Escudero et al. [‘00]使用 POS 標記和目標詞的周圍詞彙這些當作特徵，利用 Naive Bayes、Lazy Boosting…等方法來做 DSO 語料的監督式詞義消歧，正確率可達 72.09%。

非監督式的方法的詞義消歧不需利用已標記好的語料，而是從大量未標示的語料庫中搜集其特徵，根據數學公式計算建立起相似矩陣並將其分群，在同一群之中的詞彙就會被標示相同的意思，此方法可透過不同種類的語料來作標記，讓語料的資源取得較為容易。

Kwong and Tsou [‘05]主要解決中文語料於語意標記(Semantic Role Tagging)中，找出句子內中心詞(Headword)的問題，其利用被辨認為中心詞的機率、詞彙位置、中心詞詞性、介系詞的相對位置來做為機率模型(Probability model)的特徵(Feature)，來找出句中的中心詞，但從實驗結果看出，若無法知道句子的 Boundary condition 時，其辨認的正確度在訓練資料集和測試資料集不同的情況下，於 Textbook Data 下只有 46.32%，於 News Data 下只有 39.64%，可見若無事先切詞而直接以機率模型去做判斷，系統的運作結果成效不彰。

Kruengkrai et al. [‘09]提出利用單字(Character)和詞彙(Word)的混合式架構，對中文句子作切詞(Word Segmentation)和詞性標示(POS Tagging)的動作，主要是以詞彙層級(Word-level)的部分，找出那些已知的詞彙，再以單字層級(Character-level)的部分，利用單字長度、單字於詞彙中的位置、和單字的詞性作為特徵，在 Online learning algorithm 的架構下去訓練系統，找出那些於詞彙層級(word-level)時無法找出的詞彙，增加切詞和詞性標示的正確性。但此方式所使用的特徵過於繁雜，除了考慮一般所使用詞彙層級(word-level)的切詞方式，還需考慮單字層級(character-level)的特徵，勢必使得時間複雜度大幅的提高。

而在目前中文的詞義處理研究中，缺乏大量精確的詞義標示資料，目前只有少數幾個規模不大的中文詞義標示語料集，如 Senseval-3 的中文詞義標記語料集儘包含 20 個中文詞，柯淑津等人[‘07]中研院平衡語料庫為對象從中摘錄文章，並對文章中的詞彙採用監督式方法做詞義標記的動作，其正確度可達 64.51%

2.3 詩作分類

在英文方面，運用於詩作分類的種類有作者分析、風格分析、體裁分析…等，其中所使用的特徵有語意標記、標題、風格、體裁、作者、年代…等 [Plaisant and Rose ‘06]，而中文的詩作分類目前主要仍以分析詩作的字詞為，認定其風格的方式主要就需要以文字做為主體，利用文字做為特徵，也可將相同的文字歸類在同一概念或類別，將詩作進行概念擷取的動作，讓分類器依據這些特徵值，來進行詩作分類的動作。但目前中文的詩作分類因有標記語意的語料庫不足，且大部分皆以白話文的詞彙為主，缺乏文言文的詞彙，使得對於詩作中的語意辨識變成分類的一大挑戰。且無一個以分類好的標準語料庫做為系統的測試基準，詩作分類前仍必需進行語料收集的工作。

Gamon[‘04]利用微軟所開發的 NLPWin 來對 Brontë 三姐妹的文學作品作語意、詞性和句型的分析，然後利用這些分析出來的特徵經由頻率的篩選之後，再利用 SVM 分類器來做作者的分類，分成三類的正確率可達 97.5%。

Koppel et al. [‘04]對 BNC(British National Corpus)中 264 首小說語料做性別分類，他先利用 BNC 的語料庫對文章內的詞彙做詞性標示，再利用文章的表面文字和詞性，使用 Exponential Gradient algorithm [Kivinen and Warmuth ‘95]做性別分類，正確率可達 79.5%。

王迺仁[‘06]提出近體詩階層式概念，將詞義相同或是相近的詞彙歸類於同一類，歸納相同語意概念的語詞為名詞類別，其部份名詞參考 SUMO(Suggested Upper Merged Ontology)及同義詞詞林的架構建置而成，且進一步利用關聯規則(association rule)探勘詩中使用詞彙類別的組合，依可信度(confidence)及支持度(support)分析詩人詩作因詞彙使用不同的風格判別規則。但其分類的特徵侷限於名詞，詞性的缺乏使其無法涵蓋詩中所想要表達的情感，且所分析的實驗資料僅限於王維的 385 首詩作，資料量的不足讓分類時所能夠參考的分類特徵不夠準確。在雙重的限制之下使得分群結果不盡理想，所使用的評估結果也顯示無法確實的將王維詩詞分類於作者所定義的六個類別中。

許嘉妮[‘07]利用宋詞斷詞器來對宋詞做斷詞的動作，並經由同義詞詞林和宋詞的相關特徵來辨識詞彙概念，然後利用情境規則來對宋詞做豪放和婉約的分類，其正確率約可達到 74.4%。

Yi et al. [‘05]運用機器學習技術分析宋詞的風格，主要是以詩中單字詞的出現率作為分類的基準，將宋詞分為豪放與婉約兩種風格。先利用基因演算法找出影響宋詞風格的主要單字詞，再運用貝氏分類演算法來建立宋詞風格分類的模式。但在宋詞中常使用多字詞和典故相關的詞彙，單字詞的詞義往往在排列組合後會產生與原來不相同的情況，因此會影響到分類時的準確度。

Li et al. [‘04]等人利用詞連接的自然語言分析方法，將詩詞風格分為豪放與婉約兩種風格。他將詞彙分為柔和、中性、強烈三個字集，再利用詞彙之間的連接關係建構最佳搜尋樹，來決定句子中的中心語和詞彙所構成的句子的語意傾向，並利用句子的語意來推導出整個詩詞的風格。但此方法在使用之前需先建立龐大的語彙間的語意架構資料庫，對於古典詩詞常會出現現代漢語的非常用字而言，顯得相當的不容易，且此方法的評分方式如同意見探勘(opinion mining)的二元傾向評分法，只能做二元的分類，對於多樣化的詩詞分類法時，此方法即無法使用。

我們將詩文分類的相關研究整理如表 4：

	英文		中文				
	[Gamon ‘04]	[Koppel et al. ‘04]	[王迺仁 ‘06]	[許嘉妮 ‘07]	[Yi et al. ‘05]	[Li et al. ‘04]	本研究
外部支援辭典	無	BNC	同義詞詞林	同義詞詞林,CKIP	無	專家制定的詞集	同義詞詞林
斷詞處理	無	無	有	有	無	無	有
概念標記	有	有	有	有	有	有	有
實驗語料	勃朗特作品 1441 篇	BNC 中的 264 首小說	王維詩作 385 首	100 首宋詞	398 首宋詞	55 首唐詩、414 首宋詞	1080 首五言絕句
訓練：測試語料比例	8：2	9:1	相同	5：5	相同	未提供	8：2
詩題分析	無	無	無	無	無	無	有
分類模組	SVM	Exponential Gradient algorithm	k-means	規則式斷詞	Naïve Bayes	詞彙連接計算	SVM
特徵選擇	頻率	頻率	無	無	基因演算法	無	Chi Square Test
類別數	3	2	6	2	2	2	6
正確率	97.5%	79.5%	未提供	74.4%	88.5%	未提供	72.4%

表 4：詩文分類的相關研究整理

第三章 詩作處理

以往對於詩作分類的研究，常拘限於表面詞彙做為特徵[Li et al. ‘04] [Yi et al. ‘05]，這樣在找尋詞彙特徵時會遇到一義多詞的問題，如以“儲藏”而言，可能就可換成“貯備”、“積存”、“儲存”等詞彙，故我們認為應將這些表面字詞歸類成單一的概念，在我們選擇詩作分類的特徵時，詩詞概念與其出現位置也是影響詩詞風格的重要因素。在此章節我們將陳述如何辨識詩作中的概念。

