國立交通大學

理學院網路學習學程

碩士論文

唐詩之詩風探勘

Style mining for Tang Poetry

研究生：王廼仁

指導教授：曾憲雄　教授

中華民國　九十五　年六月

唐詩之詩風探勘
Style mining for Tang Poetry

研 究 生：王迺仁　　Student：Nai-Jun Wang

指導教授：曾憲雄　　Advisor：Shain-Shyong Tseng

國 立 交 通 大 學

理學院網路學習學程

碩 士 論 文

A Thesis

Submitted to Degree Program of (E-Learning) (Applied Science and Technology)

College of Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Degree Program of (E-Learning) (Applied Science and Technology)

June 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年六月

唐詩之詩風探勘

研究生：王迺仁　　指導教授：曾憲雄 博士

國立交通大學理學院碩士在職專班

摘要

詩是中國文化偉大的文學創作，尤其在唐朝時最為盛行。詩為韻文之一種，講究音韻諧和，所用文字是古漢語，與現在的白話文差異很大，且漢語文字多有同義與歧義的問題。對詩作文字進行風格探勘，字詞的概念繁瑣，由專家學者來分析處理已是不容易，雖然文學評論者將唐詩依作者區分為不同風格的詩作，但區分的條件或規則並不明顯，要從詩句文字中找出詩風分類的規則更是困難。

本論文運用資料探勘技術中關聯規則探勘盛唐時期文人詩作，找出詩人創作詩時所偏好使用的名詞組合規則。第一階段是**名詞擷取**：文人在詩中常描寫景觀及事物變化用詞大多為名詞，所以從詩作擷取出名詞作為基本的分析資料。第二階段是**名詞概念歸納(concept generalization)**：本論文提出唐詩名詞概念階層(concept hierarchy)將詞義概念繁雜的名詞概念歸納成概念精簡的名詞類別，建置成唐詩名詞類別集。第三階段是**名詞使用差異性分群(clustering)**：比較詩作名詞類別使用的差異性(dissimilarity)來分群，找出文人風格詩作群集。第四階段是**名詞使用關聯規則探勘(association rule mining)**：利用關聯規則探勘分析風格詩作群集使用名詞類別的組合，依可信度(confidence)及支持度(support)找出詩人詩詞創作名詞使用的風格規則。從分析的結果發現，實驗方法將王維詩作分為邊塞詩群及山水詩群等，可與詩詞專家的評論相互佐證。並將文學評論者所整理之山水派唐詩經名詞擷取、名詞概念歸納後，與實驗結果比較，與山水詩群相近且名詞使用的規則也相同。

此實驗方法可以客觀找出詩作因名詞使用不同的風格判別規則，供詩詞研究者分析或歸類詩作，或供學習者了解詩作中名詞使用所表現的風格意涵。

關鍵字：唐詩、風格探勘、概念階層、分群、關聯規則探勘

Style mining for Tang Poetry

Student: Nai-Jun Wang Advisor: Dr. Shian-Shyong Tseng

Degree Program of E-learning

College of Science

National Chiao Tung University

Abstract

The Chinese poetry is the greatest Chinese literature creation especially in Tang Dynasty. It is a kind of rhymed article which is concerned with the harmonious rhyme. It is a challenge to analyze Tang poetry because the ancient Chinese language used in the poem is different from the modern Chinese language. Besides, there are many synonyms and ambiguity in Chinese vocabulary. Therefore, the analysis of the poetry style is difficult for Tang Poetry.

In this thesis, we propose a Data Mining approach to discover the style of the poetry for the poet. Firstly, in **the noun retrieval process**, the nouns in the poem are treated as the features of the poetry style and retrieved for data analysis. Secondly, in **the noun concept generalization**, we transform the nouns by Tang noun concept hierarchy to the corresponding concept. We construct the Tang concept hierarchy table from the transformed results. Thirdly, in **the style clustering of poetry**, we cluster all the poetry by comparing the nouns dissimilarity of poetry. Finally, in **the noun association rule mining**, we apply the association rule mining to discover the noun association of poetry. During these processes of the style mining for Tang poetry, we obtained the noun sets of poetry and the noun association rules. These results are verified by the experiment, and we can distinguish the poetry style from these association rules.

In the future, the noun association rules of Tang poetry that are discovered objectively by the style mining can support the researcher for further research.

Keywords: Tang Poetry, style mining, concept hierarchy, clustering, association rule mining

誌謝

這兩年來的學習，是我寶貴的人生經歷，每當回想，每一個走過之足跡，都讓我感動。在學習的歷程中，最感謝的是恩師 曾憲雄博士，幫助我在研究領域上培養獨立思考、自我突破及掌控與解決問題的能力。其次，感謝專班 莊祚敏、 林登松、 黃大原、 陳明璋、 袁媛等博士在課程設計及教學過程中，一直引領著我前進，不斷地給予我鼓舞與激勵。也感謝專班助理呂孟嫻小組及助教們的協助，讓我能順利地完成在校學習、研究時階段性的過程。

在論文口試期間，感謝 莊祚敏、 楊錦潭和 曾秋蓉等博士提供在研究方法與內容闡述上的建議，讓我深刻體會研究過程比成果更為豐碩，所給予的精闢見解與建議，讓本篇論文更加完備。

求學生涯中，感謝實驗室的學長們蘇俊銘、楊哲青、王慶堯、林順傑、曲衍旭、李威勳等，適時地給予協助與建議，讓我能完成研究及論文的寫作。專班及實驗室研究伙伴，感謝你們在我研究歲月中所留下的美好回憶，這份誠摯的情誼，將永懷於我心。

對於家人的支持，這一份感謝，非言語所能形容。在面對工作、課業與研究壓力時，家庭是我最大的動力，你們總是陪伴在我身邊。玉慧殷勤的盼望、無怨無悔的付出，孩子品翰、品倫天真的笑臉，總是我持續研究的精神支柱，為我打氣加油，也在不安、沮喪時給我安慰，謝謝你們。

僅將本篇論文的完成，獻給每一位給予我幫助及支持我的人。

目錄

圖目錄

表目錄

演算法目錄

一、 緒論

中華數千年來的文化全由文人紀錄下來，在文學創作中除了保留豐富的歷史文化，也展現古人的智慧。詩是中國古典文學的瑰寶，尤其在唐朝時，政治開明，科舉考試以詩取士，詩的創作受到了提倡。詩篇中不論一句五字或七字等，皆講究音韻上的和諧，文人以其智慧選用文字或詞彙來描寫景物的變化或嘲諷政局等，內容蘊含大量知識與風格。

由於時代變遷，現代人對古典文學的認識不多，或深入研究的門檻變得更高。然而，近年來興起數位典藏計劃，將古典書籍電子化，除了能永久保存之外，更可以有效提升知識的累積、傳承與運用。結合網際網路，提供使用者能在線上查詢與閱讀，使古典文學得以推展。

唐朝詩作中，詩人為講求文字內涵優美及音韻的協調，使用不同字詞來描寫相同事物，造成中文字詞的概念多且繁瑣，由專家學者來分析處理得知詞義已是不容易，且文學評論者對某些唐詩加以分析或評鑑，將唐詩依作者區分為山水田園、邊塞等不同風格的詩作，但其區分的條件或規則並不明顯或只是隨評論者自身對某些詩作喜好而來區別，要從詩句文字中找出詩風分類的規則更是困難。

本論文運用資料探勘技術分析盛唐時期文人詩作，找出詩人創作詩時所偏好使用名詞的組合規則作為風格判別的依據。唐詩詩風探勘過程可分為四個階段，第一階段是**名詞擷取**：文人在詩中常描寫景觀及事物的變化，為了表達自己的情感，多針對特定的景物來描寫，而景物用詞大多為名詞，所以本文從詩作擷取出名詞作為風格分析的基本資料。第二階段是**名詞概念歸納**：文人擅長使用不同的語詞來描寫同一事物，但電腦無法判斷詞義，造成分析詩作用詞的關聯性時的困難。本論文提出唐詩名詞概念階層(Tang Poem noun concept hierarchy)將詞義相同或近似的詞彙歸為同一類。從詩中所擷取的名詞對應至此概念階層，能使詞義概念繁雜的名詞概念歸納(concept generalization)成概念精簡的名詞類別，建置成唐詩名詞類別集。第三階段是**名詞使用差異性分群**：詩人創作詩作雖有其主要的風格，

也會有多樣風格的創作，主要表現在其詩作中所使用名詞的異同。先利用詩作名詞使用的差異性(dissimilarity)來分群(clustering)，找出文人風格詩作群集。第四階段是**名詞使用關聯規則探勘**：利用關聯規則探勘(association rule mining)分析詩人風格詩作群集中使用名詞類別的組合規則，設定可信度(confidence)及支持度(support)找出詩人創作時名詞使用的風格規則。

從實驗的結果分析發現，王維詩作可分為邊塞詩群、仕官詩群、田園詩群、山水詩群及思鄉懷人詩群。邊塞詩群中多動物、社會、政法、地貌等類與空間及機具類名詞組合規則；仕官詩群中多時間、建築物、植物、數量、單位及空間與材料等類名詞組合規則；田園詩群多用建築物、植物、文教、時間與空間等名詞組合規則；山水詩群則多地貌、建築物、植物、氣象、時間及空間等類名詞組合規則；思鄉懷人詩群使用人、建築物、植物及時間與空間等類名詞組合規則，可與詩詞專家的評論相互佐證。

並將詩詞專家評論的唐朝山水詩作群，經名詞擷取、名詞概念歸納後，取各類名詞使用率的平均值作為群中心，與實驗中各風格群群中心比較差異性，作準確性評估，與實驗山水詩群非常接近。並分析名詞使用的關聯規則，也與實驗分析相近。

本論文的貢獻：

一、結合詩詞專家的格律知識，建構格律斷詞模組，從詩作中擷取出名詞。

二、將唐詩詩作中所使用的名詞依詞義並參考同義詞詞林架構，建置唐詩名詞概念階層。

三、對唐朝文人詩作進行風格探勘：應用分群技術區分詩人風格詩作群集，再針對詩作名詞使用，使用關聯規則探勘分析風格詩作群名詞使用的組合規則。

本論文架構：第一章為緒論，第二章為研究背景，第三章介紹唐詩詩風探勘的方法，第四章是詩風探勘的實驗與分析，第五章為結論與展望。

二、研究背景

針對唐朝詩作進行風格探勘，必須具備古漢語的知識，且應用本體論來表示字詞的詞義，進而探討詩作用詞的關聯性。在本章中介紹詩的結構與特性、本體論及前人對詩詞文字和風格的相關研究。

2.1 唐詩格律與詩風簡介

中國詩詞是中國文學最具特色的創作，不僅文字優美，也講究語言音律。韻文是使用韻語的文學作品，而韻語是在文字用詞上使用音素相近的語詞，雖然文字聲調高低長短不同，但因具備相同或相近語音的語詞，連綴起來便會產生音韻諧和且令人深感共鳴。詩是韻文的一種，其中包含詩經、五言詩、七言詩、樂府、漢詩、律詩、絕句等形式。詩中文字多用是古漢語，漢語文字都是單音詞，但到了唐朝，詩作所用的文字詩大量使用雙音詞（雙字詞），故稱為近代漢語與古代漢語區隔，而現今所使用的白話文，則稱為現代漢語。[1]

千百年來，中國詩詞經古文人不斷地研發及創作，不僅在文字用詞上的考究，更要求能有音韻上的變化及整首詩在吟唱時格律的合諧，這些音韻與格式上的要求，稱為格律。以絕句詩的格律規則為例來說明：

(1) 字數及句數：每首絕句詩的句數為四句；若為八句，則稱為律詩；若是八句以上且每兩句排排對仗，則稱為排律或長律。

(2) 字詞的音韻：每一句末字稱為韻腳，而押韻是指韻腳所用的韻要相同。第二、四句一定要押韻，一韻到底，不能換韻。第一句押韻者，稱為入韻；不押韻者，稱為不入韻。且每首詩的韻腳不能重複。

(3) 平仄的安排：要求單一詩句文字音韻平仄交錯，第一、二句與第三、四句對句音韻平仄對立，第二、三句鄰句音韻平仄相黏。

所以，只要知道第一句第二字的平仄，依據以上格律規則分析就可以知道這

一首絕句詩的格律。若第一句的第二字為平，則這首詩是「平起」的格律；若第一句的第二字為仄則是「仄起」的格律。[2]

漢語詩的音律是長短調遞用，且都是以兩個字（音）組成一個節奏單元，也稱為音節，節奏的重點落在音節的第二音上，稱為音節點或節奏點。當詩句字數為偶數，能被 2 整除，得出整數個音節，如四言、六言詩。

