# 行政院國家科學委員會專題研究計畫　成果報告

# 建立醫學教育上的生醫結構式與非結構式資料之知識建構與管理系統研發(I)

計畫類別：個別型計畫
計畫編號：NSC92-2524-S-038-001-
執行期間：92 年 05 月 01 日至 93 年 04 月 30 日
執行單位：臺北醫學大學醫學資訊研究所

計畫主持人：蔣以仁
共同主持人：劉致和

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫涉及專利或其他智慧財產權，1 年後可公開查詢

中　華　民　國 93 年 9 月 9 日

# 建立醫學教育上的生醫結構式與非結構式資料之知識建構與管理系統研發

Building a biomedical structural and unstructured data collection and management system for medical education

蔣以仁

I-Jen Chiang [ab]

[a] 台北醫學大學醫學資訊研究所 .[b] 台灣大學醫學工程學研究所

主持人: 蔣以仁 計畫編號: NSC-92-2524-S-038-001

## 一、中文摘要

本計畫旨在建立一知識蒐錄整理並管理知識的平台，能蒐集文獻並配合實際案例做搜尋，協助醫學院學生或住院醫師進行臨床學習。檢驗、診斷與處方等為結構化的資料，病歷報告、醫囑、及文獻等皆為文字是非結構式資料，要協助醫學院學生或駐院醫師執行有系統的學習，就在於將這些資訊有序化成組織式的規則知識，搭配臨床個案，成為 Evidence-based 之 case-based 式之學習，並將所擷取的知識以視覺化知識網絡圖加以呈現，期盼臨床醫師能透過該知識建構及管理系統，在面對病患時能提供快速且精確查檢資料與相關概念分析圖，藉由一些相關文獻的佐證以作出對病患最佳的處置。

本研究系統架構中有兩個重要模組：（1）為自動分類的訓練模組（2）為階層式知識分類模組，皆以貝氏(Bayesian)定理為主要方法結構；以其概念將文件向量化對所使用的詞庫進行比對找出字詞關聯性，利用系統中的文件自動分類技術，經臨床專家指導式學習所產生的分類規則，準確進行文獻的分類，提供使用者能在眾多資料中精確得到所需文獻。

本研究採用較高標準，所以並未將所有資料庫非相關文獻納入分母（樣本母數）計算，而是將經由關鍵詞檢索後所製成的資料庫中擷取訓練樣本（305 篇文獻），經專業臨床醫師對此系統進行樣本訓練後，對測試樣本（108 篇文獻）進行評估，結果發現觀察其正確率高達 95.4％。

**關鍵詞**：知識管理, 文件探勘, 醫學教育, 醫學文獻

## ABSTRACT

Medical knowledge needed for physicians in clinical practice is getting complex and rapidly expanding nowadays. In order to provide an optimal patient cares in clinical practices, physicians need to get concise and precise evidences from information in time. This framework will design and develop a knowledge management system for the structured data management.

**Keywords**: Knowledge Management, Text Mining, Medical Education, Medical Literatures.

## 緒論

因醫學上大部分的診斷與處置，常在極高不確定性下進行決斷，當醫師面對病患下達診斷決策時，須伴隨著各種可能性，附隨著相對之機率值，作可能性推估，以決定下一步的檢驗或治療處置；就此，條件機率型的判斷過程嚴然成型，貝氏推論模式正洽好可滿足此一在不確定因素狀況下進行推論之方法，本計畫將以貝氏圖形推論模式正式建構整個醫學教育上的生醫結構式與非結構式資料之知識建構與管理系統研發。

健康照顧的品質一直是現代我們所追求的，其好壞則決定於醫師的決策，然而決策的成效則完全仰賴於醫學知識 [1]。生物醫學知識的大量分散於迅速的累積的醫學文獻中，醫學文獻是透過許多嚴謹的臨床實驗撰寫而成的非結構性資料，故其中蘊藏大量的知識。[2]隨著電腦科技以及網際網路的普及，醫學文獻資料庫也已由傳統光碟轉變成線上電子格式，並可透過網