對於詞彙概念標記的解決流程如下：

1. 對詩作輸入後進行斷詞處理。
2. 利用同義詞詞林對詞彙進行語意標記。
3. 找出詩作中同義詞詞林所無法標記的詞彙進行處理。
4. 對於部分的語料做人工概念標記，再利用這些語料，對多義詞進行概念岐義的處理。

3.1 語料前置處理

我們所使用的詩作語料庫，是將維基文庫其五言絕句目錄中所陳列的詩題為主，因為於維基文庫[1]中所記載之資料常有缺字的現象，故我們再於“【新詩改罷自長吟】全唐詩檢索系統”[2]下取得相關資料，建成五言絕句語料庫，其格式如表 5 所示：

[1] http://zh.wikisource.org/
[2] http://cls.hs.yzu.edu.tw/tang/database/index.html

詩題	作者	詩文
春怨	金昌緒	打起黃鶯兒，莫教枝上啼。 啼時驚妾夢，不得到遼西。

表 5：五言絕句語料庫範例

有了初步的語料庫後，我們再根據之前所定義的六個類別：詠物述志、山水田園、情愛閨怨、贈別思友、邊塞征戰、社會民生，對這些詩作進行人工的分類，我們請三位研究所學生來對詩作進行分類的動作，而若出現標記者之間的歧義，我們則採多數決的方式，讓每一個詩作有單一的類別，為標記者標記詩詞類別的情況：

詩作類別標識狀況	詩作數
全部人相同	343
多數人相同	497
皆不相同	240

表 6：標記者類別標記情況

其中標識者標記皆不相同的情況，即為詩詞可能具有同屬不同類別的定義或是其語意較不明顯，如“行背青山郭，吟當白露秋。風流無屈宋，空詠古荊州。”這首詩可同時被歸類為“山水田園”和“社會民生”類，故會產生標記者不相同的情況。

而我們為了統一輸入斷詞系統的格式，將詩句在資料庫記載中有分岐的部分去除，只保留單一的版本。且維基文庫中也會有詩作重覆的情形，我們在語料庫中

也將其他重覆的部分刪除。最後在五言絕句語料庫共有 1080 首詩作，其語料範例與各類別的詩作數目如表 7 所示：

分類	詩作數
詠物述志	144
山水田園	363
情愛閨怨	126
贈別思友	257
邊塞征戰	58
社會民生	132
小計	1080

表 7：五言絕句語料庫各類別詩作數目

在進入詩作的分類前，詩作必需處理成適當的格式，才能夠進行後續的概念辨識和特徵選取，而對於中文的詩文部分需先經過斷詞的處理，在此我們使用陳紹宜 [‘10]所開發的“啟發式規則斷詞系統”，將未處理的詩作其詩文的部分做斷詞的動作。此系統主要利用五項啟發式規則來當作其斷詞的準則：

1. 句型規則：利用詩詞中不同的句型規則來做為斷詞的模式，以五言絕句為例，一句有五個字，可根據不同的句型規則切分詞彙為 2/3、2/2/1、2/1/2…等九種型式(2/2/1 的格式表示 兩字/兩字/一字)，且依據詩作中使用的頻率制定其使用優先權[許清雲 ‘97]。
2. 已知詞彙：利用辭典中有收錄的已知詞彙，已知詞彙較多的句型規則有較高

的斷詞優先權。

3. 最長詞彙：若已知詞彙中字數較長的詞彙較多，則符合的句型規則有較高的斷詞優先權。
4. 專有名詞：指人名、地名…等獨特個體的名詞，專有名詞較多的句型規則有較高的斷詞優先權。
5. 典故：指詩作中所使用到的典故有關的詞彙，利用典故資料庫查詢而得，典故詞彙較多的句型規則有較高的斷詞優先權。

“已知詞彙”是用來決定“句型規則”中最重要的規則，而“句型規則”則是用來切分詩作的重要工具，若句型規則中的已知詞彙較多的話，此句型規則就有較高的斷詞優先權，若有未知詞彙在所有的句型規則中時，“最長詞彙”的規則就可用來決定那一個句型規則有較高的斷詞優先權，若沒有符合最長詞彙的情形時，就可利用“已知詞彙”、“專有名詞”、“典故”的數目來決定哪一個句型規則優先權較高，最後比較詩作第一句和第二句間較高優先權的句型規則，即可得到斷詞所使用的句型規則。

範例 1：半朽臨風樹，多情立馬人。

這句詩作利用“已知詞彙”和“句型規則”的優先權，選出的句型規則第一句為 2/2/1，第二句也為 2/2/1，故最後的切詞結果為“半朽，臨風，樹，多情，立馬，人”。

範例 2：山窮水盡人，柳暗花明村。

這句詩作並為所有詞彙皆為“已知詞彙”，故使用“最長詞彙”的規則，而在這個例子第一句為 4/1，第二句為 4/1，最後的切詞結果為“山窮水盡，人，柳暗花明，村”。

我們將此系統的切詞字數和句型規則頻率統計如下：

詞彙字數	個數	不重覆個數
一字詞	4345	1106
二字詞	8554	6513
三字詞	49	45
小計	12948	7664

表 8：系統切詞字數統計

	最高頻率規則	詩作數	次高頻率規則	詩作數
第一句	2/2/1	730	2/1/2	258
第二句	2/2/1	786	2/1/2	211
第三句	2/2/1	747	2/1/2	261
第四句	2/2/1	740	2/1/2	252

表 9：句型規則頻率統計

由表 8 和表 9 可得知，此系統傾向於將詩詞切成一字詞和二字詞，對於三字詞的切詞數較少，而此系統的斷詞的 F-score 為 69.15%，系統的優點為利用句型規則來做切詞符合近體詩的創作原則，若使用的句型規則正確對於韻文就能做準確的切詞，其缺點為若句中遇到資料庫中未收錄的詞彙其數量很多時，就會產生切詞的錯誤。如：“功蓋/三分國，名成/八陣圖。江流/石/不轉，遺恨/失吞/吳。”此為正確的切詞情情，但系統的切詞結果為“功蓋/三分/國，名成/八/陣圖。江流/石不/轉，遺恨/失吞/吳。”，此結果是由未知詞彙所產生的錯誤。

3.2 詞彙語意處理

在詞彙語意處理中，在 3.2.1 節我們先對詩作中的詞彙作同義詞詞林的語意比對，接下來找出詩作中同義詞詞林所無法標記的詞彙進行處理，而於 3.2.2 節中對於部分的語料做人工概念標記，再利用這些標識好的語料，對多義詞進行規則式的概念岐義的處理。

3.2.1 語意辭典比對與未知詞彙處理

將詩作的詩文做好斷詞之後，我們先利用同義詞詞林[‘97]做初步的斷詞的概念標記。經過初步標記的斷詞其格式如表 10 所示：

詞彙	階層代號	階層概念
鴛鴦	Ah08	夫、妻、夫妻
鴛鴦	Bi11	禽獸、禽

表 10：初步概念標記範例

當詩作斷詞中出現同義詞詞林所無法標記雙字以上的詞彙時，我們會利用中研院八萬目詞[3]和 E-Hownet[4]和典故資料庫，尋找此詞彙是否為一般漢語所使用的詞彙，若在這些詞庫中仍無法找到該詞，我們會根據該詞於所有詩作斷詞的出現頻率，來決定是否該保留該詞彙，若皆不符合以上的條件，我們會將其拆成單字詞，分別於同義詞詞林中尋找。對於未知詞彙的解決流程如表 11：