《酬諸公見過，王維》

屏居藍田，薄地躬耕。

● ● ● ● （●表示該字為音節點）

《田園樂，七首之一，王維》

厭見千門萬戶，經過北里南鄰。

● ● ● ● ● ● （●表示該字為音節點）

當詩句字數為奇數，便不能被 2 整除，結果句子最末字單獨成為一個音節，如五言、七言詩。例如：

《鹿柴，王維》

空山不見人，但聞人語響。

● ●● ● ●● （●表示該字為音節點）

《輞川閒居贈裴秀才迪中，王維》

渡頭餘落日，墟里上孤煙。

●● ● ●● ● （●表示該字為音節點）

《奉和聖製從蓬萊向興慶閣道中留春雨中春望之作應制，王維》

渭水自縈秦塞曲，黃山舊繞漢宮斜。

● ● ●● ● ● ●● （●表示該字為音節點）[3]

四、六言詩句中語氣以兩字為一頓（音節），而第二字、第四字及第六字為句子的音節點。五、七言詩除了第二字、第四字及第六字為音節點外，句未第五字或第七字自成一音節點。吟詠詩作時，偶數字數的詩句音調及語氣在句子末字能完整結束，但詩人常為了讓句子語氣的變化或詩中意念的延伸，而多出一字，所以唐朝詩作多為一句五字（五言）或七字（七言）。文人有時為了強調語氣或使語調變化，會將音節點後移在句子的最後一字，而在第末三字的第一字產生一頓，

使得詩句音韻多了變化。

詩人創作時，考慮配合詩句音節，產生二字一斷（一個音節）的規則，每一個音節中，文字用詞多為雙字詞或由兩個單字詞所組成，而句末尾三字則有三字詞或前二後一、前一後二的字詞變化。本文利用此規則建置成格律斷詞模組，協助系統在名詞擷取時能切分出正確的詞彙。

文人創作因生活或個性不同，造成詩作風格多樣，如陶文鵬於〈明月松間照詩佛：王維作品賞析〉書中所介紹盛唐時期的詩人王維。王維，字摩詰，生於武后長安元年（西元 701 年），卒於肅宗上元二年（西元 761 年），蒲州人。父親早喪，與母親及弟妹五人過著較清寒的生活。母親崔氏一生篤信佛教，王維深受她的薰陶，也虔誠奉佛，佛教的思想及思維方式，對王維的生活與創作產生極大的影響。政治生涯中，目睹朝廷奸邪專橫、政治腐敗，內心不滿，也不願同流合污，飽嘗生活的艱辛與世態的炎涼，於是採取亦官亦隱、全身避禍的生活方式。[4]

王維精通音律、擅長書畫，詩作文字不僅優美，更兼具濃厚的畫意和音響之美。詩的題材廣泛，風格多樣，邊塞詩、遊俠詩和詠史詩是早期的作品，在朝為官時也寫應制詩、酬和詩，另有不少的送別詩、贈友詩和思鄉懷人詩，於半官半隱時的田園詩和山水詩，及晚年的禪詩等，其中以山水詩的創作數量最多。詩歌體裁有五言古詩、七言古詩、五言律詩、七言律詩、五言絕句、七言絕句、雜詩、六言詩等。

綜上所述，要將王維等詩人多樣風格的詩作分群歸類是件不容易的事，本論文依專家建議從文字使用的角度，預期分為六大詩群：描寫大漠風光、歌頌將士奮勇殺敵的邊塞詩群，內容包含邊塞詩、遊俠詩和詠史詩等；擔任官職時歌功頌德及生活描寫的仕官詩群，內容包含應制詩與酬和詩；抒發離恨鄉愁、或情感真摯的思鄉懷人詩群，包含思鄉懷人詩、送別詩及贈友詩；刻畫山水勝景的山水詩群；勾勒農村寧靜風光的田園詩群；及寓寄佛教禪宗教義的禪詩群。

2.2 格律知識庫與格律檢查系統

羅鳳珠等人所建置【倚聲填詞】格律自動檢測索引教學系統[5]，希望借助電

腦快速準確的檢索功能，將每一詞牌的格律輸入電腦，再輔以詞韻之韻字，各韻字之例詞、例句。讓使用者填詞時，可以透過系統檢索詞牌、詞韻的功能，選定作詞所用的詞牌、詞韻；填詞時，系統能自動檢測不合格律的字句，提醒使用者修改，同時也可以檢索例詞、例句及唐宋詞人作品，作為修改之參考，對於詞的教學、研究與習作有莫大的幫助。然而，使用者須具備基本詩詞的知識，例如聲調、詞牌等，雖有整合檢索工具，但對初學者仍有相當的門檻，不易跨入學習。

蕭斯聰等人提出近體詩系統的理論架構[6]，近體詩格律由於具有明顯的規則性，如果利用「規則式專家系統」推論的機制進行近體詩文字平仄聲調、用韻格式等的檢查，可以克服初學者在使用詩詞輔助學習系統上的不便，有效地幫助使用者跨越「中國古典詩詞」學習上的門檻。

楊哲青等人也提出近體詩專家系統[7,8]，在系統架構中需要一個有推論功能的引擎，以及能處理詩句文字的詩詞系統。依據使用者所創作之詩句，由詩詞系統查詢古字聲韻資料庫，取得詩句文字的平仄及韻目後，再交由推論引擎依據專家系統知識庫中的規則進行格律推論，以判斷是否合乎近體詩之格律要求。而所提出的中國詩詞專家系統，其中對於絕句詩的規則庫只是將文人所整理出的十種絕句詩譜格律建置成規則，系統在判斷時，只能判別詩作符合或不符合詩譜格律，且有一字多韻的問題待解。

王迺仁等人分析近體詩的絕句詩格律發現[9]，詩詞的格律知識是有架構及層次的，利用知識物件模型來加以表示，將近體詩的格律知識導入專家系統，便可用來保存中國近體詩的知識及作為教學使用，使近體詩知識不致有失傳之疑慮。藉由詩句格律規則類別規劃、詩句格律狀態分析、格律規則表示及規則驗證等絕句詩知識擷取的過程，建置成絕句詩知識庫(Knowledge Base)，以作為專家系統之知識推論的來源依據。結合專家系統的推論功能及網際網路的服務，讓詩詞創作者或學習者，輸入詩句後，系統能即時分析詩句的格律，並回應格律特徵給使用者，作為創作近體詩的輔助工具。而系統僅只於分析絕句詩的格律，而建置成絕句詩格律知識庫。

2.3 詞彙分析和檢索

俞士汶與胡俊峰[10]利用統計的方法對唐宋詩語料進行詞彙擷取，應用「共現度」、「結合強度」等統計參數的計算方法，並與傳統的「互信息」方法進行了比較。詞彙的提取與分析當然離不開對詞義的理解，領域專家對詞義的理解自有優勢。但是，許多現代漢語中的詞（如：可以、上學等）在古詩詞中還不是詞，而古詩中的一些詞（如：弱冠、小槽等）由於社會環境的變化，在現代漢語中已經很少這樣使用。僅僅依靠領域專家是很難進行大規模調查與分析的。統計手段的引入，就能夠有一個相對客觀的標準來判定古漢語中的詞。也提出詩人在創作時的用詞特徵是詩人風格分析的重要參考。

綜上所述，利用統計方法可以快速且大量分析詩詞文字的組合，得出古漢語的詞彙，但詩詞字詞因字數受限，文字簡明而意涵豐富，不能只用統計方法來提升字詞分析的準確性。應結合詩詞專家的知識，輔以字詞分析的技術，才能建立詞義明確標示的詞彙庫。

2.4 本體論(ontology)

本體論(ontology)為用來描述與定義各種知識的語言，以便達到知識分享共用的目的，以語言資訊轉換、概念階層的連結、詞義的區分與詞義關係的連結等技術為主要核心技術，讓文字的處理不只提供查詢，也紀錄下更多文字內容細節及結構。SUMO (Suggested Upper Merged Ontology, 建議上層共用知識本體)[11]是結合英文 WordNet 架構所建置的本體論上層架構，中文化的部份是由中央研究院歷史語言研究所維護，加上台灣地區的語言使用的經驗，參考其 SUMO 概念架構，結合不同領域的知識本體，提供跨領域的資訊檢索，並可衍生出其他特殊領域的知識本體。但 SUMO 把概念視為節點，概念架構分類極為詳細，概念與概念間不單是上下隸屬的階層關係，也可以相互參考引用，在詞語的分類上沒有明確的界定，部份詞語間無法區隔及歸類，使得資料探勘上不易處理使用。

同義詞詞林[12]主要選收現代漢語詞彙，也收錄部份常見的古語詞，根據漢語的特點與實用的原則，依詞義歸類詞彙，並考慮詞類，共分成 12 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，表 1為第一至四大類及以下中類目錄編號及名稱表，同義詞詞林共收錄詞彙約七萬個。

表 1　同義詞詞林大、中類目錄編號表（名詞部份）

編號	目錄名	編號	目錄名	編號	目錄名
A	人	Bd	天體	Da	事情、情況
Aa	泛稱	Be	地貌	Db	事理
Ab	男女老少	Bf	氣象	Dc	外貌
Ac	體態	Bg	自然物	Dd	性能
Ad	籍屬	Bh	植物	De	性格、才能
Ae	職業	Bi	動物	Df	意識
Af	身份	Bj	微生物	Dg	比喻物
Ag	狀況	Bk	全身	Dh	臆想物
Ah	親人、眷屬	Bl	排泄物、分泌物	Di	社會、政法
Ai	輩次	Bm	材料	Dj	經濟
Aj	關係	Bn	建築物	Dk	文教
Ak	品性	Bo	機具	Dl	疾病
Al	才識	Bp	用品	Dm	機構
Am	信仰	Bq	衣物	Dn	數量、單位
An	丑類	Br	食品、藥品、毒品		
B	物	C	時間和空間		
Ba	統稱	Ca	時間		
Bb	擬狀物	Cb	空間		
Bc	物體的部分	D	抽象事物		

陳書磊[13]藉著觀察詩中文字組合的必然性，定義部份字詞使用及字組合成詞彙的規則，可以把詩作中每句所使用的字詞標示，並建置成漢語詩本體論的例句，而提供檢索及相似度的計算。主要參考的工具書是同義詞詞林，先將詩句所使用的單字先對應至同義詞詞林大類的架構，再依漢語使用的特性所建置字詞的組合規則，分析詩句文字，並標示出其語意架構。系統並提供詩句語意結構的相似度比較，供使用者輸入相關詞彙，找出符合且依相似度高低建議的詩句。在檢索功能上雖然可以針對詞彙來進行，但若格式與詩句格式架構相差太大，則可能會檢索出錯誤的例句或無法檢索。且所建置字詞語意處理規則，也需要驗證並增加。

同義詞詞林收錄以現代漢語常見的一般語詞為主，也有一些古語詞，所以，可以作為唐詩詞彙庫的參考工具書。且同義詞詞林對詞義分類架構明確，從大類到小類以至於以下的詞群，概念階層清楚且完整，在本論文中依據同義詞詞林對名詞詞義的分類階層來，建置唐詩名詞概念階層，供歸納繁多的詞義。

2.5 資料探勘

資料探勘技術主要是在擁有大量資料的資料庫中，找尋隱藏的資訊，並利用視覺化或非視覺化的方式呈現，顯示資料各種屬性的關聯性。而此類資訊不同於一般的統計資料，能讓領域專家判讀並解釋資訊中所代表的現象或意義，進而提供修正系統的意見。

在資料探勘研究領域中，依據資料分析應用需求，其方法可分為三大類：關聯規則探勘(association rule mining)、分類與預測(classification and prediction)及分群分析(clustering analysis)。

其中，關聯規則探勘可從資料屬性的數量關係上，找出重複機率較高的組合模式，作為資料相互之間關聯性的規則，也就是找出一組屬性或事件與另一組屬性或事件結合起來的規則。例如：奶油、牛奶→麵包(30%，2%)，意義為購買奶油和牛奶的顧客，同時間也購買麵包的可信度(confidence)為 30%，此商店所有購買行為中，同時購買奶油、牛奶及麵包的支持度(support)為 2%。可信度及支持度可表示為下列式子。

$$confidence(A \Rightarrow B)=P(B|A) \quad (1)$$

$$support(A \Rightarrow B)=P(A \cup B) \quad (2)$$

關聯規則探勘最著名的演算法為 Apriori 演算法，這個演算法可分成兩個步驟：第一步驟是找出頻繁項目集(frequent itemset)，其次數不低於所訂定之最低支持度(minimum support)；第二步驟是從頻繁項目集中，找出不低於所訂定之最低可信度(minimum confidence)的關聯規則。[14,15]

分群分析可從大量資料中，藉著資料間屬性的相似性(similarity)或差異性(dissimilarity)，整合成相似資料群集。較著名的分群演算法是 *k* 均值法(*k*-means)，可依資料屬性的差異性來分群，在第三章中將會有針對詩作差異性分群的詳細討論。

