

歷史警學案例之義類關鍵詞標註方法

古永昌

中央警察大學教務處

unicon@mail.cpu.edu.tw

摘要：歷史的功能在於讓後人借鏡，除了具有鑑古知今的功能外，更有「讀史明智」的說法。歷史警察知識庫目的在於將歷史警學文獻作現代警察實務應用，在初期則以《折獄龜鑑》為例，並嘗試提供「義類關鍵詞」檢索概念。本在著重知識庫規劃與系統環境說明，以及義類關鍵詞標註技術處理介紹，期待有更多的研究人員能投入歷史警學與現在警學接合交融的工作中，並在實務應用上更具前瞻性。

關鍵詞：警察實務案例、關鍵詞抽取、義類關鍵詞

綱目：

- 一、前言
- 二、歷史警學知識庫規劃與系統環境
- 三、義類關鍵詞標註技術
- 四、結論

一、前言

歷史的功能在於讓後人借鏡，亦即有所謂的「鑑古知今」或「鑑往開來」。在悠悠歲月，文明進展的過程中，前人將其生活經歷、人我際遇及互動過往，或以刻板、或以書簡記錄下來供後人參閱，尤有甚者，出版專書或有他人專作注譯、解說，故有西漢司馬遷的「史記」、北宋司馬光的「資治通鑑」或更多的經典史料能在後世流傳久遠，無非就是彰顯歷史文獻對後人的教化功能，因此，「讀史明智」。

在時間長足的堆積下，歷史文獻種類繁多且功用不一。對歷史警學文獻而言，曾榮汾(2007)在「歷史警學文獻的整理與實例」乙文中，根據《叢書子目類編·史部·政書類》及《中國法制史書目》，彙整了有檢驗類、治獄類、判牘類、成案類、慣例類、審判紀錄類、讞獄故事類等，計 105 部可供警學發展參考。然而，人力時而有窮，諾大的警史文獻真要拿來援引，卻也是有心無力。基此，資訊管理的功用在此處卻可以得到發揮。一般而言，資訊管理的功用則是在管理觀念的牽引下，透過資訊科技將繁雜無序的資料，作有意義的整理與分析成錯落有致、脈絡可尋的資料集，甚至發展成知識庫作更廣泛的應用。

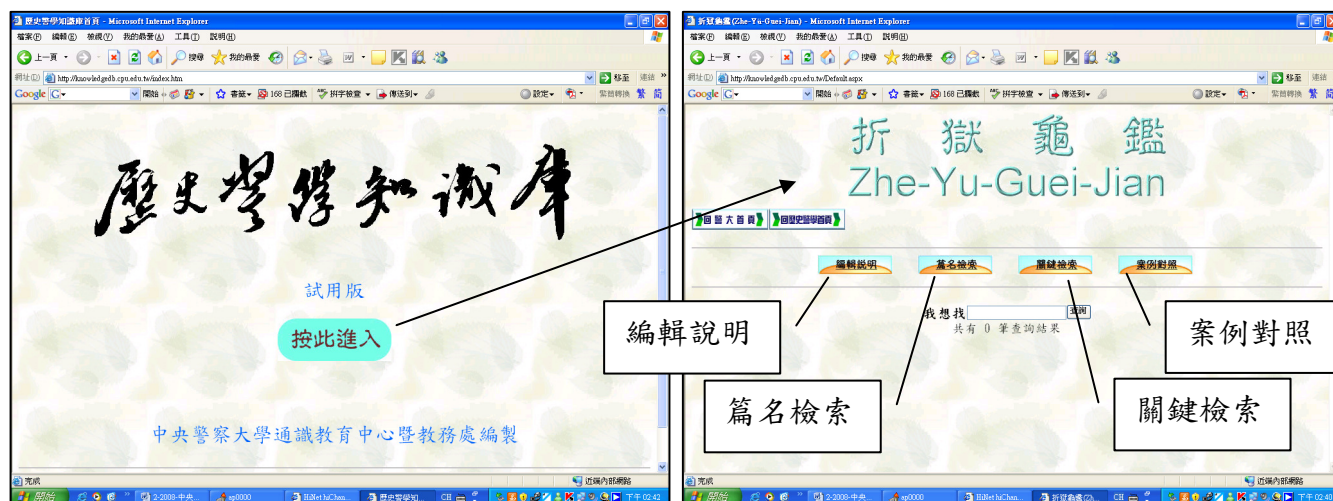
「歷史警學知識庫」為曾榮汾(2005)於「整理《折獄龜鑑》的觀念析述」中提出，並在「歷史警學文獻的整理與實例」作警學文獻整理與實例說明，企能將所有古今警學相關的知識彙整成一個綜合資料庫，貫穿「歷史」、「現代」和「未來」。「歷史」部分為容納歷史文獻和過去的案例，「現代」部分則整合現代文獻和案例，至於「未來」部分，則朝向融彙前二者知識，進而推導出開創性的學理與可能的應用模式。據此，筆者亦提出警學知識庫建置藍圖(古永昌等, 2007)，期能在資訊管理的觀念及技術的輔助下，將歷