[3]八萬目詞，中文詞知識庫小組(1995)。台北南港：中央研究院資訊科學所。
[4]E-Hownet，中央研究院資訊所詞庫小組(2007)。台北南港：中央研究院資訊科學所。

輸入：同義詞詞林無法辨別之詞彙

輸出：需手動標記概念的詞彙與不標記概念的單字詞

步驟 1：檢查未知詞彙是否出現於八萬目詞、E-Hownet、典故資料庫中，

若詞彙存在，將其加入同義詞詞林並手動標記概念。

步驟 2：若詞彙不存在於步驟 1 的三個資料庫中，

步驟 2.1：檢查它的出現頻率是否大於等於二，

若成立就將詞彙加入同義詞詞林中，並手動標記概念。

步驟 2.2：若不符合，將詞彙拆成單字詞，並於同義詞詞林中尋找，

若單字詞仍不存在，則放棄標記。

表 11：未知詞彙處理

經未知詞彙處理流程後，需做人工標識詞彙的數目佔全部語料比例的 10.58%，而經過人工標識的詞彙範例如表 12：

詞彙	階層代號	階層概念
蘆花	Bh02	花、花卉
芭蕉	Bh07	水果
金殿	Bn23	皇宮、祠堂、佛殿
隴西	Cb08	地方、地點
石徑	Cb27	道路、路線
芳心	Df05	心意、心情、愛情
州牧	Di15	地位、職位
短歌	Dk27	詩、詞、賦、令

表 12：人工標識詞彙範例

詞彙經過未知詞彙處理後，最後仍會剩下不標記概念的單字詞，對於同義詞詞林所無法標記的單字詞部分，在詩作中同義詞詞林所無法辨識的單字詞佔全部語料比例的 4.48%，對於分類結果不會產生太大影響，故本研究不處理同義詞詞林無法辨識的單字詞部分。

3.2.2 啟發式規則概念歧義處理

對於詩作中所出現的多義詞，我們需將每個詞彙指定單一的概念之後當作詩作分類的特徵，在此我們提出了三個啟發式的方式循序來解決，即共同出現詞彙、共同出現概念、短距離概念階層。

方式 1：共同出現詞彙優先

1. 我們利用 104 首五言絕句從其中建立共現詞彙資料庫。對每類的詩作的每一詞彙，記錄其概念及概念出現次數，我們以“情愛閨怨”類別中的“玉面耶溪女，青娥紅粉妝。一雙金齒屐，兩足白如霜。”這首詩作為例，其中的詞彙將其記錄成< 東 , (S_1,5) , (S_2,2) >, < 足 , (S_3,2) > ， < 女 , (S_4,3) , (S_5,1) >,...，其中 S_i 為概念，S_i 後面為詞彙搭配 S_i 這個概念的出現次數。

2. 此外對同類別中的每一詞彙，記錄共同出現於同一首詩作中的詞彙及次數，例如 < (東, (S_1,5)), (女,(S_4,3)), C_1, 2> , < (足, (S_3,2)), (女, (S_4,3)), C_1, 3>，其中的 C_i 為類別，C_i 後面為該類別此詞彙配對的出現次數。

3. 當有新的詩作進行語意標註時系統將對詩作中兩兩詞彙去尋找共現詞彙資料庫，若找到相符的詞彙配對便可標記其概念。若有多筆搜尋結果，則先比較

兩個詞彙於該類別此詞彙配對的出現次數，若出現次數相同，再比較詞彙 1 於該類別的概念的出現次數，若仍相同，再比較詞彙 2 於該類別的概念的出現次數。

方式 2：共同出現概念優先

1. 我們利用 104 首五言絕句從其中建立共現概念資料庫。對每類的詩作的每一詞彙，記錄其概念出現次數，我們以“情愛閨怨”類別中的“美人怨何深，含情倚金閣。不嘆復不語，紅淚雙雙落。”這首詩作為例，其中的詞彙將其記錄成< 語 , (S_1,5) , (S_2,2) >, < 淚 , (S_3,2) >, < 紅 , (S_4,3) , (S_5,1) >,…，其中 S_i 為概念，S_i 後面為詞彙搭配 S_i 這個概念的出現次數。

2. 此外對同類別中的每一概念，記錄共同出現於同一首詩作中的概念及次數，例如 < (S_1,5), (S_3,2), C_1, 2> , < (S_2,2), (S_3,3), C_1, 1>，其中的 C_i 為類別，C_i 後面為該類別此詞彙配對的出現次數。

3. 當有新的詩作進行語意標註時系統將對詩作中兩兩概念去尋找共現概念資料庫，若找到相符的詞彙配對便可標記其概念。若有多筆搜尋結果，則先比較兩個概念於該類別此概念配對的出現次數，若出現次數相同，再比較概念 1 於該類別的概念的出現次數，若仍相同，再比較概念 2 於該類別的概念的出現次數。

方式 3：短距離概念階層優先

在近體詩中，在奇數句和偶數句上下兩句之間常會有對仗的情形發生，在此兩句中相同位置的詞彙詞性是相對的，如以下詩句：

明月 / 松 / 間 / 照 ，

(名詞) (名詞) (位置詞) (動詞)

清泉 / 石 / 上 / 流 。

(名詞) (名詞) (位置詞) (動詞)

根據此種特性，再加上同一個類別中的詩作，在相近的詩句間，常使用相同概念的詞彙來描述同一件事情，故可推斷上下句的其相對位置的詞彙其概念會是相近的，我們參考了許嘉妮[‘07]所提出的“上下文語意距離”，針對詩作中的岐義詞，先於同義詞詞林中找出其所包含的各種概念的位置(階層代號)，然後分別計算上下句的其相對位置詞彙的概念距離，例如：Aa01 與 Aa02 的概念距離=2，Aa01 與 Ab02 的概念距離=4。

利用以上所敘述的特性，我們用範例將“共同出現概念”這個方法的流程敘述如下：

1. 對於同義詞詞林自動標記的概念，將每個詞彙的擁有的概念表列，我們以“秋天思婦心，雨夜愁人耳。”這個詩作中的“思”和“愁”為例：

詞彙	階層代號	概念
思	Df01	思想
思	Gb06	掛念
愁	Ga02	煩悶

表 13：同義詞詞林自動的標記概念表列

2. 我們對於一首詩中，出現在相同位置的詞彙，比較其概念階層距離，，而對

應的位置以絕句為例，第一句對應第二句，第三句對於第四句，對於出現在相同位置的詞彙，將兩個詞彙的概念列出，並兩兩比較其概念階層距離，若其概念階層距離為最小，便標記其詞彙概念，在“秋天思婦心，雨夜愁人耳。”這個詩句中的“思”和“愁”都在詩作中第三句和第四句的第三個字，故我們如表 14 比較其概念階層距離：

詞彙 1	詞彙 2	概念 1 (階層代號 1)	概念 2 (階層代號 2)	概念階層 距離
思	愁	思想(Df01)	煩悶(Ga02)	6
思	愁	掛念(Gb06)	煩悶(Ga02)	4

表 14：詞彙對應的概念階層距離表

“愁”的階層代碼與概念為煩悶 (Ga02)，而“思”在標示時會有思想 (Df01)、掛念 (Gb06)這些階層的岐義，利用概念階層距離，可看出兩字在第二個配對中的概念階層距離較小，故將“思”的概念標記為“掛念”，“愁”的概念標記為“煩悶”。