文學作品通常是人對述事或寫景的邏輯思維表現，詩也不例外，且詩作中文字、句數有限，字詞間呈現高度的關聯性和次序性。且詩人創作風格多樣，不同的景物描寫，詞彙使用自然也就不同，表現出詩作風格的差異性。本論文提出運用分群的技術，先將詩人眾多詩作依名詞使用的差異性分群，並利用關聯規則探勘技術找出風格詩作群集中名詞使用的關聯規則。

2.6 詩風探勘

楊哲青[8]應用資料探勘技術分析詩人資料及詩作，提出建置詩作風格知識的方法。先依照詩詞專家的分析與意見，建立蘇軾生平資料，並對詩作依風格予以分類。針對近體詩格律並結合專家知識，整合成格律規則集，對詩作進行斷詞、斷句，利用詮釋資料技術(Metadata Technology)建立詞彙知識庫。運用關聯規則探勘分析詩作詩句平仄及韻目的使用關聯規則，萃取出詩人潛在創作風格。將所建置的知識庫搭配詩作系統，可達成協助使用者詩作創作的目的。在風格知識上只針對詩人詩句聲韻的使用來分析。

易勇[16]等人運用機器學習(machine learning)技術分析宋詞的風格，主要是分析宋詞中所使用單字詞的出現率，建立豪放或婉約風格分類模式。因宋詞中虛詞使用甚多，會影響到分類的準確性，所以，利用基因演算法(genetic algorithm)找出

影響宋詞風格的主要特徵文字群集，再運用 Naïve Bayesian 分類法建立宋詞風格分類的模式。但詩詞常引用典故或使用多字詞，且詞義與其組合的單一文字不同，若能考慮詞彙的使用，更能建立精確的分類模式。

王迺仁[17]等人在文學與資訊國際會議曾提出近體詩階層式概念，將詞義相同或近似的詞彙歸為同一類，使概念繁雜的名詞整理成概念精簡的名詞類別。近體詩階層式概念是將詞林典腋中古人描寫事物的詞彙及唐詩中部份名詞參考 SUMO (Suggested Upper Merged Ontology)及同義詞詞林的架構而自行建置而成，有階層式的架構將事物的概念由大至小細分下來。利用近體詩階層式概念將唐詩所使用到的詞彙概念歸納(concept generalization)，進一步利用關聯規則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。因近體詩階層式概念是自行建置，階層架構不完整且分類規則因人為判別而不一致，致使資料分析準確率不高。

綜合以上的研究，具有豐富意義的詞彙是詩作風格分析的重要關鍵。因此，本論文分析唐朝詩作所使用的名詞，依詞義建置唐詩名詞概念階層，並使用資料探勘的技術，針對唐朝詩人王維在全唐詩中所收錄的詩作，進行風格探勘。

三、 唐詩之詩風探勘

在本論文中，利用資料探勘(Data Mining)的技術分析唐朝文人詩作文字或字詞的使用，找出詩人創作時字詞使用的規則。研究架構可分為資料前處理(data preprocessing)與資料探勘等兩大部份，如圖 1所示。資料前處理過程中包含兩個階段：1、名詞擷取：將唐朝詩人詩作所使用的名詞擷取出來。2、名詞概念歸納：將詩作所使用的名詞依詞義分類，建置成唐詩名詞概念階層(concept hierarchy)，將繁雜的字詞意涵歸納成精簡的名詞類別。資料探勘過程中也包含兩個階段：1、名詞使用差異分群：依名詞使用的差異性並利用分群(clustering)的技術將詩人詩作歸類成名詞使用風格相近的詩作群集。2、名詞使用關聯規則探勘：使用關聯規則探勘(association rule mining)分析同一風格詩作群詩作名詞使用與風格的關聯規則。在本章各小節中作詳細的說明。

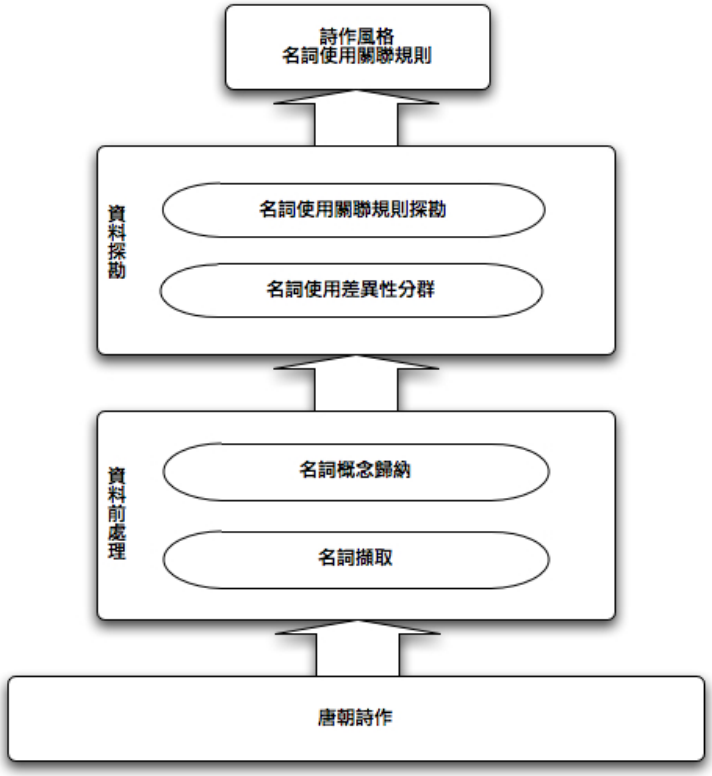

圖 1　研究架構

3.1 名詞擷取

唐朝文人將其知識及情感表達在所創作的詩作中，以景色、事物變化的描寫來表達對時事變化的感傷與情懷。在詩作中，描寫景物或事物的字詞就是以名詞為主，且名詞類詞彙意義實在，能單獨充當句子的主要成分。所以，本論文中主要的重點是分析詩人詩作，找出名詞使用與詩人創作風格的關係。

要將詩作中的名詞擷取出來，必須要有大量的唐詩詞彙，來分析詩作文字，並篩選出名詞。依據唐朝詩體文字使用格律要求音調變化，提出唐詩格律斷詞模組，將斷出來的語詞查詢唐詩詞彙庫，並擷取出詩作中的名詞，整理成唐詩名詞集，做為分析的基本資料，如圖 2所示。

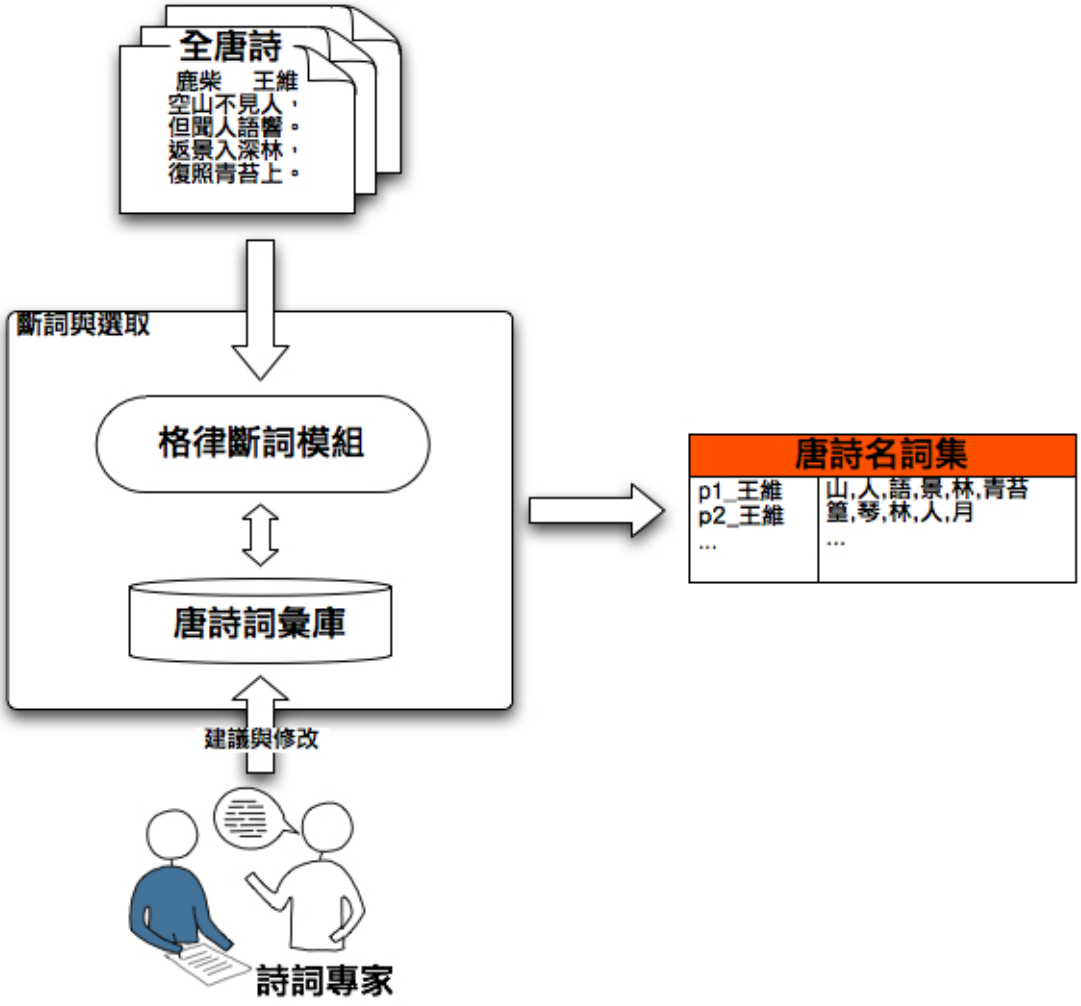

圖 2　名詞擷取程序圖

唐朝文學盛行詩，文人寫作不僅講究文字意涵的優美，也重視吟詠時語氣的變化。吟詠詩作時，偶數字數的詩句音調及語氣在句子末字能完整結束，但詩人常為了讓句子語氣的變化或詩中意念的延伸，而多出一字，自成一個音節，以五言詩為例，依字數分成 2-2-1 三個音節。文人有時為了強調語氣或使語調變化，會將音節點後移在句子的最後一字，而在第末三字的第一字產生一頓，此時五言詩可能變成 2-1-2 三個音節。詩人配合音節字數選用詞彙，組合成一首聲韻優美的詩作。