際網路隨時查檢，但同時也使得醫事研究人員面臨資料量過多的問題，根據統計，每20年醫學文獻的量將增加一倍[3,4]，因此若能利用文件探勘技術，經由專業領域人士對於文獻探勘系統進行訓練，對於大量文獻進行自動彙整、自動分類以及概念分群，並透過此醫學文獻知識組織及管理的平台提供查檢、瀏覽，以及對醫學詞彙間概念關聯性進行預測，讓醫事研究人員能快速由大量文獻資料中得到下達決策前所需的可能性推估。文件探勘的定義為「從非結構性或半結構性的文字中發掘出所隱含有用或是有意義的片段、模型、方向、趨勢或規則」，也可定義為「分析文件並由其中擷取重要資訊的過程」，唯有經過探勘的階段，才能將資料或資訊轉化而為知識，否則所有的資料或資訊都將只是缺乏意義的數字與符號，而無法被應用。

要如何從醫學文獻內容中找到有用的知識，就是將文件探勘運用於醫學資訊中的重要議題，醫護人員在臨床上經常須面臨疾病的巨大改變，諸如過去AIDS、到前一陣子的SARS，未曾面對的疾病，則必須依賴新的臨床醫學知識的散佈來達成，這些知識必須經由新發表的醫學文獻之蒐集獲得。因此要增加醫師面對病人以正確下決斷的能力，其中之一的挑戰就是訓練醫師決策支援的程序，資訊系統正好提供相當的支援[5, 6, 7]，透過系統的自動分類，幫助臨床醫師能快速且精準的檢索到所需文獻，並以知識網絡圖將知識間的關聯性加以表達。

為了增加臨床的決策品質，就必須仰賴醫師本身過去的經驗，教科書、文獻、回顧、以及專家所提供的證據[8,9,10]，除了經驗的傳承、實驗以及文獻的佐證外，針對其特定領域的環境需求，辨識出使用者所需組織的知識概念架構、資訊需求，以建構使用者導向的分類模組，期許達到快速搜尋所需佐證資料，以進行臨床決策的目的。

傳統的醫學教育是不斷透過各項實習主題進行，卻往往無法達到預期成效[11]，因此，對於臨床醫師長期教育以及專業資訊技術，必須將資訊檢索以及支援決策加以結合，以作為一個醫學知識系統平台之框架，而此系統必須具備幫助醫師對於非結構性的醫學文獻進行知識彙整以及管理之效能。

一個優秀的醫學知識管理系統必須能將各種不同的資訊整合，以供臨床醫師快速檢索資訊，並及時透過系統對於病患照護作出最佳處置。在一個醫學知識管理系統中，文獻自動分類是一個相當重要的主題，對於醫學文獻、案例報告以及教學手冊等醫學資訊進行概念分類，以符合臨床醫師的分類概念，以求精確快速進行資訊檢索。

## 文獻探討

### (一)自動分類

由於傳統由人工進行的文件分類工作往往除必須耗費相當多時間及精力外，對於文獻內容的知識推論及分析更是昂貴及困難，透過系統學習進行自動演算分類所獲取的領域知識價值優於傳統人工分類。文獻自動分類在近些年是廣受討論及研究的議題，自從 1961 年時 Maron’s 在 ACM 中提出文件自動分類後，陸陸續續有一些文件相關分類的應用出現，例如：Harold Borko 與 Benick 則是將文件以人工先進行分類，並經由計算訓練文件關鍵詞詞庫向量及測試文件向量內積值，內積值可作為分類的依據，值越大表示相似性越大。[12] Linear Discriminant Analysis (LDA)是透過統計模組的方式對於文件進行學習，由原始模組對其維度高低所進行分類，並可萃取其相關資訊。[13]Category Discrimination Method（CDM）對於正向及負向關聯作為分類方式，以找到最佳關聯權重為特色，其精確度（Precision）與回收率(Recall)高達 74.2%.。[14]

### (二)字詞關聯性

傳統的文件探勘技術多以關鍵詞彙及其概念對文獻進行分析，但是，如同 Feldman 及同僚所發現的[15]，亦可透過對文獻進行資料探勘找出字詞間的關聯性規則，以探索出所隱含的知識，文件的自動分類主要是利用文字在文件中所出現次數的多寡、文字詞性、結合機率來做分類，並說明文件中所含的重要詞彙也就是所謂的關鍵詞可作為分類的依據。利用字詞關聯性是對於非結構性的醫學文獻進行分類