史警學文獻結合現代警學實務作最有價值整合及應用。考量時間與人力分配，無法直接就 105 部歷史警學文獻進行整理，故僅先《折獄龜鑑》一書為例，並分作二部分。第一部分將在曾榮汾(2008)在「歷史警學知識庫的第一步—《折獄龜鑑》的整理與成果」乙文中提出整體體例設計與警察語言檢索方法；第二部分則是由筆者進行歷史警學知識庫建置說明與關鍵詞標註技術。

二、歷史警學知識庫規劃與系統環境

如前所述，本文所指之歷史警學知識庫建置規劃係以《折獄龜鑑》這部文獻作日後其他警學文獻的體例參照樣版。在體例設計中有二個重點，一為警學文獻之現代用語解說；二則為透過關鍵詞檢索方式將文獻文本串連起來，以綜合觀之與相關文本比較。「警學文獻之現代用語解說」指的是以現代用語對歷史文獻進解說，此種解說非僅將原本為文言文的文獻進行翻譯成現代之白話文，而必須對文獻內容進行考究詮釋，以及採用「警學觀點」進行詮釋解說，如此方符合警學應用要求。因此對於解說者，除了必須具備深厚的文史學養，更要求警學背景知識(police domain knowledge)，此種警學背景知識須長期的警學研究投入、對警史沿革探究理解、或充實的警察實務經驗方能竟其功。在《折獄龜鑑》則以對警史沿革素有研究及具深厚文史學養的曾榮汾教授作肇始，以拋磚引玉的態度期能吸引更多的警學研究人員參與，而能豐富歷史警學知識庫的質與量。

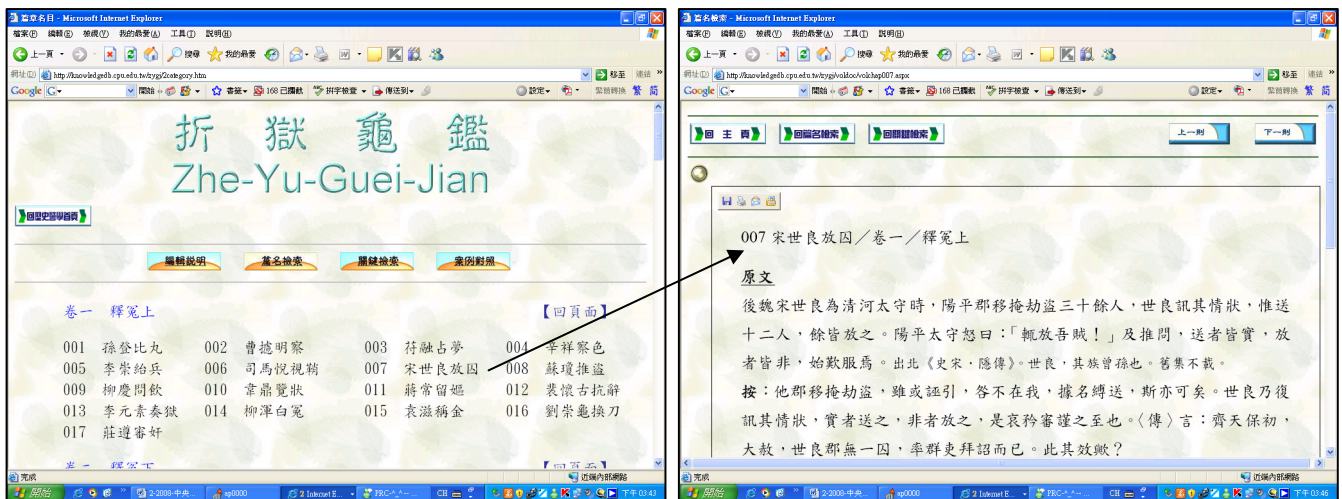
在《折獄龜鑑》頁面的配置上，約有四個部分，包括編輯說明、篇名檢索、關鍵檢索及案例對照，如圖一所示。