因此，利用唐詩字詞格律的特性，建立格律斷詞模組，並將斷出來的字詞查詢由詩詞專家所提供的唐詩詞彙庫，若有符合者，便是音韻和諧且有特定意涵的唐詩詞彙。

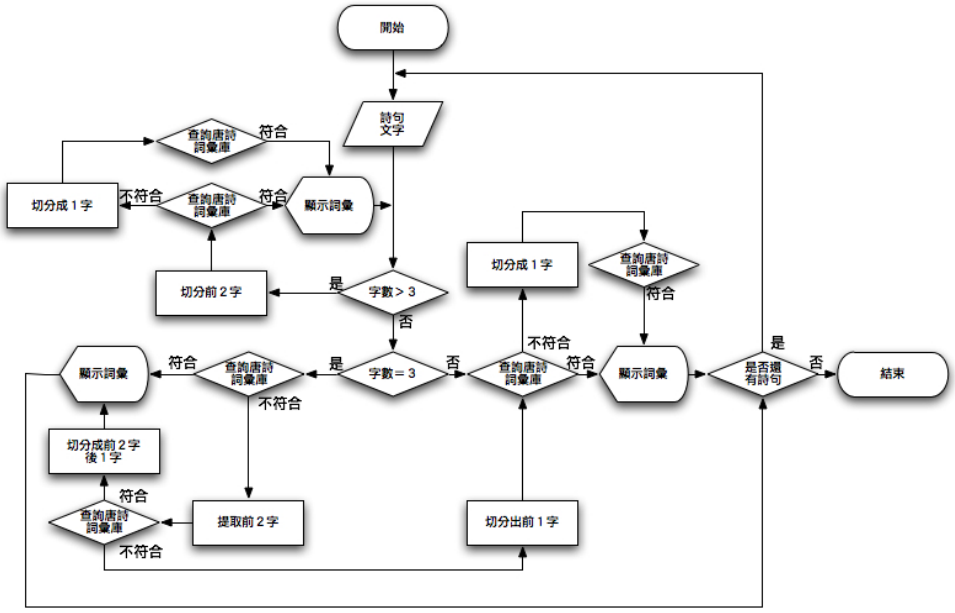

圖 3　名詞擷取流程圖

演算法 1　名詞擷取演算法

Algorithm: Noun retrieval algorithm. It is a process that retrieving nouns from a poem.
Symbol Definition:
Poem: one poem in **PoemSet**
Q*(*X*):** query noun ***X from Tang poem noun corpus and ***X*** ∈ Tang poem noun corpus
Input:
PoemSet: all poems created by Wang-Wei(王維)
Sentence: the union set of segmented vocabularies, e.g., $Sentence_i=\{S_1 \cup S_2 \ldots \cup S_n\}$
Output: *N*: noun set from *S*
Method:
for all $Poem_i$ in **PoemSet**
{ **for** each $Sentence_j$ of $Poem_i$
{ **for** each S_j of $Sentence_j$
Step1: if the length of $Sentence_j$ ＞3 then {
Step2: segment the first 2 Chinese chars from $Sentence_j$ into $\{S_i \cup S_j\}$
$Sentence_j = S_j$
Step3: if ($Q(S_i)$ = true) then add noun S_i to *N*
Step4: else { segment S_i into two single Chinese char S_1 and S_2
if ($Q(S_1)$ = true) then add noun S_1 to *N*
if ($Q(S_2)$ = true) then add noun S_2 to *N*
}
}
Step4: else if the length of $Sentence_j$ = 3 then {
Step5: if ($Q(Sentence_j)$ = true) then add noun $Sentence_j$ to *N*
Step6: else { segment the first 2 Chinese chars $Sentence_j$ into $\{S_i \cup S_j\}$
Step7: if ($Q(S_i)$ = true) then { add noun S_i to *N*
if ($Q(S_j)$ = true) then add noun S_j to *N*
$Sentence_j = S_j$
}
Step8: else { segment the first Chinese char $Sentence_j$ into $\{S_k \cup S_l\}$
Step9: if ($Q(S_k)$ = true) then add noun S_k to *N*
$Sentence_j = S_l$
}
}
}
Step10: else if the length of $Sentence_j$ = 2 then {
Step11: if ($Q(Sentence_j)$ = true) then add noun $Sentence_j$ to *N*
Step12: else { segment $Sentence_j$ into two single Chinese char S_1 and S_2
if ($Q(S_1)$ = true) then add noun S_1 to *N*
if ($Q(S_2)$ = true) then add noun S_2 to *N*
}
}
}
}

名詞擷取的流程如圖 3所示，並依演算法 1 進行名詞擷取。當詩句文字輸入後，先以兩字為一音節的規則切出前二字，並查詢唐詩詞彙庫中，若有符合的詞彙且是名詞，則予以標示；若無符合的詞彙則再切分成兩個單字詞來進行查詢及標示的工作。重複上述的步驟直到文字字數為末三字或末二字時，則進行下一個步驟。因詩句末三字可能有三字詞、前二後一、前一後二及都是單字詞的可能性，必須依序考慮，對應詞彙庫標示出可能的名詞。依詞長最大為優先，先考慮三字詞，再依詩句格律要求，依序考慮前二後一、前一後二及三個單字詞的可能性查詢詞彙庫，符合則切分詞彙並標示。

以王維鹿柴中的詩句「空山不見人」來說明名詞擷取的流程，如圖 4所示。先切分出前二字「空山」，查詢唐詩詞彙庫，沒有符合的名詞，而再切分成「空」、「山」二字，經查詢後，「山」為名詞。而句中的末三字「不見人」則必須較多的字詞組合的可能性。首先查詢彙中否有「不見人」，若有則標示為名詞。而詞彙庫中沒有「不見人」這個詞彙，則再切分成「不見」、「人」，將「不見」這詞查詢唐詩詞彙庫，發現沒有符合的名詞；再切分成「不」、「見人」，並查詢詞彙庫，仍是沒有符合的名詞，則再細分成「不」、「見」及「人」，經查詢後，只有「人」是名詞。所以，這句「空山不見人」這句中的名詞為「山」及「人」兩字。

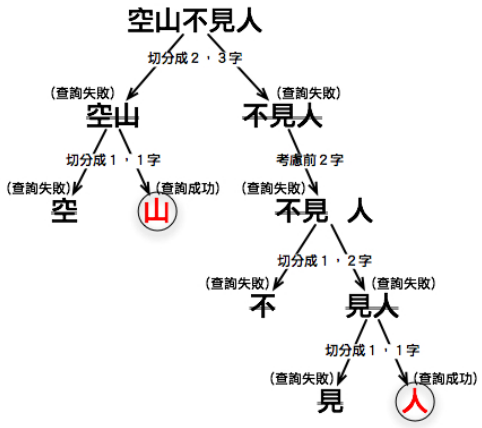

圖 4 「空山不見人」名詞擷取過程

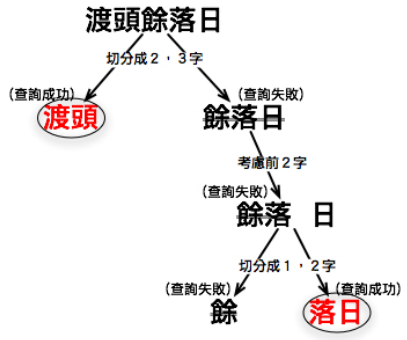

圖 5 「渡頭餘落日」名詞擷取過程

再以輞川閒居贈裴秀才迪詩句「渡頭餘落日」為例來說明，如圖 5。先切分出前二字「渡頭」，查詢詞彙庫後可成功標示為名詞。末三字「餘落日」查詢詞彙庫失敗後，而切分成「餘落」、「日」，查詢後因沒有「餘落」這詞彙，而改切分成「餘」、「落日」。「落日」經查詢成功後，可標示為名詞。所以，從「渡頭餘落日」中可以擷取出「渡頭」、「落日」等名詞詞彙。

在上述的例子中，可以發現，唐詩詞彙庫中所收錄的詞彙可能有單字詞、雙字詞及三字詞，三字詞在詞彙庫詞彙數所佔的比率為 3%，雙字詞則佔 56%，單字詞佔 41%。詞彙收錄至詞彙庫時，不只是收錄名詞詞彙，且收錄部份標示為非名詞特定詞彙，可以增快名詞擷取的效率，且也因此增強名詞擷取的準確率。例如圖 4中，將「不見」收錄至詞彙庫並標示為非名詞，因減少查詢的次數而增進名詞擷取效率。

3.2 名詞概念歸納

詩人對文字詞語有豐富的知識，擅長以不同詞彙描寫同一事物，例如：峰、嶺、嶽、巒與巔等字，都是山的同義詞或近義詞。致使語詞文字繁多，且電腦會因不同的文字表現，而視為不同的資料，造成資訊量過多，無法分析出文字使用的知識。

在本文中提出唐詩名詞概念階層，歸納語義概念相似的語詞為名詞類別。將語義繁雜的字詞轉換為概念精簡的名詞類別，且不會破壞詩作的意涵表現。利用

資料探勘的技術，分析詩作中名詞類別的使用，找出與詩人創作風格相關的知識。名詞概念階層建置是參考同義詞詞林分類的架構，共分為四層，如圖 6所示。

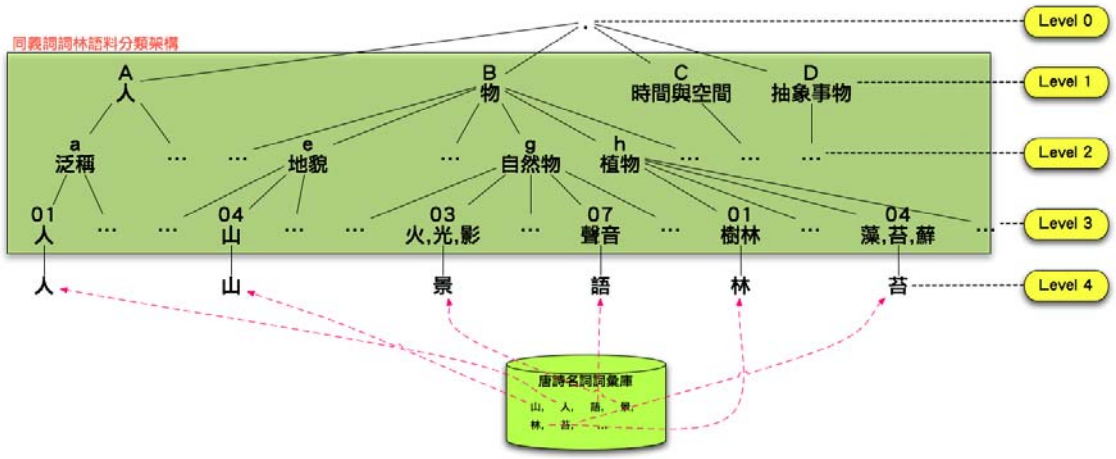

圖 6　唐詩名詞概念階層

其中將同義詞詞林中人（A）、物（B）、時間與空間（C）及抽象事物（D）等四類為名詞做為階層的第一層（Level 1，編號為A、B、C與D），此四大類底下的中類為第二層（Level 2，編號為a、b、c、d……），同義詞詞林的小類為唐詩名詞概念階層的第三層（Level 3，編號為 01、02、03、04……），且將唐詩名詞詞彙歸在第四層（Level 4）。因同義詞詞林中收錄部份古語詞，將唐詩名詞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，做為名詞在唐詩名詞概念階層中的位置編碼，名詞概念階層編碼如圖 7所示。例如，「人」在名詞概念階層的序號為 21，歸類在人（A）大類，泛稱（a）中類，第（01）小類之中。仍有部份的字詞有歧義的問題，再由詩詞專家建議及修改。

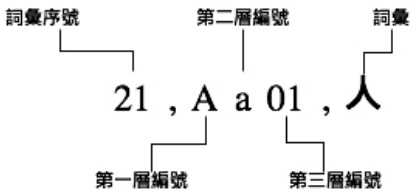

圖 7　名詞概念階層位置編碼

以王維詩作鹿柴、竹里館、送別、雜詩及相思等五首詩作為例來說明，並以全唐詩中王維詩作的次序來編號。將表 2中的名詞查詢唐詩名詞概念階層，將名詞在概念階層中的位置編碼，整理如表 3。

表 2　唐詩名詞集

詩名	編號	詩中所使用到的名詞
鹿柴	p311	山,人,人,語,景,林,青苔,上
竹里館	p323	篁,琴,林,人,明月
送別	p336	山,中,日暮,柴扉,春,草,明年,王孫
雜詩	p345	君,故鄉,故鄉,事,來日,綺窗,前,梅,花
相思	p349	紅豆,南國,秋,枝,君,物

表 3　唐詩名詞類別集

詩名	編號	詩中所使用到的名詞位置編碼
鹿柴	p311	Be04, Aa01, Aa01, Dk11, Bg03, Bh01, Bh04, Cb03
竹里館	p323	Bh01, Bp13, Bh01, Aa01, Bd02
送別	p336	Be04, Cb05, Ca29, Bn04, Ca19, Bh03, Ca18, Af06
雜詩	p345	Aa03, Cb15, Cb15, Da01, Ca12, Bn04, Ca11, Bh02, Bh02
相思	p349	Bh06, Cb02, Ca19, Dn08, Ah08, Ba01

若要以唐詩名詞概念階層第二層來探勘詩作的風格，則必須對詩作所使用的名詞作概念歸納（concept generalization)。將在概念階層第四層的名詞詞彙找出其所對應的第二層的位置編碼，例如：「人」是歸在第三層位置編碼「Aa01」之下的詞彙，其第二層的位置編碼為「Aa」。因此，原本是數千個不同的詞彙，經詞義概念歸納後，可以將眾多詞義不同的詞彙歸納成數十個詞義精簡的名詞類別，再藉由資料探勘的技術找出詩作中名詞類別使用的組合關聯規則，作為風格判斷的規則。例如，p311（鹿柴）使用到名詞類別Aa（人，泛稱）有 2 個，Be（地貌）有 1 個，Bg（自然物）有 1 個，Bh（植物）有 2 個，Cb（空間）有 1 個，Dk（文字）有 1 個，如表 4所示。在表中以這五首詩中所使用到的名詞概念階層第二層名詞類別作為資料欄位，如：Aa、Af、Ah等 15 個欄位，並計算詩作中各類名詞的使用個數，最後並加總計算出詩作名詞使用的總數量。

表 4　唐詩名詞類別表

PID	Aa	Af	Ah	Ba	Bd	Be	Bg	Bh	Bn	Bp	Ca	Cb	Da	Dk	Dn	總計
p311	2	0	0	0	0	1	1	2	0	0	0	1	0	1	0	8
p323	1	0	0	0	1	0	0	2	0	1	0	0	0	0	0	5
p336	0	1	0	0	0	1	0	1	1	0	3	1	0	0	0	8
p345	1	0	0	0	0	0	0	2	1	0	2	2	1	0	0	9
p349	0	0	1	1	0	0	0	1	0	0	1	1	0	0	1	6

3.3　名詞使用差異分群

文學評論者常依文人來區分詩風的派別，大都是以文人所有作品中較多且風格類似的詩作來論。而詩人也會嘗試創作不同風格的詩作，若將詩人所有的詩作來做分析，可能會因屬性資料過於分散，而造成分析不出詩作群中名詞共同使用組合的規則。本論文提出利用詩作名詞使用的差異性來分群，先找出詩人因名詞使用相近的風格詩作群，再進行風格規則探勘。

將名詞概念階層的位置編號作為名詞維度座標，則可以將詩作表示為名詞概念空間的點。例如，在名詞概念空間中有 p 個名詞維度$\{x_1, x_2, ..., x_p\}$，則第 i 首詩作可表示為$\{x_{i1}, x_{i2}, ..., x_{ip}\}$，第 j 首詩作可表示為$\{x_{j1}, x_{j2}, ..., x_{jp}\}$。其任兩首詩作的差異性（dissimilarity)可以式 3 來表示。

$$d(i,j) = \sqrt{\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + ... + \left|x_{ip} - x_{jp}\right|^2} \qquad (3)$$

較著名的分群演算法是 *k* 均值法(*k*-means)，可依資料屬性的差異性來分群。如以下演算法 2：

演算法 2　*k* 均值法

Algorithm: *k-means*. The k-means algorithm for partitioning based on the mean value of the objects in the cluster.