的規則，透過此關聯性規則能夠由所彙集之文獻資訊中萃取出所隱藏的知識。

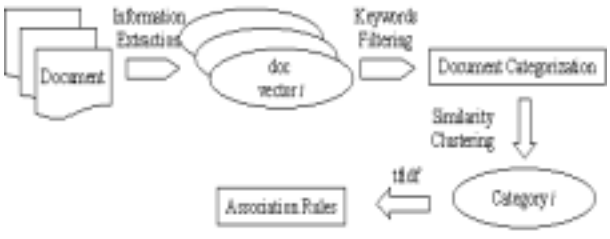
對於詞彙的選擇，需進行詞量、詞類、詞義、詞間關係以及先組合等控制，具有控制詞彙的優點，但在內容分析上最常遇到的問題是類目區分時往往缺乏該領域專家進行指導，或其分析類目無法達到使用者需求，所以跨學科文獻需配合使用者習慣及方式，以使用者的觀念進行概念分類[16]，形成一個具客制化的分類類目。

(三)類目架構

在文件探勘的技術中，如何將各類文獻精準的依主題概念分類成為其首要工作，為了達到有效且快速取得有用資訊的目的，文獻自動分類成為文件探勘系統的重要評估項目。在建立文件探勘自動分類的系統學習中發現其知識來源則取決於自然語言處理與控制辭彙索引典[17]，在資訊檢索的發展趨勢上，希望能將主題法及分類法共同整合為一體[18]，透過經專家訓練過的主題類目所形成的架構分類。

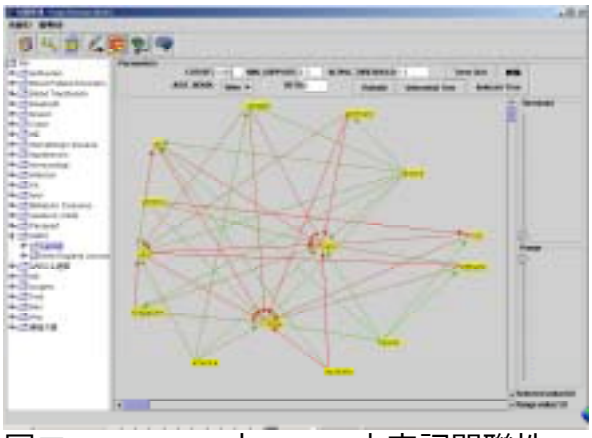
## 系統簡介

本計畫所使用系統流程簡述如圖一，是一套完全以 Java 開發 專為從事知識管理與分析等領域的專業人員所設計的專業知識發掘工具，主要在對非結構性之文字資料進行分析。Clever Craft 對擁有大量文字資料的使用者、提供利用文件探勘之貝氏網路演算法進行分析，對於文獻中概念字詞的關聯性進行分析，並以語義網路圖呈現所隱藏的知識。

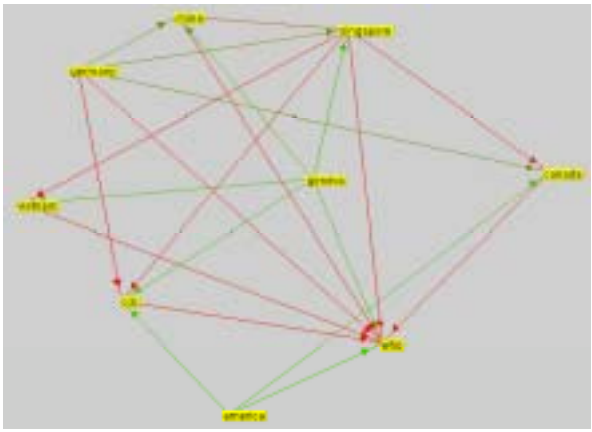



圖一. 系統流程

例如，我們由 *New England Journal of Medicine* and *the Lancet* 期刊所收集有關 *Severe Acute Respiratory Syndrome* (SARS). 的文獻資料，經由此文件探勘系統之演算法分析，並經由文獻自動分類發現，我們著重在 C.D.C., W.H.O.以及 SARS.等重要字詞以及字詞彼此忽略的影響，而導致其他概念關聯性不高的字詞被淘汰，我們可得到與 SARS 相關字詞在 *the Lancet* 期刊中，依關聯性規則結果如圖二所示，而在 *New England Journal of Medicine* 期刊中，所得結果如圖三所示。