3. 若最小的概念階層距離相同時，便比較詞彙和其概念的配對於人工標識概念資料集中出現的頻率，選取頻率較高的配對來標識詞義。

3.3 詞彙歧義消解實驗

對於我們所提出的詞彙標記方法，我們對於系統將每個詩作所標記的概念其平均所標記的概念個數做了以下統計：

分類	概念數 (含/不含重覆概念)		詩作數	平均概念數 (含/不含重覆概念)	
詠物述志	2228	2154	144	15.47	14.96
山水田園	5599	5372	363	15.42	14.80
情愛閨怨	1902	1829	126	15.10	14.52
贈別思友	3880	3275	257	15.10	14.49
邊塞征戰	882	855	58	15.21	14.74
社會民生	1984	1891	132	15.03	14.33
小計	16475	15376	1080	15.22	14.79

表 15：平均概念統計

觀察表 15 可了解，每一首詩作中大約會有一個概念是重覆的，故在同一類的詩作或同一首詩作中，也許作者所使用的詞彙不同，但與該類別相關的概念會被重覆的使用。

詞彙被標記概念數	詞彙數	詞彙被標記概念數	詞彙數
1	6747	11	115
2	2578	12	70
3	2021	13	86
4	1255	14	38
5	1584	15	53
6	672	16	0
7	515	17	0
8	463	18	70
9	184	19	0
10	23	20	1

表 16：詩作詞彙以同義詞詞林標記的概念數

而對於 1080 首詩中每個詞彙去比對同義詞詞林，平均概念數為 3.03 個，若扣除單一概念的部分，則每一個詞彙被標記的平均概念數為 4.44 個，表示每一個有岐義標記問題的詞彙平均會被標記 4 個以上的概念，詞彙被標記概念數可參考表 16。

我們用 104 首詩建立訓練語料的配對資料庫，再拿另外的 104 首詩作用於測試語料，測試語料的詞彙數共有 1603 個，以下表 17 為歧義消解的實驗結果：

	單一概念	共同 出現詞彙	共同 出現概念	短距離 概念階層	人工標記
比對詞數/ 標記詞數 (涵蓋率)	1603/391 (24.39%)	1212/167 (10.42%)	1045/152 (9.48%)	893/547 (34.12%)	346/346 (21.58%)
正確標記 詞數	391	127	122	272	346
正確率	100.00%	76.05%	80.26%	49.73%	100.00%
召回率	24.39%	10.48%	11.67%	30.46%	100.00%

表 17：詞彙標記歧義消解實驗結果

其中“單一概念”這類型的詞彙表示在同義詞詞林中，只會有單一的概念，故我們不計算其概念正確率，而“人工標記”這類型的詞彙在人工標記概念時只會標記單一的概念或是不處理的未知單字詞，故我們也不計算其概念正確率。

從實驗結果可看出，“共同出現詞彙”和“共同出現概念”這兩個啟發式規則其標記正確度明顯地優於“短距離概念階層”，而這幾個方法發生錯誤標記的原因整理如下：

1. 在測試資料集中並非每首詩作皆有對仗的情形，若沒有對仗的詩作，使用“短距離概念階層”時無法標記正確的概念。

2. 詩人於對仗時會要求上下句相同位置上的詞彙詞性相對，但語意有可能不相

近，若遇到此種情況下的詩作也會使得“短距離概念階層”的標記產生錯誤。如以詩作《南行別弟》中的第一、二句為例，“萬里人南去，三春雁北飛。”當中的“人”與“雁”，雖然其詞性是相同的，符合對仗的規則，但語意不相近，此狀況這兩個字便有可能標記相近的但非正確的概念。

3. “共同出現詞彙”和“共同出現概念”在配對資料中，若多義詞的其中一個意思頻率較高時，去標記頻率較低的詞義就會產生錯誤，如“君家何處住，妾住在橫塘。停船暫借問，或恐是同鄉。”這首詩作中的“君”這個詞彙，利用“共同出現詞彙”所標記的意思為“皇帝、后妃”，但其正確的意思應該是“男人”。

在能夠標記的詞彙數上面，“共同出現詞彙”和“共同出現概念”這兩個方法就不如“短距離概念階層”能夠辨識的數量，因這兩個方法需詩作的詩文中有訓練資料集相符合的詞彙或是概念的配對才可進行標記，而短距離概念階層不會有此限制，故形成可標記數量的差異。

第四章 詩作分類

在此章節中我們將會介紹如何對於第三章處理過的詩作進行詩作風格的分類，4.1 節中會提到有關於我們對於詩作的觀察而選取出來的分類特徵，以及如何利用相關特徵選取的理論來進行特徵篩選的動作，4.2 節中會介紹我們的實驗平台，還有在實驗中所使用的工具 Libsvm[Vapnik et al. ‘95] [Chang et al. ‘01]，接下來介紹我們對於尋找特徵組合所使用的演算法，及詩作分類的實驗結果和討論。

4.1 分類特徵

對於在詩作分類中所使用的特徵，我們將其列表如下：

1. 詩題單字詞：將詩題切分成單字詞作為特徵，若非詩題與內容不符的情況下，詩作中的詩題代表了詩作的主題[羅鳳珠 ‘08]，故詩題對於風格分類有相當大的幫助，我們希望藉由分析詩題的文字來尋找其和類別的關係。
2. 詩題雙字詞：將詩題切分成雙字詞作為特徵，理由同上所述。
3. 詩題單字詞概念：利用詩題中的單字詞，其所標記的概念作為特徵。
4. 詩題雙字詞概念：利用詩題中的雙字詞，其所標記的概念作為特徵。
5. 詩文詞彙：直接利用詩文中的詞彙作為特徵，這個特徵是之前與中文詩作分類相關的研究[Li et al. ‘04] [Yi et al. ‘05]經常使用的特徵，我們將其加入特徵中與“詩文詞彙概念”這個特徵做比較。
6. 詩文共同出現詞彙：使用在同一首詩的詩文中，共同出現的詞彙，我們希望尋找在同一類別中，除了經常使用的單一詞彙外，是否會有經常使用的詞彙配對。

7. 詩文詞彙概念：利用詩文中的詞彙，其所標記的概念作為特徵，純粹利用表面字詞來做為特徵時，會使得單一特徵對於分類的代表性下降，故我們認為應將這些表面字詞歸類成單一的概念，也就是對這些字詞進行概念分類，以概念集合來做為詩作的主要特徵。
8. 詩文共同出現概念：使用在同一首詩的詩文中，共同出現的概念，我們希望尋找在同一類別中，除了經常使用的單一概念外，是否會有經常使用的概念配對。
9. 詩文詞彙概念和位置：利用詩文中的詞彙，其所標記的概念和詞彙出現在詩文中的第幾句作為特徵。加入位置的資訊，我們希望尋找詩人於創作詩作時是否會於特定的位置使用有關於該類別的概念，驗證詩人創作詩作時的模式。

在決定使用的特徵後，接下來必需選取這些特徵中具代表性的部分，而有些在單一類別中出現頻率較高的詞彙或是概念，未必就是此類別具代表性的特徵，例如像「人」、「我」這類在每一個類別都會出現的詞彙，若以這些詞彙當作關鍵字，對於整個訓練與分類過程幫助不大，且會降低分類時的正確率，在考量特徵選取時，需以能正確表示類別性質的特徵為主，常用的特徵選取方法有 TF-IDF、資訊增益(Information Gain)、卡方檢定(Chi-square test)…等，在 Yang et al. ['97]實驗中，卡方檢定和資訊增益相較於其他的方法有良好的分類正確度，經我們對於兩個方法的測試後考量所選出的特徵，以卡方檢定來做為選取特徵值的方法。

對於每一個特徵 F_i，本研究所使用的卡方檢定公式如下：

		特徵 F_i	
		有	無
類別	有	A	B
	無	C	D

$$\chi^2 = \frac{n(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{4.1}$$

n：該類別所有的詩作數

A：屬於該類別且有特徵 F_i 的詩作數

B：屬於該類別且沒有特徵 F_i 的詩作數

C：不屬於該類別且有特徵 F_i 的詩作數

D：不屬於該類別且沒有特徵 F_i 的詩作數

除了使用卡方檢定外，我們另外訂定了一項特徵選取公式，稱為特徵類別比例(Feature Class Ratio)，這個公式用來計算單一特徵佔不同類別間的比例，藉以了解特徵對於該類別是否具代表性，特徵類別比例之公式如下：