Input: The number of clusters k and a database containing n objects.

Output: A set of k clusters that minimizes the squared-error criterion.

Method:

(1) arbitrarily choose k objects as the initial cluster centers;

(2) repeat

(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

(4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;

(5) until no change;

在詩作中為了描寫景物或突顯所要表達的意境，詞義相似的名詞使用的個數較多，且不同的詩作，其名詞使用的個數並不相同。所以，在進行名詞使用差異性分群前，先將每首詩作所使用到的名詞類別個數轉換成名詞類別使用率。以 3.2 中所舉的例子來說明，將表 4依名詞類別使用率來轉換，其中p311（鹿柴）共用 8 個名詞，而Aa（人，泛稱）使用 2 個，其名詞類別使用率為 2/8（0.250），而Be（地貌）使用 1 個，名詞類別使用率為 1/8（0.125），最後詩作各類名詞的使用率總合必為 1。如此轉換後，仍可利用名詞類別的使用率來突顯詩作中所要強調的景色描寫或事物的表達。

表 5　唐詩名詞類別使用率表

PID	Aa	Af	Ah	Ba	Bd	Be	Bg	Bh	Bn	Bp	Ca	Cb	Da	Dk	Dn	總計
p311	0.250	0.000	0.000	0.000	0.000	0.125	0.125	0.250	0.000	0.000	0.000	0.125	0.000	0.125	0.000	1
p323	0.200	0.000	0.000	0.000	0.200	0.000	0.000	0.400	0.000	0.200	0.000	0.000	0.000	0.000	0.000	1
p336	0.000	0.125	0.000	0.000	0.000	0.125	0.000	0.125	0.125	0.000	0.375	0.125	0.000	0.000	0.000	1
p345	0.111	0.000	0.000	0.000	0.000	0.000	0.000	0.222	0.111	0.000	0.222	0.222	0.111	0.000	0.000	1
p349	0.000	0.000	0.167	0.167	0.000	0.000	0.000	0.167	0.000	0.000	0.167	0.167	0.000	0.000	0.167	1

在此，使用*k*均值法(*k*-means)計算詩作彼此間名詞類別使用率的差異性來分群。3.2 節例子中，將王維五首詩作分為兩群，表 6為是利用表 5來計算任兩首詩名詞類別使用率的差異性。

表 6　唐詩名詞使用差異性矩陣表

PID	p311	p323	p336	p345	p349
p311	—	—	—	—	—
p323	0.41	—	—	—	—
p336	0.53	0.63	—	—	—
p345	0.39	0.49	0.31	—	—
p349	0.48	0.56	0.42	0.36	—

在表 6中，縱軸與橫軸分別表示詩作，其所對應的交點是任意兩首詩作的名詞使用差異性。若將這五首詩作分為兩群，則設定*k*=2。一開始就選擇差異性最大p323 與p336 作為群中心，則從以表 6中的差異性，可以看出，p311 會歸在以p323 為群中心的第一群；而p345 及p349 則會歸在以 p336 為群中心的第二群。計算各群名詞類別使用率的平均值作為群中心，重新計算這五首詩作名詞類別的使用差異性並歸類至差異性最小的群中，與之前詩作分群配置不變而停止分群，分群結果如下表 7所示。其中p311、p336 與各群中心差異性最小，而選為群代表。

表 7　分群結果表

群編號	群代表	詩作
1	p311	p311, p323
2	p336	p336, p345, p349

3.4　名詞使用關聯規則探勘

利用所建置的名詞類別集，以關聯規則探勘（association rule mining)，找出作者在詩作語詞表現上的名詞類別使用組合。進行關聯規則探勘分析時，把每首詩所使用的名詞類別集視為一筆資料處理的內容，從同一位作者所創作的詩作集

中，找出大量且共同的名詞類別使用組合。

以 3.3 節表 7王維詩作分群結果的第二群來說明，包含p336、p245 及p349 等三首詩作，其中詩作中所使用到名詞概念階層第二層的名詞類別如下表 8所示。

表 8　唐詩名詞類別使用表

編號	詩中所使用到的名詞類別
p336	Be, Cb, Ca, Bn, Bh, Af
p345	Aa, Cb, Da, Ca, Bn, Bh
p349	Bh, Cb, Ca, Dn, Ah, Ba

在關聯規則探勘中，利用支持度(support)及可信度(confidence)作為規則選擇的依據。本論文中，支持度是欲分析的名詞類別組合在詩作群出現的機率，例如在表 8中，Cb（時間）在這三首詩中都有出現，其支持度為 3/3（1.00）；Bn（建築物）只有兩首詩使用到，其支持度為 2/3（0.67）；Ca（時間）、Cb（空間）在三首詩中都使用到，其支持度為 3/3（1.00）。可信度是欲分析的名詞類別組合出現在特定條件名詞類別組合詩作群中的機率，例如在表 8中，使用Ca（時間）、Cb（空間）的三首詩中全都使用Bh（植物），則可信度為 3/3（1.00）；使用Ca（時間）、Cb（空間）及Bh（植物）的三首詩中，只有兩首使用Bn（建築物），則可信度為 2/3（0.67）；從另一個角度來看，使用Cb（空間）、Bh（植物）及Bn（建築物）的兩首詩中，全都使用Ca（時間），則可信度為 2/2（1.00）。

表 9　唐詩風格規則分析表

編號	條件	結果	支持度	可信度
1	Bh(植物) Cb(空間)	Ca(時間)	1.00	1.00
2	Ca(時間) Cb(空間)	Bh(植物)	1.00	1.00
3	Bh(植物) Ca(時間)	Cb(空間)	1.00	1.00
4	Bn(建築物) Bh(植物) Cb(空間)	Ca(時間)	0.67	1.00
5	Bn(建築物) Bh(植物) Ca(時間)	Cb(空間)	0.67	1.00
6	Bn(建築物) Ca(時間) Cb(空間)	Bh(植物)	0.67	1.00
7	Bh(植物) Ca(時間) Cb(空間)	Bn(建築物)	0.67	0.67

以王維 p345《雜詩，三首之二》為例，說明詩作名詞類別使用的關聯規則。

《雜詩，王維》

君[Aa] 自 **故鄉**[Cb] 來，

應知 **故鄉**[Cb] **事**[Da] 。

來日[Ca] **綺窗**[Bn] **前**[Cb] ，

寒 **梅**[Bh] 著 **花**[Bh] 未。

從第 1 至 3 條規則的文字組合來看，來日、故鄉及梅花的組合，表現出思念故鄉的意象，是首思鄉抒懷詩。

表 9中第 1 至 3 條規則，時間（來日）、空間（故鄉、東）及植物（梅、花）等類名詞組合出現且支持度與可信度都是 1.00 可作為這群詩作風格名詞類別使用判別的規則。更進一步去看，第 4 至 7 條規則中，建築物（綺窗）與時間、空間及植物的名詞類別使用組合支持度是 0.67，表示建築物類名詞在詩作群中使用率較低，但從其可信度來判斷得知，使用建築物、植物及空間的文字組合後，詩人也會使用時間類的名詞的機率比使用植物、時間與空間後再使用建築物類名詞的機率來得高，因此，在這詩作群中，詩人創作風格中可能有「因使用建築物、植物及空間等類名詞，而又會使用時間類名詞」的詞彙使用關聯規則。在規則選擇過程中，觀察名詞類別組合支持度較高的規則，選擇可信度較高者作為名詞使用的風格判別規則。

四、 實驗過程與分析

在本論文中，選擇以全唐詩中王維的詩作來進行分析，共有 385 首。為了能分析詩作的名詞使用風格規則，建置唐詩名詞概念階層是最主要的工作，並利用所建置的名詞概念階層來擷取詩作名詞、名詞概念歸納，並進行名詞使用差異分群及名詞使用關聯規則探勘等詩風探勘階段，找出不同詩風詩作群中名詞使用的關聯規則。在以下的小節中，作詳細的說明。

4.1 唐詩名詞概念階層的建置

在前人研究中發現詩作使用的詞彙可作為風格探勘的特徵，若要對唐詩進行風格探勘，在資料處理與分析上，必須仰賴大量的名詞詞彙。擷取語句文字與已知的名詞詞彙比對後擷取出名詞，依名詞詞義來進行概念的歸納與分析，進而能分析詩作名詞使用的風格規則。所以，唐詩名詞概念階層建置工作是本論文中最為重要且需要詩詞專家的知識。在本文中提出建置唐詩名詞概念階層的方法，可分為以下三個步驟。

第一步驟：將詩詞專家提供唐詩中部份動物、植物等已標示詞義的名詞，依同義詞詞林的分類架構，建置成唐詩名詞概念階層的初期架構，如圖 8所示。雖然同義詞詞林收錄的詞彙是以現代漢語為主，其明確且完整的分類架構，是值得參考引用的。在建置唐詩名詞概念階層時，仍有些古字詞未收錄在同義詞詞林內，但因有詩詞專家的詞義標注，可輕易以人工判別的方式，新增至名詞概念階層中。例如：菊花、蘭、荷及梅等花卉名詞，可在同義詞詞林查到所對應的位置編碼Bh02（花），然而另有鬱金香、蘿等花卉類名詞未收錄在詞林內，但專家已明確地歸類為花卉，所以在建置唐詩名詞概念階層時，可以人工方式加入階層Bh02 的分類之下。

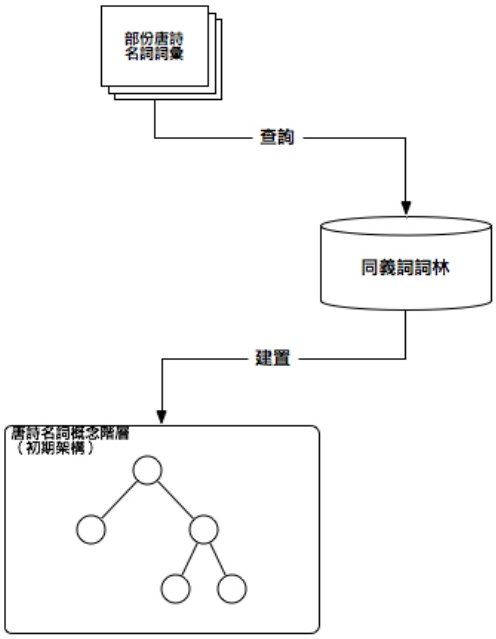

圖 8　唐詩名詞概念階層的建置（第一階段）

第二步驟：唐詩所用詞彙及語義與現代漢語有所差異，不能直接依查詢同義詞詞林來斷詞。所以，將詩句文字比對已建置的唐詩名詞概念階層中名詞詞彙，先行把已確認的唐詩名詞擷取出來，再把未確認的字詞比對同義詞詞林，可以提高斷詞的準確度。在此步驟中，將王維詩作經名詞擷取後產生的詞語，查詢第一階段建置的唐詩名詞概念階層，挑選出已知的名詞詞彙，若找不到對應的詞彙，則去查詢同義詞詞林資料庫，有符合詞義的名詞就新增至唐詩名詞概念階層中，如圖 9所示。