圖一：歷史警學知識庫《折獄龜鑑》頁面功能配置

其中：

1. 編輯說明：係介紹《折獄龜鑑》，以及說明編輯理念；
2. 篇名檢索：係對《折獄龜鑑》內各文本以現代用語解說，讓讀者可以很快地掌握文本內容，及思索可資借鑑之處，如圖二所示；
3. 關鍵檢索：係提供預設的關鍵詞供檢索，將具有相同的關鍵詞的篇名聚集在一起，以綜合觀之及參照，如圖三所示；關鍵詞給定的方法係採用「義類關鍵詞」的概念處理，「義類關鍵詞」的定義在曾榮汾(2008)中有詳細說明，本文中不再贅述，惟將在下一節中提出關鍵詞的標註方法。
4. 案例對照：係將現代的案例與文獻中相同的案例作對應，可以用來觀察不同年代對於案例處理方式的不同與借鏡。但因各項配套資源並未具備，「案例對照」功能暫無法啟用，須待更多的研究人員投入及協助收集現代案例及作適當的與文獻案例作串連。



圖二：《折獄龜鑑》之「篇名檢索」實例；以「宋世良放囚」為查詢值



圖三：《折獄龜鑑》之「關鍵檢索」實例；以「上吊」為查詢值

其次在系統環境規劃上，則如下：

1. 採用 Windows 2003 Server 為作業系統，並以 Microsoft IIS6.0 為網站伺服器；
2. 安裝 SQL Server 2005 並建置歷史警學知識庫之資料庫環境；
3. 安裝 .net framework，採用之程式語言為 vb.net 及 asp.net，以便利知識庫網頁編製以及動態查詢。

三、義類關鍵詞標註技術

如曾榮汾(2008)所言，「義類關鍵詞」為從現在警務實用角度出發，利用「警察語言」去標註每一文本的關鍵詞。「義類關鍵詞」不只是從文獻文本內容中的文字去摘錄，更是就文本內容情節去與現代警務需求作關聯。雖在歷史警學知識庫建置之初，可採取借用領域專家(domain expert)的方式進行關鍵詞或義類關鍵詞標註工作，但日後納入大量歷史警學文獻後，人工的標註方式將會耗費大量的人力、物力。基此，本文中將思考採用資訊技術協助關鍵詞標註的方法，進行文獻文本關鍵詞抽取與義類關鍵詞標註。

進行關鍵詞抽取與義類關鍵詞標註的工作，原則上有三個核心步驟需要考慮，包括斷詞與關鍵詞抽取、領域專家挑詞與標註義類關鍵詞、關鍵詞分類，以下將以《折獄龜鑑》為例分別介紹各步驟的處理方式。