$Feature\_Class\_Ratio_i$：特徵於類別 i 所佔之比例

$Uni\_Class\_Frequeny_i$：特徵在類別 i 出現的頻率

All_Class_ Frequeny：特徵在全部的類別出現的頻率

$$Feature\_Class\_Ratio_i = \frac{Uni\_Class\_Ratio_i}{All\_Class\_Ratio} \tag{4.2}$$

整合以上兩個公式，我們特徵選取的方法如下：

步驟一：以特徵出現的詩作數為單位，計算卡方檢定的值，並刪除卡方檢定的值太低，對該類別缺乏鑑別度的特徵。

步驟二：以特徵出現的頻率為單位，計算特徵類別比例的值，若特徵類別比例低於訂定的門檻就將該特徵刪除。

4.2 分類實驗

我們發展與測試系統的平台為 Microsoft Windows XP，並使用 Python 做為開發工具。而實驗語料為我們所收集的 1080 首五言絕句，我們將其中的 863 首當作訓練資料集，217 首當作測試資料集來進行分類實驗。

我們在本研究中的分類主要使用支持向量機作為我們的理論基礎，支持向量機是由 Vapnik et al. [‘95]等人所提出以統計學習理論(Statistical Learning Theory)為基礎，針對資料分類、迴歸與圖形辨識的機器學習工具，其應用領域包括影像辨識、資訊探勘、文件分類…等。

而在分類器方面，我們使用 Chang et al. [‘01]所開發的 Libsvm 來作為詩作分類的分類器，在其中的核心函數部分我們選用 RBF，於參數 gamma 和 cost 的部分，則是利用 LIBSVM 中的 grid 程式來反覆測試，找出最佳的 gamma 和 cost，再經由分類器來對訓練資料集來找出最佳的超平面。

在特徵的組合方面，我們共有 7 個特徵，若直接以窮舉的方法來選取特徵組合相當的耗時，故我們使用 Le et al. [‘04]所提出的 Forward Sequential Selection

Algorithm 來做特徵選取，這個方法大致上是先令一個特徵的集合 SF(Selection Feature)為空集合，然後對每一個特徵做分類實驗，挑一個具有最高分類正確率的特徵 F_i 放進 SF 中，接著將 F_i 之外的每一個特徵都放進 SF 中看哪個得到的分類正確率最高來決定第二個要放入 SF 中的特徵，如此反覆直到最後正確率不再增加為止，最後 SF 即可得到一個不錯的特徵組合，我們將演算法詳述如表 18：

Step1：產生一特徵集 $PF = \{F_1, F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9\}$ 和特徵組合 $SF = \{\ \}$。

Step2：對於每一個於 PF 中而不在 SF 中的特徵 F ：

Step2.1：將 F 加入 SF 中。

Step2.2：使用 SF 跑詩作分類，得到正確率 $Eval(SF)$。

Step2.3：若 $Eval(SF)$ 大於最佳正確率 $BestEval$ ，

則讓 $BestF = F$ ， $BestEval = Eval(SF)$。

Step3：若 $BestF$ 不為空，則將 $BestF$ 加入 SF 中。

Step4：直到 $BestF$ 或 $SF = PF$ 時，停止演算法，可得一特徵組合 SF 。

表 18：Forward Sequential Selection Algorithm

利用 Forward Sequential Selection Algorithm 和 SVM 分類器，我們將特徵組合挑選的數據表列，先對單一特徵做實驗找出最佳的分類特徵：

單一特徵	正確率
F1. 詩文詞彙概念	53.00%
F2. 詩文詞彙概念和位置	41.47%
F3. 詩文詞彙	35.48%
F4. 詩題單字詞	64.52%
F5. 詩題雙字詞	53.00%
F6. 詩文共同出現詞彙	33.64%
F7. 詩文共同出現概念	39.17%
F8. 詩題單字詞概念	51.61%
F9. 詩題雙字詞概念	43.32%

表 19：分類結果第一回合

F1~F9 分別為代表 7 個特徵，而根據表 19 的實驗結果我們在第一回合可選出“詩題單字詞”這個特徵，將這個特徵放入特徵集合中，SF = {F4}再進入第二回合的實驗：

雙特徵	正確率
F4+F1	69.59%
F4+F2	63.59%
F4+F3	62.21%
F4+F5	64.52%
F4+F6	61.75%
F4+F7	61.29%
F4+F8	64.98%
F4+F9	61.29%

表 20：分類結果第二回合

在第二回合中，我們選出“詩文詞彙概念”這個特徵，將這個特徵放入特徵集合中，SF = {F1,F4}，再進入第三回合的實驗：

三特徵	正確率
F4+F1+F2	69.12%
F4+F1+F3	66.82%
F4+F1+F5	71.43%
F4+F1+F6	71.89%
F4+F1+F7	69.12%
F4+F1+F8	70.51%
F4+F1+F9	72.35%

表 21：分類結果第三回合

在第三回合中，我們選出“詩題雙字詞概念”這個特徵，將這個特徵放入特徵集合中，SF = {F1,F4,F9}，再進入第四回合的實驗：

四特徵	正確率
F4+F1+F9+F2	70.97%
F4+F1+F9+F3	67.28%
F4+F1+F9+F5	71.89%
F4+F1+F9+F6	71.89%
F4+F1+F9+F7	68.66%
F4+F1+F9+F8	70.97%

表 22：分類結果第四回合

可看出第四回合無法選出具有更高正確率的特徵，故 Forward Sequential Selection Algorithm 停止，以第三回合選出的 SF = {F1,F4,F9}為最佳的特徵組合，此特徵組合的正確率為 72.35%，而我們使用全部的特徵所做的分類正確率為 70.51%，可驗證此演算法可得到較佳的特徵組合。從以上的實驗的結果我們可以看出，詩題的字詞與詩文的詞彙概念做為特徵，對於分類有相當大的幫助，若只對於詩文的表面詞彙來做為特徵，對於分類的結果並不理想。對於特徵的分析結果，我們將其陳述如下：

1. 詩題的字詞和其概念、詩文的詞彙概念做為特徵，對於分類效果較佳。由此結果可知若將表面詞彙轉換成概念後，如同將詞彙做初步的分類，可讓特徵值減少並集中，強化該特徵對於類別的重要性，增加分類的正確性。

2. 用詩文的表面詞彙做為特徵，不如詩題的表面詩彙的分類效果。但對一般語料而言，此結果是相反的，一般語料中內容的詞彙作為特徵相較於標題而言分類效果較佳，歸究原因應是詩作語料中詩文內容的詞彙數較少，故只藉由計算頻率的方式較不易達到在一般語料中特徵選取的效果，反而是詩題較能代表詩作的類別，故效果較佳。

3. 特徵的條件較多，使得符合的詩作數目變少，使分類效果較不理想。因詩作語料的詞彙數不多，若增加特徵條件會使得單一詩作符合特徵的機率降低，單一詩作可被用於分類的特徵減少，分類正確率便會下降。

對於我們所找出最佳的特徵組合 SF = {F1,F4,F9}，我們將其各分類的分類結果統計於表 23：

	C1	C2	C3	C4	C5	C6	小計
訓練資料集個數	115	291	101	205	46	105	863
測試資料集個數	29	72	25	52	12	27	217
分類 True Postive	19	71	14	35	7	11	157
分類 False Negative	10	1	11	17	5	16	60
分類結果	34	98	16	45	8	16	217
Precision	55.88%	72.45%	87.50%	77.78%	87.50%	68.75%	72.35%
Recall	65.52%	98.61%	56.00%	67.31%	58.33%	40.74%	72.35%
F-score	60.32%	83.53%	68.29%	72.16%	70.00%	51.16%	