圖二 *the Lancet*.中 SARS 之字詞關聯性



圖三 *New England Journal of Medicine* 中 SARS 之字詞關聯性

## 資料來源

文獻資料主要來自 Transfusion、Transfusion Medicine、Transfusion Science、Journal of Pediatrics、Archives of Diseases in Childhood Fetal and Neonatal Edition等期刊，由Journals@OVID電子資料庫（來源：新光醫院圖書室網站 http://library.skh.org.tw ） 及 SDOS-ES 、Blackwall Science（來源：台北醫學大學圖書室網站http://library.tmu.edu.tw）分別以關鍵字『transfusion and newborn』、『transfusion and fetal』、『transfusion and pediatrics』進行檢索，其原始檔案格式有

HTML及PDF並將檢索結果彙整，本研究由上述資料庫中共收集了約1736篇相關於輸血醫學以及小兒科（含括胎兒、新生兒、嬰兒以及幼童）文獻，並由這些文獻中擇取文獻作為訓練及評估系統學習所得的分類規則。

**類目選擇**

依據由NCBI所使用的MeSH為基本架構，MeSH是以主題概念作主軸的階層式分析主題標目，在醫學領域中最受推崇的主題標目，是採用傳統的控制詞彙可增加對某概念的用詞間劃一性，協助使用者抓住概念重點，在小兒輸血領域中，共求得7大類48小類，但其缺點則是其詞彙與使用者概念並非完全相容，控制詞彙用詞往往也不夠新穎，也因領域知識所需類目不盡相同，而導致類目過多及不足現象，經由資深血庫工作人員以及臨床小兒科主治醫師共同對於其有關於輸血醫學及小兒科相關類目進行增減，共選擇10大類21個專業術語作為分類架構，如圖四所示，以符合專業醫師對樣本文獻的分類類目。

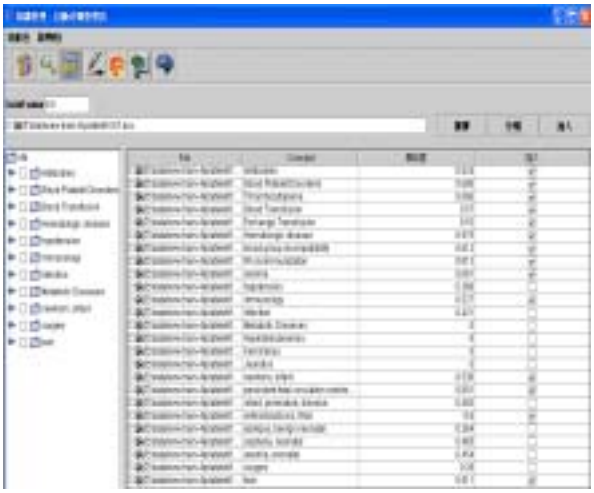
| |
|---|
| Antibodies |
| Blood Platelet Disorders |
| Thrombocytopenia |
| Blood Transfusion |
| Exchange Transfusion |
| Hematologic Disease |
| Blood Group Incompatibility |
| Rh Isoimmunization |
| Anemia |
| Hypotension |
| Immunology |
| Infection |
| Infant, Newborn Disease |
| Persistent Fetal Circulation Syndrome |
| Infant, Premature Disease |
| Erythroblastosis, Fetal |
| Epilepsy, Benign Neonatal |
| Asphyxia, neonatal |
| Anemia, Neonatal |
| Surgery |
| Twin |

圖四　　階層式類目

**系統自動分類**

該系統依據研究者對文件相似度值(cut off)設為參數值進行分類，研究者須將欲分類的文獻資料（可為單篇、多篇及一個資料夾）藉【瀏覽】按鈕納入系統，再由【分類】按鈕進行自動分類，系統會針對研究者所輸入之 cut off 值，進行類目歸類，其系統介面如圖五所示。

圖五　　自動分類介面



**指導式學習**

主要為訓練系統建立概念分類，經專業臨床醫師對此系統進行人工指導式學習樣本訓練（訓練樣本數為305篇文獻）後，力求每類目皆有相關文獻，經由系統自動分類，所得到的測試樣本（測試樣本數為108篇文獻）結果，再交由該位專業臨床醫師對系統所做的測試樣本進行評估，其所得結果以評斷該系統經指導式學習訓練後所得分類概念與該臨床專家的專業概念相似程度，以評估系統經指導式學習後的可信度。