(一)斷詞與關鍵詞抽取

在中文的文本裡，詞與詞之間是沒有明顯的區隔，詞可由一個以上的相鄰中文字所組成。與英文文件的相異點在於：中文文本如果沒經過中文斷詞技術前置處理，將無法拿來作後續文本分析處理，因此發展適當的中文斷詞技術是重要的。過去有許多的斷詞技術不斷的被發表出來，而最常見的中文斷詞技術主要可以分為三大類：詞庫式斷詞(word identification)(Fan&Tsai, 1988; Sproat&Shih, 1990)、統計式斷詞法(statistical word identification)(Chen&Liu, 1992)與綜合前面兩種斷詞方法的混合式斷詞法(hybrid word identification)(Nie et al, 1996)。詞庫式斷詞法通常配合詞庫或辭典一起運作，根據一些規則逐步排除不可能的詞語組合，達到較好的斷詞結果，但由於受到詞庫品質的影響，當句子中出現新生的詞彙，將使斷詞正確性降低；若要提高斷詞正確性，則應不斷新增詞庫詞彙，如此則會大幅降低實際斷詞時的效率(Chen et al, 2000)。統計式斷詞法是基於對語料庫(corpus)語詞統計訓練，以鄰近字元同時出現的頻率(bi-gram)高低作為斷詞依據，不依賴詞庫或是文法就能夠擷取文本中的詞組，優點在於執行效率高，但所斷出的詞容易因使用者語彙習慣而有影響。混合式斷詞法則是綜合以上兩種方法，利用詞庫找出不同組合的詞彙，再利用詞彙的統計資訊找出最佳的斷詞組合，如 Yeh 與 Lee (1991)提出以聯併(Unification)為基礎的斷詞法，先利用詞典搜尋可能的斷詞組合，接著利用構詞規則簡化斷詞組合，再以一階馬可夫機率模式排列出所有可能的結果，然後依照機率值排列所有可能的斷詞組合，最後使用 HPSG 剖析器(Head-driven Phrase Structure Grammar Parser)逐一過濾這些斷詞組合，確認該斷詞組合是否合於文法(Yeh&Lee, 1991)。

在本文中，採用的則是 Chien (1997)提出了適合應用在中文字擷取的 PAT-Tree-Based 方法，屬於統計式斷詞法，其方法為將文字分成三種概念，包括詞彙樣式(Lexical Patterns, LP)、完整詞彙樣式(Complete Lexical Patterns, CLP)、重要詞彙樣式(Significant

Lexical Patterns, SLP)等。其中 LP 定義為詞彙必須有一個字元以上，並且在某類文本中有一定程度出現次數；CLP 定義為詞彙必須是 LP，並且還具有完整的意思；而 SLP 定義詞彙必須是 LP 及 CLP，並且還需是明確的或重要的。如以「折獄龜鑑」字樣為例，二字詞之折獄、獄龜、龜鑑、三字詞之折獄龜、獄龜鑑及四字詞之折獄龜鑑都是 LP，而折獄、龜鑑、折獄龜鑑是 CLP，而龜鑑、折獄龜鑑則是明確且重要的，因此為 SLP。

(二)領域專家挑詞與標註義類關鍵詞

在經過文本斷詞後，為了找出《折獄龜鑑》與警務工作有關的詞彙，初期必須透過領域專家進行文本中警務用語的詞彙挑彙，且由該領域專家就《折獄龜鑑》中文本作義類關鍵詞標註，其處理程序以《折獄龜鑑》中《卷一／釋冤上／003 苻融占夢》作說明。

《苻融占夢》¹之文本關鍵詞為：占卜、謀殺；義類關鍵詞為：釋冤、超能力、鬼神、冤獄、刑案、通姦、迷信、夢境、洗冤、測字。

案例解說

前秦的苻融擔任職司隸校尉時，有個京兆人叫董豐的，外出求學，三年後歸來，路過妻子娘家，當夜就住了下來。哪知當天夜裏，妻子竟然被賊殺死。妻舅懷疑是董豐殺的，就把他押送到官府。董豐熬不過嚴刑拷打，只好認罪。苻融在核查案情時感到可疑，就問董豐道：「你出外和返家的路上，可曾發生過什麼奇怪異常的事？有沒有請人占卜過？」董豐說：「當初要離家啟程時，夜裡曾夢見騎著馬向南渡河，接著回頭向北渡河，然後又再次從北向南渡河。這時馬就停在水中，鞭策牠也不走。我低頭一看，見水中有兩個太陽，馬身的左側是白的，溼的；右側是黑的，乾的。醒過來以後，心裡很不安，總覺得不吉祥。想不到動身要回來的那天夜裡，又做了相同的夢。我去請筮者解夢。筮者說：『你要擔心吃上官司，這幾天要遠離枕頭，避免沐浴洗髮。』那天回到妻子娘家後，妻子準備好熱水，讓我洗髮；夜裡遞給我枕頭，讓我枕著睡。因為我牢記筮者的告誡，都沒有接受。妻子便自己洗了髮，枕著枕頭睡了。」苻融說：「我知道凶手是誰了。《周易》上講，坎為水，離為馬。（中略）馬身左側溼，溼就是水，左水右馬，是個『馮』字；兩個日，是個『昌』字。大概是叫『馮昌』的人殺死了婦人吧？」於是進行搜查，果然逮捕了馮昌，加以審問後，馮昌全部招供，認了犯行。他說：「本來與董豐的妻子約好謀殺董豐，以新洗頭髮和枕枕而眠為記號，哪知陰錯陽差，誤殺了婦人。」