C1：詠物述志　C2：山水田園　C3：情愛閨怨

C4：贈別思友　C5：邊塞征戰　C6：社會民生

表 23：各分類的分類結果

我們從表 23 中可看出，類別中詩作數較多的類別如山水田園，會有較高的 F-score，而詩作數較少且詩中語意較不明顯的類別如社會民生，F-score 就相對較低，我們將分類錯誤的詩文範例和原因陳列如下：

1. “望南山雪懷山寺普上人：靜宜樵隱度，遠與車馬隔。有時行藥來，喜遇歸山客。”

這首詩被系統判斷的類別為“山水田園”，但正確的類別為“贈別思友”，因他於詩題關於“山”的元素較多，而詩文的特徵也較不明顯，故產生分類的錯誤。

2. “歌舞：遏雲歌響清，回雪舞腰輕。只要君流眄，君傾國自傾。”

這首詩被系統判斷的類別為“詠物述志”，但正確的類別為“情愛閨怨”，因為此首詩詩題為“歌舞”，詩文中的前兩句又包含了“清”、“輕”等形容事物的詞彙，而後半段才較明顯的寫出整首詩的主題為愛慕之情和對人物的描述，故系統不易判別。

第五章 結論

本論文提出並且製作了一個近體詩的詩作分類系統，經過實驗數據的分析可顯示，本系統能夠有效的對於使用者所輸入的詩作來做詩作的風格分類。本論文主題從研究、設計到實作系統，其貢獻分別如下：

1. 建立 1080 首五言絕句的語料庫，在其中有 208 首具有人工標註。
2. 解決詩作中詞彙概念的標記歧義問題。
3. 根據近體詩的特性，提出詩作中的分類特徵，並利用特徵選取的方法來找出對詩作有效的分類特徵。
4. 建置近體詩的分類系統，並使用五言絕句的語料庫來對系統加以測試。

對於本論文未來的研究方向，有下列幾個方向：

1. 利用網路來對現有的辭典做詞彙擴充，建構詩詞的本體論資料庫。因現在我們所使用的語料庫大部分所收錄的詞彙以白話文為主，對於文言文的詞彙其收錄仍不足，故希望能有效地利用網路上的資源，如維基文庫或是相關的詩詞資料庫，利用現有的辭典的架構將其詞彙加以擴充。

2. 可嘗試其他不同的詩作分類特徵，如詩文中的詞性組合，或是加強現有的分類特徵，如詩題是否也可加入切詞器來讓詩題做為特徵的結果更為準確，或是共同出現概念可經由資料探勘的方式來尋找其關聯性。

3. 除了支持向量機外，還可使用其他的分類模型來進行詩作分類，比較其分類結果的正確度。

4. 此分類系統可經由修改後應用於如五言律詩、七言絕句等其他不同的近體詩格式，藉由對不同格式韻文的研究可比較其結果，並分析不同的分類特徵對於不同格式的影響。

參考文獻

[1] Anna Korhonen, Yuval Krymolowski, Nigel Collier (2006), “Automatic Classification of Verbs in Biomedical Texts.” In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Sydney, Australia, pp. 345-352.

[2] Canasai Kruengkrai, Kiyotaka Uchimoto, Jun’ichi Kazama,YiouWang, Kentaro Torisawa, Hitoshi Isahara (2009), “An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging.” In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, pp. 513-521.

[3] Catherine Plaisant, James Rose (2006), “Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces.”Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, Chapel Hill, NC, USA, pp. 141-150.

[4] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines (2001). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[5] Corinna Cortes and Vladimir Vapnik (1995), "Support-Vector Networks", Machine Learning, Vol. 20, pp. 273-297.

[6] Gerard Escudero and Lluís Màrquez and German Rigau (2004),“An Empirical Study of the Domain Dependence of Supervised Word Sense Disambiguation

Systems.”Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpor, Hong Kong.

[7] Huan Liu and Rudy Setiono(1995), “Chi2: Feature selection and discretization of numeric attributes.” In Proceedings of the Seventh International Conference on Tools with Artificial Intelligence, Washington, DA, USA, pp.388-391.

[8] Ian Niles and Adam Pease(2003), “Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology”, In Proceedings of the 2003 International Conference on Information and Knowledge Engineering, Las Vegas, p.p. 23-26

[9] Jyrki Kivinen amd Manfred K. Warmuth (1995),“Additive versus exponentiated gradient updates for linear prediction”, Proceedings of the twenty-seventh annual ACM symposium on Theory of computing, Las Vegas, Nevada, United States, pp. 209-218.

[10] Keh-Jiann Chen, Shu-Ling Huang, Yueh-Yin Shih, Yi-Jun Chen(2005), “Extended-HowNet: A Representational Framework for Concepts” , In Proceedings of IJCNLP-05 Workshop on Lexical Semantic, Jeju Island, South Korea, p.p 1-6.

[11] Le Cuong Anh, Shimazu Akira. (2004), “High WSD Accuracy Using Naïve Bayesian Classifier with Rich Features”. PACLIC 18, Waseda University, Tokyo, pp. 105-113.

[12] Liang-Yan Li, Zhong-Shi He, Yong Yi (2004), “Poetry stylistic analysis technique based on term connections.”, In Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, China, vol.5, pp. 2713- 2718.

[13] Michael Gamon (2004), “Linguistic correlates of style: authorship classification with deep linguistic analysis features”, The 20th International Conference on Computational Linguistics, Geneva, pp. 611-617.

[14] Moshe Koppel, Shlomo Argamon, and Anat R. Shimoni (2003),“Automatically Categorizing Written Texts by Author Gender.”Literary and Linguistic Computing, Volume 17, Number 2, pp 401-412.

[15] Oi Yee Kwong, Benjamin K. Tsou (2005), “Data Homogeneity and Semantic Role Tagging in Chinese.” In Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition, Ann Arbor, Michigan, pp. 1-9.

[16] Roberto Navigli (2006), “Consistent Validation of Manual and Automatic Sense Annotations with the Aid of Semantic Graphs.” Association for Computational Linguistics, Vol. 32, No.2, pp. 273-281.

[17] Xiaojun Wan (2009), “Co-Training for Cross-Lingual Sentiment Classification.” In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, pp. 235–243.

[18] Yang, Y., Pedersen J.P (1997), “A Comparative Study on Feature Selection in Text Categorization”. Proceedings of the Fourteenth International Conference on Machine Learning, Nashville, TN, USA , pp. 412-420.

[19] Yee Seng Chan, Hwee Tou Ng (2007), “Domain Adaptation with Active Learning for Word Sense Disambiguation.” In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, pp. 49-56.

[20] Yong Yi, Zhong-Shi He, Liang-Yan Li, Tian Yu, Elaine Yi (2005), “Advanced studies on traditional Chinese Poetry style identification.” In Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, China, vol.5, pp. 2936- 2939.