## 結果

### （一）系統可信度評估

分析對測試文件進行內部的準確性與一致性（inter-coder agreement）評估，採用 Kappa Statistics 工具，Kappa 值主要是常用來評估原始對系統進行測試的臨床醫師與訓練後系統對測試文件結果的可信度和一致性。因此，可作為彼此測試者之間對文件內容所產生的分類是否一致。因此計算

Kappa 值來進行評估與分析彼此差異性，如表一。計算公式如下：

$$K = \frac{Po - \Pr}{1 - \Pr}$$

Po 為觀察值，Pr 為隨機值，1 為觀察值最大值。

表一　　Kappa 值評估內部一致性的好壞

| Kappa | 可信度評價 |
|---|---|
| 0.00 | 拙劣 |
| 0.01-0.20 | 微弱 |
| 0.21-0.40 | 可靠 |
| 0.41-0.60 | 可信 |
| 0.61-0.80 | 重要 |
| 0.81-1.00 | 完美 |

經專業臨床醫師對此系統進行樣本訓練（305 篇文獻）後所得到的測試樣本（108 篇文獻）結果，再交由同一位專業臨床醫師對系統進行測試，將測試樣本進行評估，臨床醫師對於各篇文件分類與系統經指導式學習後進行分類，其所得結果如表二所示。

表二　　臨床醫師與系統分類效度評估

| | | 專家 | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| 系統 | Yes | 98(90.7) | 0 (0.0%) | 98(90.7) |
| | No | 5 (4.6%) | 5 (4.6%) | 10(9.2%) |

當採相似度為 0.5（cut off＝0.5）進行測試時，臨床醫師所認為文章經系統進行指導式學習分類後，我們發現 108 篇文獻當作測試樣本時，醫師認為有 103 篇文獻可以被歸類，但有 5 篇文獻醫師並無認為有適合的類目可進行歸類；正確被歸於該類目中的文獻有 98 篇(89.8%)，臨床醫師覺得系統分類類目不妥的有 5 篇(4.6%)，然而另有 5 篇文獻(4.6%)，其相似度皆低於 0.5，並未被系統進行分類，恰巧為醫師認為非為小兒輸血領域相關文獻。所以將醫師覺得系統分類正確的 98 篇文獻加上醫師覺得無法被分類的 5 篇文獻，可觀察其正確率高達 95.4％

Observed agreement = (98+5)/108 = 0.954
Random agreement = 0.907 * 0.954 + 0.092 * 0.046 = 0.869

Kappa = (0.954 - 0.869)/(1 - 0.869) = 0.649

因此，從表格中可看出，若是測試【Clever Craft】產生的 Kappa 值大於 0.6，表示此文件分類系統經訓練後與臨床醫師的概念具一致性。

（二）、各類目精準度

本系統經小兒科主治醫師訓練後，系統與專家對測試樣本各類目進行精準度測試，由系統分類所得各篇文件應屬之類目（若內容為跨類目，則不限類目數量）交由該臨床醫師評斷其分類準確性，結果如表三所示，發現其精確度相當高，也就是說當系統經由一位臨床醫師進行指導式學習後，其分類概念相當雷同於當初對系統進行訓練人員的概念，所以該系統的指導學習式分類模組其可信度頗高。

表三　　類目精確度評估

## 討論及未來發展

在本計畫中，主要是透過自動化分類的文件探勘系統，將有關於小兒輸血的醫學文獻做精確的分類並以視覺化方式呈現，以利臨床醫師能快速搜尋所需資訊，所以致力於提升系統自動化分類之效能。

結果顯示，分類是依據物件關係將其排序分組的行為，分類的精確度與其類目的選擇能否讓使用者能精確清楚明白並符合所需甚為重要，尤其是以一個具絕對理論基礎的醫學知識領域而言，影響更鉅，使得各類目的劃一性、獨特性、特定以及直接性顯現，則其精確度平均值會相對提升，更是達到我們提供醫事人員能快速從相關文獻中精確找到所需文獻的目的，所以該系統的文件自動分類是相當值得推崇的。