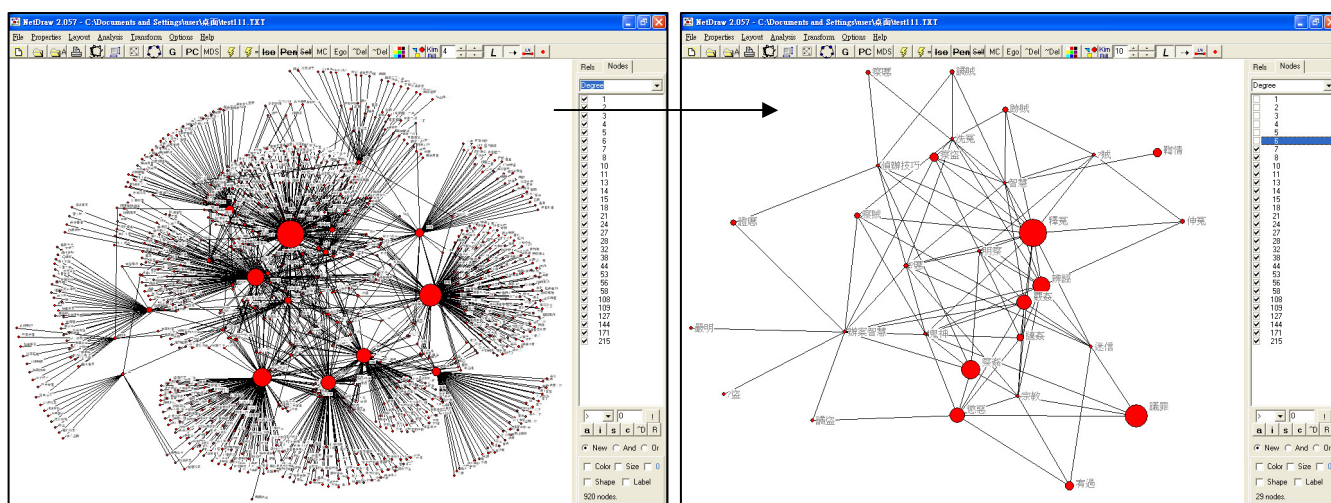
按：古人為了查案，方法真多。無論卜筮、怪異之事，只要是線索，都會盡心去查。因為只要至誠地去哀憐無辜，冥冥之中，就會獲得神助。也因此馮昌只好認罪，董豐之冤得以洗雪。不過「馮」字的「馬」邊不是三點的「水」字，而是兩點的「欠」字；「昌」字的「日」下也不是「日」字，而是「曰」字。苻融是根據猜想來推測，結果都應驗了。這正是魏國周宣所謂的「神靈動君使言」（神明使君如此說話），難道不是至誠哀憐無辜才得以如此的嗎？

¹ <http://knowledgedb.cpu.edu.tw/zygj/voldoc/volchap003.aspx>

(三)關鍵詞分類

關鍵詞為提供檢索文獻文本用，其必須具有代表性，故除了收錄各文本所訂的主題名目，如釋冤、辨誣、鞠情、議罪、宥過、懲惡、察姦、覈姦、摘姦、察慝、證慝、鈎慝、察盜、跡盜、譎盜、察賊、跡賊、譎賊、嚴明及矜謹等外，還包括因警察實務需求之義類關鍵詞，如在執法用語上常用的有辦案智慧、偵辦技巧、侵佔、偵辦、密告、收賄、偽告文書、執法態度、眼線、鑑識、賄賂、收賄、證物、破案線索、嚴刑逼供、誣告、行政疏失、心防、文書鑑定、用刑過當；及在事件上常用的有宗教、鬼神、迷信、神棍、逃兵、除籍、婆媳、屠夫、棄屍、設局、弑親、自殺、官商勾結、爭奪財產等。