[21] 王廼仁，“唐詩之詩風探勘”，國立交通大學，碩士論文，2006 年 6 月。

[22] 古遠清，詩歌分類學，高雄：復文圖書出版社，1991 年 9 月。

[23] 朱我芯,「深秋猿鳥來心上，夜靜松杉到眼前」─華文詩歌情境再現，第五屆全球華文網路教育國際研討會，台北，2007 年 6 月。

[24] 李支舜，高考古詩詞鑑賞與應考指導，上海辭書出版社，2007 年 7 月。

[25] 柯淑津，黃居仁，洪嘉馡，劉詩音，簡卉伶，蘇依莉，“中文詞義全文標記語料庫之設計與雛形製作”，第十九屆自然語言與語音處理研討會，2007 年 9 月，台灣大學，台灣。

[26] 梅家駒等編著，同義詞詞林，臺灣東華書局股份有限公司，1997 年 3 月。

[27] 許清雲，部編大學用書-近體詩創作理論，臺北市：洪葉文化，1997。

[28] 許嘉妮，“詞風與情境判斷專家系統”，國立交通大學，碩士論文，2007 年 6 月。

[29] 陳紹宜，“建構一個中文對聯創作的知識評價架構”，國立交通大學，碩士論文，2010 年 6 月。

[30] 楊昌樺，陳信希，“以部落格文本進行情緒分類之研究”，第十八屆自然語言與語音處理研討會，新竹，台灣，2006 年 9 月。

[31] 羅鳳珠，“植基於中國詩詞語言特性所建構之語意概念分類體系研究”，第九屆海峽兩岸圖書資訊學學術研討會，武漢大學，2008 年 7 月 3-6 日。

[32] 龔霽芃，唐詩分類鑑賞，江西人民出版社，2003 年 12 月 1 日。

附錄

特徵選取結果-詠物述志

詩題單字詞
χ^2 > 6.5 , Feature_Class_Ratio ≥ 0.3
詠、 獵、 竹、 鶴、 蟬
詩題雙字詞
χ^2 > 6 , Feature_Class_Ratio ≥ 0.5
酒店、 從獵、 畫鶴、 詠院 、 春雪
詩題單字詞概念
χ^2 > 5 , Feature_Class_Ratio ≥ 0.5
(Bi18,昆蟲)(Dm04,商店)(Br12,飲料、茶、酒、乳酪)(Ih03,擴大、伸長、收縮)(Hg19,繪畫、製圖、雕刻)
詩題雙字詞概念
χ^2 > 5 , Feature_Class_Ratio ≥ 0.5
(Dm04,商店)(Bh02,花、花卉)(Bi11,禽獸、禽)(Bk13,骨骼、指甲、甲殼、鱗)(Hh04,唱歌、跳舞、演奏)
詩文單一詞彙
χ^2 > 5 , Feature_Class_Ratio ≥ 0.5
借、 惜、 音、 鶯、 酒家
詩文共同出現雙詞彙組
χ^2 > 13 , Feature_Class_Ratio ≥ 0.5
(時、似)(未、風)(曉、聲)(花、家)(寒、長)
詩文詞彙概念
χ^2 > 5 , Feature_Class_Ratio ≥ 0.4
(Di14,風俗、習慣、風氣)(Fa06,拿、撮、揪、挎、夾)(Je05,給予、寄予、加以、冠以)(Bp10,籃、簍、籠、籮筐、簸箕)(Ed14,固定、靈活、生動、平板)(Hc14,建議、討論、決定、制訂)
詩文共同出現雙概念組
χ^2 > 13.1 , Feature_Class_Ratio ≥ 0.5
(Ca24,節氣 ,Eb01,多、少、繁多、稀少)(Ab02,老人、成年人、老小 ,Bn01,建築、房屋)(Di14,風俗、習慣、風氣 ,Ca29,傍晚、晚上、夜間、晝夜)
詩文詞彙概念和其所在位置
χ^2 > 6.5 , Feature_Class_Ratio ≥ 0.4
(Di14,風俗、習慣、風氣 ,2)(Db10,關係、友誼、緣分 ,4)(Gb08,了解、認識、體會、不解 ,2)(Bk11,毛、髮 ,2)(Fa18,穿、戴、佩帶 ,3)

特徵選取結果-山水田園

詩題單字詞
χ^2 > 3 , Feature_Class_Ratio ≥ 0.3
江、 行、 溪、 田、 雲
詩題雙字詞
χ^2 > 7.5 , Feature_Class_Ratio ≥ 0.5
江行、 無題、 雜詠、 藍田、 秋浦
詩題單字詞概念
χ^2 > 1.5 , Feature_Class_Ratio ≥ 0.5
(Be05,海洋、江河、溪澗)(Hg11,揮筆、記錄、留言、附筆)(Ie13,實行、舉行)(Bf02,風、雲)(Hd26,打獵、捕魚)
詩題雙字詞概念
χ^2 > 1.5 , Feature_Class_Ratio ≥ 0.5
(Da24,氣候、收成)(Bn21,塔、亭、閣、台)(Bh08,藥草)(Bn11,路、胡同、橋)(Bf02,風、雲)
詩文單一詞彙
χ^2 > 3.5 , Feature_Class_Ratio ≥ 0.5
山、 江、 野、 峰、 村
詩文共同出現雙詞彙組
χ^2 > 13 , Feature_Class_Ratio ≥ 0.5
(人、江)(山、飛)(來、去)(歸、雲)(江、花)
詩文詞彙概念
χ^2 > 5 , Feature_Class_Ratio ≥ 0.4
(Be04,山、坡)(Hj62,上來、下去)(Be05,海洋、江河、溪澗)(Bo22,船、筏子、飛機)(Bh01,樹木、竹子)
詩文共同出現雙概念組
χ^2 > 13.1 , Feature_Class_Ratio ≥ 0.5
(Be05,海洋、江河、溪澗 ,Bh02,花、花卉)(Bf02,風、雲 ,Be04,山、坡)(Aa01,人、人民、眾人 ,Be05,海洋、江河、溪澗)
詩文詞彙概念和其所在位置
χ^2 > 5 , Feature_Class_Ratio ≥ 0.4
(Be04,山、坡 ,1)(Hj62,上來、下去 ,3)(Be04,山、坡 ,2)(Be05,海洋、江河、溪澗 ,1)(Be03,灘、岸 ,2)

特徵選取結果-情愛閨怨

詩題單字詞
χ^2 > 6.5 , Feature_Class_Ratio ≥ 0.3
怨、詞、女、妓、娘
詩題雙字詞
χ^2 > 6 , Feature_Class_Ratio ≥ 0.5
子夜、女詞、越女、娘詞、嫁娘
詩題單字詞概念
χ^2 > 5 , Feature_Class_Ratio ≥ 0.5
(Hi21,責備、批評、攻擊)(Ah04,父、母、父母、父子)(An07,妓女、男妓、鴇母、龜奴、嫖客)(Bn04,門、窗、門窗)(Id16,掉轉、彎轉)
詩題雙字詞概念
χ^2 > 5 , Feature_Class_Ratio ≥ 0.5
(Eb24,長久、永遠)(Ca29,傍晚、晚上、夜間、晝夜)(Bn04,門、窗、門窗)(Al03,俊傑、人才)(Ca21,月)(Af07,太監、宮女、女官)
詩文單一詞彙
χ^2 > 5 , Feature_Class_Ratio ≥ 0.5
妾、女、郎、嫁、妝
詩文共同出現雙詞彙組
χ^2 > 13 , Feature_Class_Ratio ≥ 0.5
(君、妾)(時、妾)(嫁、有)(生、秋)(畫眉、人)
詩文詞彙概念
χ^2 > 5 , Feature_Class_Ratio ≥ 0.4
(Ah08 , 夫、妻、夫妻)(Hj51 , 戀愛、訂婚、結婚、離婚)(Ab01 , 男人、女人、男女)(Dc03 , 容貌、姿勢、步伐、裝束)(Bp33 , 飾物、首飾)
詩文共同出現雙概念組
χ^2 > 13.1 , Feature_Class_Ratio ≥ 0.5
(Hh04 , 唱歌、跳舞、演奏 , Bq03 , 衣服、上裝、褲子、裙子) (Bq02 , 絲綢、呢絨 , Bp29 , 帳子、蓆子、簾子、帳幕) (Ah08 , 夫、妻、夫妻 , Hj01 , 生活、居住)
詩文詞彙概念和其所在位置
χ^2 > 6.5 , Feature_Class_Ratio ≥ 0.4
(Ah08 , 夫、妻、夫妻 , 2)(Ah08 , 夫、妻、夫妻 , 3) (Ab01 , 男人、女人、男女 , 1)(Bp33 , 飾物、首飾 , 4) (Hj34 , 理髮、梳頭、燙髮、刮臉、打扮 , 4)