初步結論可作為日後文件探勘系統以及自動分類進一步研究的佐證。茲說明如下：

一、以「使用者為導向」的分類類目

醫學研究文獻中的詞彙並非一般性用語，文獻中常出現一些特定的醫學專業術語，在內容分析上最常遇到的問題是分析類目無法達到使用者需求，所以跨學科文獻需配合使用者習慣及方式，以使用者的觀念進行概念分類，形成一個具客制化的分類類目。

二、專家詞庫斷詞佐以系統自動斷詞

在關鍵詞彙篩選上，除了系統利用向量法自動進行定詞外，仍需佐以專家對於詞彙進行人工詞彙控制並加以定詞，來彌補自然語言處理系統自動斷詞中不如人工建置詞典來得精準的缺點，並透過具控制詞彙的專家斷詞及系統自動斷詞的關鍵詞進行比對篩選排序，以達分類更加之效能。

三、知識網絡圖是具瀏覽功能的超索引

透過視覺化呈現來組織醫學知識，使臨床醫師能藉由瀏覽方式獲得所需文獻的目的，以語義距離來組織字詞，連結線及連接語則是用來表達概念之間的關係。其特徵是將知識結構化，並發展出語義的描述機制，以及著重知識關聯性，讓知識不斷的自動累積，對醫學資訊做了比較和關聯的加值動作，而成為呈現醫學知識架構的系統。

參考文獻:


[1] W. M. Tierney, M. E. Miller, J. M. Overhage, C. J. McDonald, Physician order writing on microcomputer workstations. JAMA 1993; 269: 379-383

[2] EBMR 實證醫學評論資料庫, Available at: http://www.hint.org.tw /family/research/ebm-1.htm Accessed Sep 1, 2003

[3] P.N. Gorman, J. Ash, and L. Wykoff, Can primary care physicians' questions be answered using the medical journal literature?. Bull Med Libr Assoc, 1994. 82(2):140-146.

[4] J. C. Wyatt, Knowledge management and innovation in medicine: how to go beyond practice guidelines? Advances in Clinical Knowledge Management, 2002; 5.

R. Smith, What clinical information do doctors need? BMJ, 1996. 313(7064): p. 1062-1068.

[5] B.E. Barnes, Creating the practice-learning environment: using information technology to support a new model of continuing medical education, Acad Med, 1998. 73(3): 278-281.

[6] M. Frize, F. G. Solven, M. Stevenson, B. G. Nickerson, T. Buskard, K. Taylor, Computer-Assisted Decision-support

Systems for Patient Management in an Intensive Care Unit. Proc. Medinfo '95 1995; Vancouver:1009-1012. 14.
[7] R. C. Shank, Case-based teaching: Four experiences in educational software design, (Technical support No. 7). Institute for Learning Sciences, Northwestern University, 1991.
[8] E.J. Huth, The underused medical literature. Ann Intern Med, 1989. 110(2): 99-100.
[9] D.G. Covell, G.C. Uman, and P.R. Manning, Information needs in office practice: are they being met? Ann Intern Med, 1985. 103(4): 596-599.
[10] P.N. Gorman and M. Helfand, Information seeking in primary care: how physicians choose which clinical questions to pursue and which to leave unanswered. Med Decis Making, 1995. 15(2): 113-119.
[11] B.E. Barnes, Creating the practice-learning environment: using information technology to support a new model of continuing medical education. Acad Med, 1998. 73(3): 278-281.
[12] H. Borko, and M. Bernick,. Automatic document classification. in ACM, 1963. 10(2):131-135.
[13] K. Fukunaga, Introduction to statistical pattern recognition (2nd edition). (New York, 1990).
[14] R. Feldman, Mining unstructured data in ACM SIGIR, pages 182-192, San Diageo, CA,1999
[15] R. Feldman, Y. Aumann, A. Amir, W. Kl´osgen, and A. Zilberstien. Text mining at the term level. In Proceedings of 3rd International Conferenceon Knowledge Discovery, KDD-97, pages 167–172, Newport Beach, CA, 1998.
[16] Saracevic, T. A Study of Information Seeking and Retrieving. Background and Effectiveness. Journal of the American Society for Information Science, 1988. 39(3):177-196.
[17] F. Sebastiani, Machine Learning in Automated Text Categorization. ACM Computer Survey, 2002.34(1):12.
[18] 索引典及其於資訊檢索上的探討，台大圖書館研究所， Available at:public.ptl.edu.tw/publish/suyan/36/text_46.html Accessed Sep 1, 2004