為了將關鍵詞作有效的分類，本文中以各文本之主題名目為分類主軸，透過視覺化圖形呈現方式將相關的關鍵詞串連結起來，以作更直觀的觀察，如圖四所示。



圖四：關鍵詞視覺化網絡

圖四中，左圖為所有關鍵詞之視覺化網絡，而右圖則為串連程度逐次篩選的結果，從篩選的過程中可以發現，如「辦案技巧」與察慝、譎賊、證慝、辨誣及察姦等有關；而「智慧辦案」則與證慝、鈎盜、嚴明、譎盜、摘姦、察姦及覈姦等有關。此種方式除可用來直接觀察各關鍵詞間的關聯外，並可作為指引讀者在學習相類的警察實務概念時，能有更廣範的參考案例，以在未來實際面對問題時能採較周延的因應作為。另一方面，在日後進行其他文獻文本關鍵詞標註時，亦可以漸收標準化標註義類關鍵詞之效。

四、結論

本文的目的在於將歷史警學文獻作現代警察實務應用。過去的案例在法治觀念或科技水準尚不及今日的情況下，尚能在依法行政、體察民情或偵辦程序上有所著重及充滿智慧展現，相對於民智大開、案情更趨複雜的今日，警察人員應思考更周延的因應作為，過去的案例或能提供相關指引參考。

《折獄龜鑑》為歷史警察知識庫之第一步，尚有大量的案例文獻待整理與建置，其相關應用亦可透過不同實務需求進行開發，期待有更多的研究人員能投入歷史警學與現在警學接合交融的工作中，並在實務應用上更具前瞻性。

參考文獻

1. 曾榮汾，「歷史警學文獻的整理與實例」，中央警察大學通識教育學術研討會論文集，2007.
2. 曾榮汾，「整理《折獄龜鑑》的觀念析述」，中央警察大學通識教育學術研討會論文集，2006.
3. 古永昌、廖學進、古美惠，黃澡備，「社會輿情收集機制與警學知識庫建置規劃初探」，中央警察大學通識教育學術研討會論文集，2007.
4. 曾榮汾，「歷史警學知識庫的第一步—《折獄龜鑑》的整理與成果」，中央警察大學通識教育學術研討會論文集，2008.
5. J.S. Chen, C.L. Hsieh and F.C. Hsu, "A Study on Chinese Word Segmentation Genetic Algorithms Approach," *Information Management Research*, Vol. 2, No. 2, pp. 27-44, 2000.
6. K.J. Chen and S.H. Liu, "Word Identification for Mandarin Chinese Sentences," *Proceeding of COLING-92, 14th Int. Conf. On Computational Linguistics*, pp. 101-107, 1992.
7. L.F. Chien, "PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval," *Proceedings of the 1997 ACM SIGIR*, pp.50-58, 1997.
8. C.K. Fan and W.H. Tsai, "Automatic Word Identification in Chinese Sentence by the Relaxation Technique," *Computer Processing of Chinese and Oriental Languages*, Vol. 2, No. 4, pp. 33-56, 1988
9. J.Y. Nie, M. Brisebois and X. Ren, "On Chinese Text Retrieval," *Proceeding of SIGIR*, pp. 225-233, 1996.
10. R. Sproat and C.W. Shih, "Statistical Method for Finding Word Boundaries in Chinese Text," *Computer Processing of Chinese and Oriental Languages*, Vol. 4, No. 4, pp. 336-351, 1990.
11. C.L. Yeh and H.J. Lee, "Rule-Based Word Identification for Mandarin Chinese Sentences-A Unification Approach," *Computer Processing of Chinese and Oriental Languages*, Vol. 5, No. 2, pp. 97-118, 1991.