特徵選取結果-贈別思友

詩題單字詞
χ^2 > 3 , Feature_Class_Ratio ≥ 0.3
送、 贈、 別、 答、 友
詩題雙字詞
χ^2 > 7.5 , Feature_Class_Ratio ≥ 0.5
送別、 奉和、 雨滴、 首獻、 對酬
詩題單字詞概念
χ^2 > 1.5 , Feature_Class_Ratio ≥ 0.5
(Hi26,贈送、贈答、捐獻)(Ie09,團聚、離別)(Aj01,朋友、恩人、仇人、對手)(Je14,接受、忍受)(Hi11,寫信、匯款、郵寄、拍電報、打電話)
詩題雙字詞概念
χ^2 > 1.5 , Feature_Class_Ratio ≥ 0.5
(Hi06,送別、告別)(Ae13,教師、學生)(Aa05,自己、別人、某人)(Aj01,朋友、恩人、仇人、對手)(Af11,名人、隱士、小人物)
詩文單一詞彙
χ^2 > 3.5 , Feature_Class_Ratio ≥ 0.5
別、 故人、 故鄉、 杯、 春草
詩文共同出現雙詞彙組
χ^2 > 13 , Feature_Class_Ratio ≥ 0.5
(何、歸)(君、別)(人、去)(別、遠)(故人、知)
詩文詞彙概念
χ^2 > 5 , Feature_Class_Ratio ≥ 0.4
(Ie09,團聚、離別)(Aj01,朋友、恩人、仇人、對手)(Hi02,訪問、進見、探望)(Cb15,本地、外地、家鄉)(Hi06,送別、告別)
詩文共同出現雙概念組
χ^2 > 13.1 , Feature_Class_Ratio ≥ 0.5
(Hi02,訪問、進見、探望 ,Hf07,離開、返回)(Ie09,團聚、離別 ,Hi02,訪問、進見、探望)(Be05,海洋、江河、溪澗 ,Hi06,送別、告別)
詩文詞彙概念和其所在位置
χ^2 > 6 , Feature_Class_Ratio ≥ 0.4
(Cb15,本地、外地、家鄉 ,4)(Aj01,朋友、恩人、仇人、對手 ,2)(Hi02,訪問、進見、探望 ,3)(Aj05,主人、客人、賓主 ,3)(Ie09,團聚、離別 ,1)

特徵選取結果-邊塞征戰

詩題單字詞
χ^2 > 6.5 , Feature_Class_Ratio ≥ 0.3
塞、 射、 僕、 軍、 蕃
詩題雙字詞
χ^2 > 6 , Feature_Class_Ratio ≥ 0.5
塞下、 張僕、 射塞、 平蕃、 蕃曲
詩題單字詞概念
χ^2 > 5 , Feature_Class_Ratio ≥ 0.5
(Hd17,開山、墾荒、燒荒、種地)(Bp29,帳子、蓆子、簾子、帳幕)(Hj03,消磨、度過、經歷)(Hj10,分家、過繼、承繼)(Hb11,侵略、併吞、騷擾)
詩題雙字詞概念
χ^2 > 5 , Feature_Class_Ratio ≥ 0.5
(Hh06,溜冰、游泳、下棋)(Bn18,營房、堡壘、烽火台)(Ak02,俠客、勇士、懦夫、懶漢)(Dh01,神、鬼、妖、魂、妖魔鬼怪)(Dk28,音樂、歌、曲)
詩文單一詞彙
χ^2 > 5 , Feature_Class_Ratio ≥ 0.5
將軍、 箭、 營、 刀、 隴頭
詩文共同出現雙詞彙組
χ^2 > 13 , Feature_Class_Ratio ≥ 0.5
(城、聲)(邊、度)(里、馬)(鳴、草)(高、刀)
詩文詞彙概念
χ^2 > 5 , Feature_Class_Ratio ≥ 0.4
(Ae10,軍官、將士、軍人、士兵)(Bn18,營房、堡壘、烽火台)(Bo29,弓、箭、矛、盾、劍)(Di11,軍隊、戰爭)(Di09,政策、制度、所有制、體制)
詩文共同出現雙概念組
χ^2 > 13.1 , Feature_Class_Ratio ≥ 0.5
(Bo29,弓、箭、矛、盾、劍 ,Fb01,走、跑)(Ae10,軍官、將士、軍人、士兵 ,Bo09,刀、剪、斧、刃)(Di11,軍隊、戰爭 ,Ka28,仍然、尚且)
詩文詞彙概念和其所在位置
χ^2 > 10 , Feature_Class_Ratio ≥ 0.4
(Ae10,軍官、將士、軍人、士兵 ,2)(Di11,軍隊、戰爭 ,1)(Bo29,弓、箭、矛、盾、劍 ,1)(Ak02,俠客、勇士、懦夫、懶漢 ,1)(Bn18,營房、堡壘、烽火台 ,1)

特徵選取結果-社會民生

詩題單字詞
χ^2 > 3 , Feature_Class_Ratio ≥ 0.3
歲、 感、 城、 州、 哭
詩題雙字詞
χ^2 > 7.5 , Feature_Class_Ratio ≥ 0.5
中作、 守歲、 正朝、 懷古、 有感
詩題單字詞概念
χ^2 > 1.5 , Feature_Class_Ratio ≥ 0.5
(Ca18,年)(Di22,責任、負擔)(Hh03,攝影、錄音、放映、廣播)(La04,打擾、勞駕、抱歉)(Bk11,毛、髮)(Hm01,檢舉、控告、訴訟)
詩題雙字詞概念
χ^2 > 1.5 , Feature_Class_Ratio ≥ 0.5
(Ca11,過去、近來)(Bn23,皇宮、祠堂、佛殿)(He13,積累、花費、浪費、節省)(Ef01,太平、安定、動蕩、混亂)(Cb25,城市、集鎮、鄉村)(Bn22,墳墓、墓穴、碑)
詩文單一詞彙
χ^2 > 3.5 , Feature_Class_Ratio ≥ 0.5
白頭、 鬢、 郡、 吏、 愁心
詩文共同出現雙詞彙組
χ^2 > 13 , Feature_Class_Ratio ≥ 0.5
(空、悲)(日、年)(葉、黃)(催、年)(愁、來)
詩文詞彙概念
χ^2 > 5 , Feature_Class_Ratio ≥ 0.4
(Ca18,年)(De01,性格、品行、道德、作風)(Hi29,感謝、報答)(Ai01,鼻祖、前人、今人、後人)(Da14,成就、功勞、過失、功過)
詩文共同出現雙概念組
χ^2 > 13.1 , Feature_Class_Ratio ≥ 0.5
(Ca18,年 ,Ca19,四季、春、夏、秋、冬)(Dn04,基數、序數 ,Da14,成就、功勞、過失、功過)(Ca18,年 ,Ga01,高興、悲傷、憂愁)
詩文詞彙概念和其所在位置
χ^2 > 5 , Feature_Class_Ratio ≥ 0.4
(Ca18,年 ,1)(Hc01,治理、辦理、處理 ,4)(Dk27,詩、詞、賦、令、曲、對聯 ,4)(Ai01,鼻祖、前人、今人、後人 ,1)(Da14,成就、功勞、過失、功過 ,4)