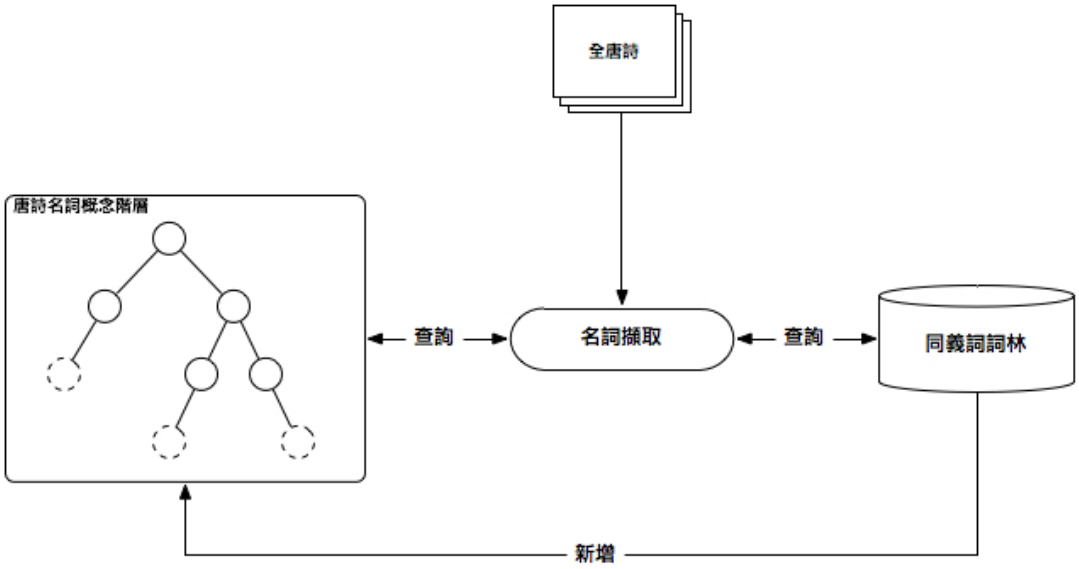

圖 9　唐詩名詞概念階層的建置（第二階段）

第三步驟：參考同義詞詞林現代漢語詞義所建置的名詞概念階層，可能仍會因歷史時間或地理空間造成字詞歧義的問題，必須由詩詞專家來驗證、增加或修改唐詩名詞概念階層，如圖 10所示。專家可依所要分析的作品來調整名詞概念階層。以“桃李”為例：在王維《酬郭給事》中所使用的“桃李”二字，若依名詞概念階層中的詞義為“門生後輩”，但在詩作中只是為“桃”、“李”的兩種果樹。所以，在此步驟中，仍需要詩詞專家判斷詩作部份名詞詞義的使用。

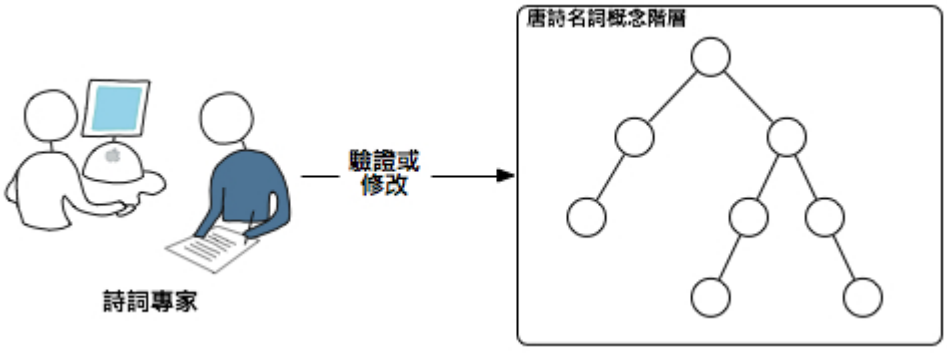

圖 10　唐詩名詞概念階層的建置（第三階段）

利用唐詩名詞概念階層內所收集的詞彙可以協助詩作名詞擷取的工作，且名詞概念階層內的詞彙量愈大時，分析詩作的名詞也就愈精確。將詩作中名詞擷取出來之後，對應名詞在概念階層中的位置編號，而整理成唐詩詩作名詞及名詞類別集。

因同義詞詞林內缺少專有名詞（人名、地名等）的歸類，致使仍有部份詞語尚未歸入唐詩名詞概念階層之中。

4.2　名詞擷取與名詞概念歸納

整合格律斷詞模組及唐詩名詞概念階層，建置成唐詩名詞擷取系統。利用唐詩名詞概念階層內所含的名詞作為在進行名詞擷取過程中參考的詞彙庫，並直接對應出概念階層位置編碼，如圖 11所示，將p13《早朝》輸入系統，系統輸出所擷選出來的名詞及唐詩概念階層位置編碼，如圖 12。

圖 11　名詞擷取系統輸入畫面－《早朝》

圖 12　名詞擷取系統輸出畫面－《早朝》

其中仍有部份名詞有歧義的問題，由人工判斷處理。例如圖 12之中“天”有

時間與空間兩個意義，由專家依此詩作內容判別，應是“空間”的“天”。

再以p61《藍田山石門精舍》為例說明，輸入畫面及輸出畫面如圖 13及圖 14。

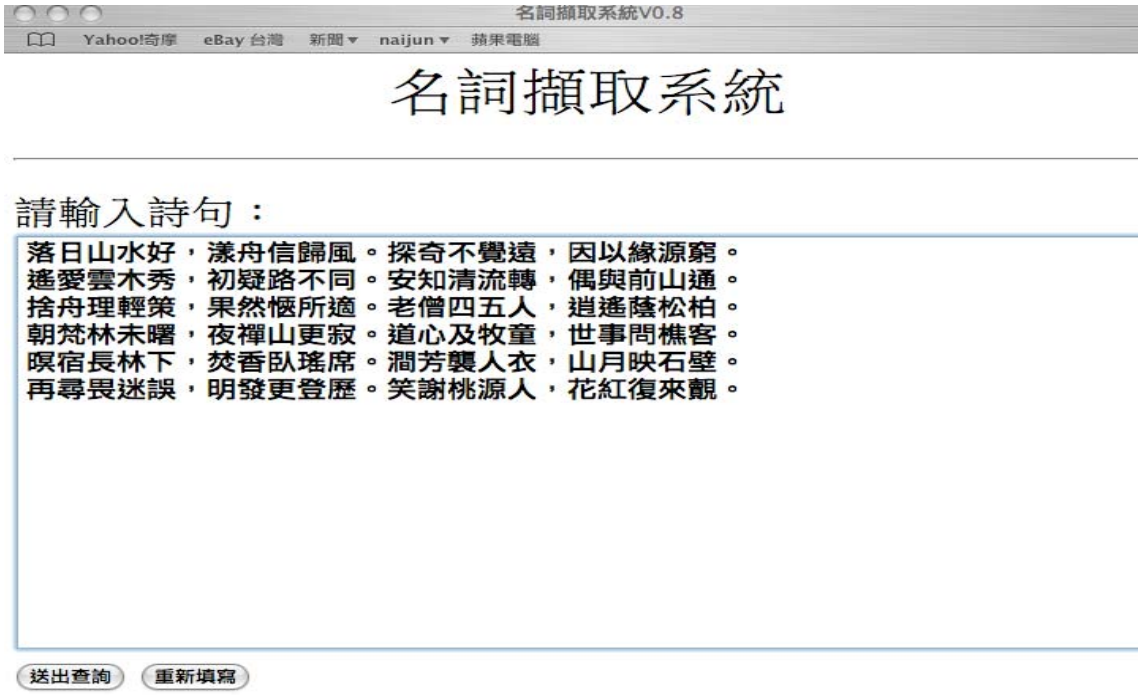

圖 13　詞擷取系統輸入畫面－《藍田山石門精舍》

圖 14　詞擷取系統輸出畫面－《藍田山石門精舍》

圖 14中“前”也有時間與空間不同的詞義，也需要專家判別。將系統所輸出的詩作名詞及名詞概念階層位置編碼整理成唐詩名詞集及唐詩名詞類別集，並進行名詞概念歸納成唐詩類別使用表，做為分群的分析資料。

4.3 唐詩名詞使用差異分析

在眾多文學評論專家的詩詞賞析中，王維雖以山水田園詩聞名，但全唐詩中所收錄王維的作品中，仍有邊塞詩、詠古詩、遊俠詩、贈別詩及應制詩等多樣的風格創作。根據詩詞專家依詩作名詞的使用情形，預先將王維詩作分為邊塞詩群、仕官詩群、思鄉懷人詩群、山水詩群、田園詩群及禪詩群等六群，因王維以山水詩而聞名，所以，預期實驗分析後，山水詩群應是分群後詩作最多的那一群。

表 10 各群名詞類別使用率摘要表

	第一群		第二群		第三群		第四群		第五群		第六群	
	名詞類別	使用率	名詞類別	使用率	名詞類別	使用率	名詞類別	使用率	名詞類別	使用率	名詞類別	使用率
人	Aa	0.03	Aa	0.03	Aa	0.07	Ad	0.09	Aa	0.02	Aa	0.09
	Af	0.03	Af	0.03	Af	0.03	Af	0.23	Af	0.02	Af	0.04
			Ah	0.02							Ae	0.03
											Ag	0.03
事	Di	0.06	Dn	0.06	Dk	0.05	Dn	0.05	Dn	0.04	Dn	0.04
					Dn	0.05	Di	0.05	Di	0.02	Di	0.03
時	Ca	0.05	Ca	0.07	Ca	0.07	Ca	0.05	Ca	0.06	Ca	0.19
地	Cb	0.14	Cb	0.12	Cb	0.09	Cb	0.05	Cb	0.17	Cb	0.11
物	Bo	0.10	Bn	0.10	Bh	0.10	Bh	0.09	Be	0.14	Bh	0.07
	Bh	0.08	Bm	0.08	Bn	0.06	Bi	0.09	Bh	0.10	Bn	0.07
	Be	0.07					Bn	0.09				

將 385 首王維詩作經名詞擷取，及名詞概念歸納至唐詩名詞概念階層第二層

後所產生的唐詩名詞類別矩陣，利用分群方法是 k 均值法針對名詞類別使用的差異性來分群，並設定 $k = 6$。

一般而言，人在分析事件時會針對人、事、時、地、物等事件屬性來看，所以摘錄出各群名詞類別使用率較高的，並依人（A）、事（D）、時（Ca）、地（Cb）、物（B）等類來分析與討論，如上

表 10。其中，各群在時間、空間類名詞使用都有一定的比率，且各群都有些

	第一群		第二群		第三群		第四群		第五群		第六群	
	名詞類別	使用率	名詞類別	使用率	名詞類別	使用率	名詞類別	使用率	名詞類別	使用率	名詞類別	使用率
人	Aa	0.03	Aa	0.03	Aa	0.07	Ad	0.09	Aa	0.02	Aa	0.09
	Af	0.03	Af	0.03	Af	0.03	Af	0.23	Af	0.02	Af	0.04
			Ah	0.02							Ae	0.03
											Ag	0.03
事	Di	0.06	Dn	0.06	Dk	0.05	Dn	0.05	Dn	0.04	Dn	0.04
					Dn	0.05	Di	0.05	Di	0.02	Di	0.03
時	Ca	0.05	Ca	0.07	Ca	0.07	Ca	0.05	Ca	0.06	Ca	0.19
地	Cb	0.14	Cb	0.12	Cb	0.09	Cb	0.05	Cb	0.17	Cb	0.11
物	Bo	0.10	Bn	0.10	Bh	0.10	Bh	0.09	Be	0.14	Bh	0.07
	Bh	0.08	Bm	0.08	Bn	0.06	Bi	0.09	Bh	0.10	Bn	0.07
	Be	0.07					Bn	0.09				

不同，例如：第一群機具（Bo）較為特殊，第二群則是材料（Bm），第三群與其它各群相較後多了文教（Dk），第四群是身份（Ad）與籍屬（Af），第五群則在地貌（Be）使用比率較高，第六群在時間（Ca）使用比率較高。

在風格分群後，各群名詞類別使用率較高的名詞詞彙如以下各群的分析，並推斷詩群的風格，在下一節中，再以關聯規則進一步探勘。

第一群：以p10 為群代表，共有 42 首詩。詩中名詞類別使用率較高的有空間（Cb）、機具（Bo）、植物（Bh）、地貌（Be）、社會及政法（Di）、時間（Ca）、人

泛稱（Aa）、身份（Af）等類，所使用的文字如表 11。在此群詩作中多使用弓、劍、城、關、國及戰等名詞，推論此群可能為邊塞詩群。

表 11　第一群詩作名詞使用分類表

類別編碼	目錄名稱	使用的名詞
Aa	人，泛稱	人…
Af	身份	官、使、君、王、侯…
Be	地貌	山、川、平原、水…
Bh	植物	花、草、樹、林…
Bo	機具	車、舟、鞍、帆、弓、劍…
Ca	時間	春、秋、暮…
Cb	空間	城、關…
Di	社會、政法	國、虜、戰…

第二群：以p13 為群代表，共有 51 首詩。詩中名詞類別使用率較高的有空間（Cb）、建築物（Bn）、材料（Bm）、時間（Ca）、數量與單位（Dn）、人泛稱（Aa）、身份（Af）、親人與眷屬（Ah）等類，所使用的文字如表 12。在這群中常用君、王、官、金、銅、玉、建章、宮、門等名詞，可能是仕官詩群。

表 12　第二群詩作名詞使用分類表

類別編碼	目錄名稱	使用的名詞
Aa	人，泛稱	人…
Af	身份	君、王、官、臣、侯…
Ah	親人與眷屬	親、夫、夫人…
Bm	材料	金、銅、玉、石…
Bn	建築物	建章、宮、門、窗、軒、園、陌…
Ca	時間	春、秋、朝、晨、曙…
Cb	空間	家、城、路…

第三群：以p1 為群代表，共有 92 首詩。詩中名詞類別使用率較高的有植物（Bh）、空間（Cb）、時間（Ca）、人的泛稱（Aa）、建築物（Bn）、文教（Dk）、

數量與單位（Dn）、身份（Af）等類，所使用的文字如表 13。從詩作群中文字使用草、花、林、門、田、言、歌、書等，應是描寫鄉野生活的田園詩群。

表 13　第三群詩作名詞使用分類表

類別編碼	目錄名稱	使用的名詞
Aa	人，泛稱	人、君、我…
Af	身份	王、官…
Bh	植物	草、花、林…
Bn	建築物	門、田、舍…
Ca	時間	春、年、時…
Cb	空間	城、天、空、家…
Dk	文教	言、歌、詩、書、語…

第四群：以 p364《少年行，四首之四》為群代表和 p367《送王尊師歸蜀中拜掃》，共有 2 首詩。因詩中文字使用與其它各群差異太大，且數量太少，可視為特例，而不討論。

第五群：以p2 為群代表，共有 147 首詩。詩中名詞類別使用率較高的有空間（Cb）、地貌（Be）、植物（Bh）、時間（Ca）、數量與單位（Dn）、人的泛稱（Aa）、身份（Af）、社會與政法（Di）等類，所使用的文字如表 14。在詩作中多出現山、水、樹、松、天、空等名詞，且詩作數量是最多的，應該是山水詩群。

表 14　第五群詩作名詞使用分類表

類別編碼	目錄名稱	使用的名詞
Aa	人，泛稱	人、君、誰…
Af	身份	官、侯、吏、臣…
Be	地貌	山、水、江、川、峰、地、原、浦…
Bh	植物	樹、松、竹、花、草、木…
Ca	時間	春、秋、朝、晨、夕…
Cb	空間	天、空、城、家、路、關…
Di	社會、政法	道、州…

第六群：以p8 為群代表，共有 51 首詩。詩中名詞類別使用率較高的有空間（Cb）、地貌（Be）、植物（Bh）、時間（Ca）、數量與單位（Dn）、人的泛稱（Aa）、身份（Af）、社會與政法（Di）等類，所使用的文字如表 15。這群詩作常君、人、山、水、花、草及家等名詞，多是思念家鄉或送人遠行的思鄉懷人詩群。

表 15　第六群詩作名詞使用分類表

類別編碼	目錄名稱	使用的名詞
Aa	人，泛稱	君、人、吾…
Af	身份	君、王…
Be	地貌	山、水、江、川、峰、地、原、浦…
Bh	植物	花、草、林…
Bn	建築物	門、扉…
Ca	時間	時、春、秋、朝、暮…
Cb	空間	城、路、家…
Di	社會、政法	道、洛陽…

表 16　第五群群中心與其它群群中心差異性比較

	第一群	第二群	第三群	第四群	第六群
差異性	0.140	0.144	0.162	0.308	0.196

其中以第五群的詩作數量最多，若以文學評論者認為王維詩作多為山水詩，則第五群詩作群可能為山水詩群。表 16為以第五群群中心與其它各群群中心的差異性比較，從群中心差異性發現，各群明顯地具有代表性。在下一節中，會以關聯規則探勘找出各詩群的名詞使用組合規則。

4.4　唐詩名詞使用關聯規則探勘

唐詩詩作經名詞使用差異性分群後，可以找出名詞使用相近的詩作群，並進一步利用關聯規則探勘，找出詩作群中名詞使用的組合規則，作為詩作風格的規則表現。利用 Apriori 方法來析詩作名詞使用的關聯規則，其中預設支持度與可信

度的門檻分別為 0.5 與 0.8，結果如下。

第一群：名詞類別使用關聯規則分析如表 17，以空間與機具的組合來看，多為關、城等與車、弓、劍等等名詞組合，輔以國、戰等社會名詞及山、川等地貌類名詞組合，多在描寫邊塞戰事準備或風光，應為邊塞詩群。

表 17　詩作群名詞類別使用關聯規則分析（第一群）

編號	條件	結果	支持度	可信度
1	Cb(空間)	Bo(機具)	0.86	1.00
2	Bi(動物)	Bo(機具)	0.62	1.00
3	Be(地貌)	Bo(機具)	0.62	1.00
4	Di(社會、政法)	Bo(機具)	0.60	1.00
5	Bh(植物)	Bo(機具)	0.60	1.00
6	Ca(時間)	Bo(機具)	0.57	1.00
7	Bi(動物) Cb(空間)	Bo(機具)	0.55	1.00
8	Bn(建築物)	Bo(機具)	0.55	1.00
9	Cb(空間) Di(社會、政法)	Bo(機具)	0.52	1.00
10	Be(地貌) Cb(空間)	Bo(機具)	0.52	1.00
11	Ca(時間) Cb(空間)	Bo(機具)	0.50	1.00
12	Bn(建築物) Cb(空間)	Bo(機具)	0.50	1.00
13	Bh(植物) Cb(空間)	Bo(機具)	0.50	1.00

從這群詩作選出《隴西行》，說明邊塞詩群名詞使用關聯規則，其中詩作名詞使用符合上表第 1 至 12 條等規則。

《隴西行》

十[Dn] **里**[Dn] **一**[Dn] 走 **馬**[Bi] ， **五**[Dn] **里**[Dn] **一**[Dn] 揚 **鞭**[Bo] 。

都護[Ae] **軍書**[Dk] 至， **匈奴**[Di] 圍 **酒泉**[Cb] 。

關山[Be] 正飛 **雪**[Bf] ， **烽戍**[Bg] 斷無 **煙**[Bf] 。

以第 1 及 3 條規則來看，**酒泉**表示**空間**中的關口險要之處，**鞭**是驅策馬的**機**

具，**馬**是供人交通及運輸的**動物**，說明情境是邊塞地區軍情告急。以第 9 條規則說明，**酒泉**是**空間**中重要的關口，**匈奴**是邊疆民族，歸類在社會類的用詞，**鞭**是驅策馬的**機具**，可看出外族來犯的情境。**酒泉**是**空間**中重要的關口，位於**地貌**險要之處，緊臨**關山**，以第 10 條規則可看出邊疆關口之險要。

第二群：在表 18中以空間與材料的名詞類別組合來看，將家、城、路等與金、玉等名詞組合，輔以建章、宮、門、窗等建築物類名詞，多為仕官時的寫景詩及應制詩，在本文中歸為仕官詩群。

表 18　詩作群名詞類別使用關聯規則分析（第二群）

編號	條件	結果	支持度	可信度
1	Cb(空間)	Bm(材料)	0.82	1.00
2	Ca(時間)	Bm(材料)	0.69	1.00
3	Bh(植物)	Bm(材料)	0.67	1.00
4	Bn(建築物)	Bm(材料)	0.65	1.00
5	Ca(時間) Cb(空間)	Bm(材料)	0.57	1.00
6	Bn(建築物) Cb(空間)	Bm(材料)	0.57	1.00
7	Bh(植物) Cb(空間)	Bm(材料)	0.55	1.00
8	Dn(數量、單位)	Bm(材料)	0.55	1.00
9	Cb(空間) Dn(數量、單位)	Bm(材料)	0.51	1.00
10	Bn(建築物) Ca(時間)	Bm(材料)	0.51	1.00
11	Bp(用品)	Bm(材料)	0.51	1.00

這仕官詩群中的代表是《早朝》，符合上表第 1 至 7 及 10、11 條等名詞類別使用規則。

《早朝》

皎潔 **明星**^{Bd} 高，蒼茫遠 **天**^{Cb} **曙**^{Ca} 。

槐^{Bh} **霧**^{Bf}暗不開， **城**^{Cb} **鴉**^{Bi} 鳴稍去。

始聞 **高閣**^{Bn} **聲**^{Bg} ，莫辨更 **衣**^{Bq} **處**^{Cb} 。

銀[Bm] **燭**[Bp] 已成行， **金門**[Dm] 儼騶馭。

其中使用到**天**空及**城**市等的**空間**類名詞，**時間**是**曙**光乍現的清晨，**高閣**是官臣所居住的**建築物**，**材料**用**銀**製成的**燭**台，是室內常見的**用品**，以第 5、6、10 及 11 等條名詞使用組合規則來說明，可想見仕官用品的奢華，從閣樓高處望去，見到早晨的曙光，描寫官臣早起為公務繁忙的景象。

第三群：在表 19中，以人、空間及植物的名詞組合來看，其中文字多為人、我等與天、空、城、家等名詞組合，輔以林、花、草等植物類名詞，多為身處園林的田園詩群。

表 19　詩作群名詞類別使用關聯規則分析（第三群）

編號	條件	結果	支持度	可信度
1	Aa(人泛稱) Cb(空間)	Bh(植物)	0.50	0.88
2	Bn(建築物)	Cb(空間)	0.53	0.86
3	Aa(人泛稱)	Bh(植物)	0.58	0.84
4	Bh(植物)	Cb(空間)	0.63	0.83
5	Dk(文教)	Cb(空間)	0.51	0.82
6	Ca(時間)	Cb(空間)	0.58	0.82

《酬諸公見過》

嗟 **予**[Aa] 未喪，哀此 **孤生**[Ag] 。屏居 **藍田**[Bn] ，薄 **地**[Be] 躬耕。

歲宴[Ca] 輸 **稅**[Dj] ，以奉粢盛。 **晨**[Ca] 往 **東**[Cb] **皋**[Be] ， **草**[Bh] **露**[Bf] 未晞。

暮[Ca] 看 **煙火**[Bg] ，負擔來歸。 **我**[Aa] 聞有 **客**[Aj] ， **足**[Bk] 掃 **荊**[Bh] **扉**[Bn] 。

簞 **食**[Br] **伊**[Aa] 何，副 **瓜**[Bh] 抓 **棗**[Bh] 。仰廁群 **賢**[Ak] ，皤然 **一**[Dn] **老**[Ab] 。

愧無 **莞**[Bh] 簟，班 **荊**[Bh] 席稿。汎汎登 **陂**[Be] ，折 **彼**[Aa] **荷花**[Bh] 。

靜觀素 **鮪**[Bi] ，俯映白 **沙**[Bm] 。 **山**[Be] **鳥**[Bi] 群飛， **日**[Bd] 隱輕 **霞**[Bf] 。

登 **車**[Bo] 上 **馬**[Bi] ，倏忽 **雲**[Bf] 散。 **雀**[Bi] 噪 **荒村**[Cb] ， **雞**[Bi] 鳴空 **館**[Dm] 。

還復幽獨，重欷累歎。

《酬諸公見過》符合上表中的規則，在詩作中使用到**予**、**我**、**伊**等對**人的泛稱**，**荒村**是**空間**類中鄉村的用詞，及**草**、**荊**、**瓜**、**棗**、**莞**、**荷花**等包含鄉野花草

及瓜果的**植物**類名詞，以第 1 條規則來說明，身處於鄉村的田野生活。輔以**藍田**及**扉**等**建築物**類中田園與屋舍的名詞，**時間**是泛著輕**霞**的日**暮**，以第 2 條規則可以想像，朝出暮回的農耕生計，遠望山水的田園生活。

第五群：如表 20中建築物與空間的名詞組合類別來看，文字多為門、舍等與天、空、城、家等名詞組合，輔以山、水、江、川等地貌類名詞及樹、松、竹、花等植物類名詞，多在描寫居處山林與逸遊於山水花草間的山水詩群。

表 20　詩作群名詞類別使用關聯規則分析（第五群）

編號	條件	結果	支持度	可信度
1	Bn(建築物)	Cb(空間)	0.62	0.99
2	Be(地貌) Bn(建築物)	Cb(空間)	0.54	0.99
3	Bh(植物) Bn(建築物)	Cb(空間)	0.50	0.99
4	Be(地貌) Bf(氣象)	Cb(空間)	0.59	0.95
5	Ca(時間)	Be(地貌)	0.61	0.94
6	Be(地貌) Bh(植物)	Cb(空間)	0.59	0.94
7	Ca(時間) Cb(空間)	Be(地貌)	0.56	0.93
8	Bf(氣象)	Cb(空間)	0.63	0.93
9	Bh(植物)	Cb(空間)	0.96	0.93
10	Be(地貌)	Cb(空間)	0.80	0.91

《歸輞川作》

谷^{Be} 口 **疏鐘**^{Bp} 動， **漁**^{Ae} **樵**^{Ae} 稍欲稀。

悠然遠 **山**^{Be} **暮**^{Ca} ，獨向 **白雲**^{Bf} 歸。

菱^{Bh} 蔓弱難定，楊 **花**^{Bh} 輕易飛。

東^{Cb} **皋**^{Be} **春**^{Ca} **草**^{Bh} **色**^{Bg} ，惆悵掩 **柴扉**^{Bn}。

《歸輞川作》是王維山水詩的佳作，名詞使用符合上表全部的規則，詩中使用**柴扉**的**建築物**類名詞，**春**、**暮**等**時間**類，**東**為**空間**類，**山**、**谷**、**皋**等**地貌**，**白雲**為**氣象**類，**菱**、**花**、**草**等**植物**類名詞，以第 2、3、4、6 及 7 條等規則說明文字

意象，詩人站在房門前遠眺一片春光山色，氣象萬千，夕陽美景無限好，引起心中些許愁。

第六群：以表 21中空間與時間的名詞類別組合規則來看，文字上以時、春、秋、朝、暮等與城、路、家等名詞組合，輔以君、人、王等人物類名詞，送人遠行、向山林寄愁懷的思鄉懷人詩群。

表 21　詩作群名詞類別使用關聯規則分析（第六群）

編號	條件	結果	支持度	可信度
1	Cb(空間)	Ca(時間)	0.75	0.97
2	Aa(人泛稱)	Ca(時間)	0.71	0.97
3	Bh(植物)	Ca(時間)	0.63	0.97
4	Bn(建築物)	Ca(時間)	0.59	0.97
5	Aa(人泛稱) Cb(空間)	Ca(時間)	0.55	0.97
6	Bn(建築物) Cb(空間)	Ca(時間)	0.51	0.96
7	Bh(植物) Cb(空間)	Ca(時間)	0.51	0.96
8	Be(地貌)	Ca(時間)	0.51	0.96

《送綦毋潛落第還鄉》

聖代[Ca]無 **隱者**[Af]， **英靈**[Ak] 盡來歸。

遂令 **東山**[Be] **客**[Aj] ，不得顧采 **薇**[Bh] 。

既至 **君門**[Dm] 遠，孰云 **吾**[Aa] **道**[Di] 非。

江淮[Cb] 度 **寒食**[Ca]， **京洛**[Di] 縫 **春**[Ca] **衣**[Bq] 。

置 **酒**[Br]臨長 **道**[Di] ，同 **心**[Df] 與 **我**[Aa] 違。

行當浮 **桂**[Bh] 櫂，未幾拂 **荊**[Bh] **扉**[Bn] 。

遠 **樹**[Bh] 帶 **行客**[Ag] ，**孤村**[Cb]當 **落暉**[Bg] 。

吾[Aa] **謀**[Df] 適不用，勿謂 **知音**[Aj] 稀。

在《送綦毋潛落第還鄉》詩中道盡詩人仕途失意，送人遠行，欲隱退鄉野的愁懷。詩中文字使用**聖代**、**春**、**寒食**等**時間**類，**江淮**、**孤村**等**空間**類，**吾**、**我**等

人的泛稱，**薇**、**荊**、**樹**等**植物**類，**扉**為**建築物**類，**東山**為**地貌**類等名詞，以第 5、6 及 7 條規則說明，政治開明，佳節將至，將更遠行，孤身在鄉野的落漠情懷。

實驗結果與預期有所差異，在詩作群的文字分析後，並無禪詩群，從整個詩作文字使用表現來解釋，王維詩作時間與空間類名詞使用甚多，在描寫時空景像變化時，已融入自身情感，且文字多以山水花草及氣象類名詞，與其它風格詩群相同，而無法區分出來。各風格群詩作文字使用及數量都不同，應調整可信度及支持度的下限，並觀察找出較具代表性的名詞類別使用關聯規則。

在實驗過程中，利用將名詞依詞義歸納至名詞概念階層第二層，並進行風格探勘，從結果發現，部份名詞類別在不同風格群中都有使用，但文字表現的意象卻是不同，例如：時間、空間及建築物等。在進行風格規則探勘時，針對部份名詞概念階層第二層名詞類別，以其之下的第三層名詞類別來進行多層次探勘(multilevel mining)，更能找出風格與特定名詞使用關聯性高的規則。

4.5 系統準確性評估

從鮑康健的〈歷代山水詩名篇賞析〉[18]中所整理的山水派唐詩共 24 首，經過名詞擷取及名詞概念歸納等階段，選擇唐詩名詞概念階層第二層名詞類別作為資料維度，轉換成名詞類別使用率表。計算每一個名詞類別的平均值，作為這 24 首山水詩的群中心，並與王維詩作各群的群中心比較其差異性，結果如下表 22。

表 22　山水詩作群中心與王維詩作群群中心差異性比較表

群集	第一群	第二群	第三群	第四群	第五群	第六群
差異性	0.152	0.175	0.202	0.345	0.076	0.225

從表中可以發現，王維詩作第五群與山水詩作群最為相近。對此 24 首山水詩進行名詞使用關聯規則探勘，名詞使用關聯規則如表 23所示。

表 23　山水詩作群名詞類別使用關聯規則分析

編號	條件	結果	支持度	可信度
1	Cb(空間)	Be(地貌)	0.96	1.00
2	Bf(氣象)	Be(地貌)	0.79	1.00
3	Bf(氣象) Cb(空間)	Be(地貌)	0.75	1.00
4	Ca(時間) Cb(空間)	Be(地貌)	0.63	1.00
5	Cb(空間) Dn(數量、單位)	Be(地貌)	0.58	1.00
6	Bh(植物) Cb(空間)	Be(地貌)	0.54	1.00
7	Bi(動物) Cb(空間)	Be(地貌)	0.50	1.00
8	Bf(氣象) Bh(植物) Cb(空間)	Be(地貌)	0.50	1.00
9	Be(地貌) Bh(植物) Cb(空間)	Bf(氣象)	0.50	0.92

對照表 20比較實驗中王維詩作第五群（山水詩群）後，可以發現，地貌、時間、空間、氣象及植物等名詞類別組合都有出現且可信度也非常地接近，可以驗證王維詩作第五群是山水詩群。然而王維山水詩群與文學評論所歸類的山水詩不同之處在於王維多用空間及建築物類名詞，突顯隱居在山林之中，才能將所見的山水景象刻畫得如此鮮明與生動。

五、 結論與展望

唐朝詩作文字精簡，詞義關聯性高，且內容多述事及寫景，景色事物多用名詞，所以應用關聯規則探勘詩作名詞使用的組合規則，作為風格判別的依據。本文中整合詩詞專家的知識與同義詞詞林名詞分類原則與架構，建構唐詩名詞概念階層，可歸類唐詩名詞詞義相同或相似為一類，並將詞義繁瑣的名詞概念歸納成詞義精簡的名詞類別，供唐詩詩風探勘使用。然而，詩人創作風格多樣，且同一風格詩作使用詞義相近的詞彙。所以，將文人詩作依名詞類別使用差異性分群，再針對單一風格詩群進行關聯規則探勘。

唐詩詩風探勘流程共有四個階段：

1、 名詞擷取：依詩句文字音韻格律的要求，及詩詞專家知識建置唐詩格律斷詩模組，結合唐詩名詞概念階層，將詩句中名詞擷選出來，整理成唐詩名詞集及唐詩名詞類別集。

2、 名詞概念歸納：將唐詩名詞類別集中的名詞位置編碼，依唐詩名詞概念階層概念歸納，並整理成唐詩名詞類別使用表。

3、 名詞使用差異性分群：將唐詩名詞類別使用表依詩作名詞使用數目轉換成唐詩名詞類別使用率表，運用分群的技術，參考詩詞專家建議，依詩作名詞使用率分為 6 群。

4、 名詞使用關聯規則探勘：利用關聯規則探勘詩人風格詩作群集中使用名詞類別的組合規則，設定可信度(confidence)及支持度(support)分析詩人創作時名詞使用的風格規則。

王維詩作經過唐詩詩風探勘後，可分為邊塞詩群、仕官詩群、田園詩群、山水詩群及思鄉懷人詩群，並得到各風格群詩作名詞使用的關聯規則。邊塞詩群中多動物、社會、政法、地貌等類與空間及機具類名詞組合規則；仕官詩群中多時間、建築物、植物、數量、單位及空間與材料等類名詞組合規則；田園詩群多用建築物、植物、文教、時間與空間等名詞組合規則；山水詩群則多地貌、建築物、

植物、氣象、時間及空間等類名詞組合規則；思鄉懷人詩群使用人、建築物、植物及時間與空間等類名詞組合規則。應用各風格群詩作名詞組合規則，分析詩作的意象可與詩詞專家的評論相互佐證。

詩詞專家評論的唐朝山水詩作群，經名詞擷取、名詞概念歸納後，取各類名詞使用率的平均值作為群中心，與實驗中各風格群群中心比較差異性，作準確性評估，與實驗山水詩群非常接近且與其它各群有著較大的差異性。進一步探勘名詞使用的關聯規則，與實驗分析中大部份的詞類組合相同。

在本文中結合專家對詩的知識建置名詞擷取及名詞概念階層，可以快速從詩作中擷選出名詞並對應至名詞概念階層，讓系統能判斷所使用名詞的詞義。根據概念階層的架構，歸納複雜的名詞詞義成精簡的名詞類別，並運用分群與關聯規則探勘技術找出特定風格詩作的名詞使用關聯規則，供詩詞專家分析詩作用詞及風格分類。

未來繼續擴增唐詩名詞概念階層的詞彙量，依概念階層的架構，針對唐朝文人詩作多層次探勘(multilevel mining)分析。依專家分析結果標示部份風格詩作，作為風格分群的起始中心，可提高詩作詩風分群的效率。並將詩風名詞使用關聯規則精煉，作為判別詩風的規則，運用此技術可將詩作自動分類或判斷詩風，協助專家作更入的研究與探討。

參考文獻

[1] 士會，詩詞挈領，萬里機構‧萬里書店，民國 90 年 4 月。

[2] 吳丈蜀，讀詩常識，萬卷樓圖書股份有限公司，民國 79 年 3 月。

[3] 蘅塘退士選輯，王進祥集解，唐詩三百首集解，頂淵文化事業有限公司發行，民國 88 年 8 月。

[4] 陶文鵬，明月松間照詩佛：王維作品賞析，德威國際文化有限公司，民國 92 年 12 月。

[5] 羅鳳珠，李元萍，曹偉政，文學上網—詩詞創作輔助工具的新嚐試【倚聲填詞】格律自動檢測索引教學系統，一九九七年 TANET97 國際學術研討會，國立臺灣大學，1997 年 10 月 21-24 日。

[6] 蕭斯聰，蘇俊銘，曾憲雄，羅鳳珠，蔡文能，近體詩格律與專家系統知識庫之研究，第十一屆國際電腦輔助教學研討會 ICCAI2003 暨中華民國電腦輔助教學研討會，國立台灣師範大學，民國 92 年 4 月 24-26 日。

[7] 楊哲青，蘇俊銘，曾憲雄，羅鳳珠，詩作風格知識庫之研究—以蘇軾近體詩為例，第一屆文學與資訊科技國際會議，清華大學、元智大學合辦，民國 92 年 12 月 9-11 日。

[8] 楊哲青，詩作風格知識庫之研究－以蘇軾近體詩為例，國立交通大學，碩士論文，民國 93 年。

[9] 王迺仁，曾憲雄，楊哲青，蘇俊銘，絕句詩格律診斷專家系統，第十屆人工智慧與應用研討會，國立高雄大學，2005 年 12 月 2-3 日。

[10] 俞士汶，胡俊峰，唐宋詩之詞匯自動分析及應用, LANGUAGE AND LINGUISTICS 4.3:631-647, 2003

[11] Niles, I., Pease, A., Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, Proceedings of the IEEE International Conference on Information and Knowledge Engineering. (IKE 2003), Las Vegas,

Nevada, June 23-26, pp.412-416, 2003.

[12] 梅家駒等編著，同義詞詞林，臺灣東華書局股份有限公司，民國 86 年 3 月。

[13] 陳書磊，詩句中的語意結構之擷取和檢索，國立清華大學，碩士論文，民國 93 年。

[14] Jiawie Han, Micheline Kamber, Data Mining: Concepts and techniques, Morgan Kaufmann, New York, 2000.

[15] Ian H. Witten, Eibe Frank, "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[16] Yong Yi, Zhong-Shi He, Liang-Yan Li, Tian Yu, Elaine Yi, Advanced studies on traditional Chinese poetry style identification, Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005

[17] 王迺仁，曾憲雄，楊哲青，蘇俊銘，羅鳳珠，詩風規則之研究-以唐朝近體詩為例，第二屆文學與信息科技國際研討會，北京大學，2005 年 12 月 8-10 日。

[18] 鮑康健，歷代山水詩名篇賞析，華成圖書出版股份有限公司，民國 92 年 11 月。