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

Biomedical structural and unstructured data collection and management system for medical education

蔡宜芬　蔣以仁　徐珮嵐　范碧琴

Yi-Fen Tsai [a], I-Jen Chiang [ab], Pei-Lan Hsu [a], Pi-Chin Fann

[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

摘要

本文旨在以一個蒐錄、整理並管理知識的平台，來蒐集文獻並自動分類。

本文將以貝氏定理為主要方向結構，建構系統中兩個重要的模組：（1）自動分類的訓練模組（2）階層式知識分類模組；以其概念將文件向量化，並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

緒論

臨床決策品質的增進，仰賴於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[1, 2, 3]。大部分醫學知識為區域性的整理。為了改進臨床決定的質量，醫師必須從課本、臨床指南、回顧、研究文章，以及專家所提供的證據去發現解答；往往需要花費許多時間、金錢、和工作量在搜尋龐大的醫學資訊術語上[4]，才能表達最理想的醫學問題解答。而生物醫學知識大量分散於迅速累積的醫學文獻中；根據統計，每２０年醫學文獻的量將增加一倍[7, 10]。

另一方面，因為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，並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；是以所有的下一步的決斷，皆是依據本次及之前得到的結論而來；就此，條件機率型的判斷過程嚴然成型；貝氏推論模式洽好可滿足此種在不確定因素狀況下進行推論之方法。

因此，我們利用文件探勘技術，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，這些知識經由新發表的醫學文獻之蒐集獲得。因此要增進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[14, 15, 16]。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。

文獻探討

■　自動化文章分類

一個有效的醫學知識管理系統必需要有效的整合大量分散異質的資源，並且提供一個快速的方法取得正確的資訊去回答在臨床照護上所遭遇的問題。

自動化文件的編目方法或分類在近些年是廣受討論及研究的議題，並且成為研究的重點達至少十年之久。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[17]。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[18]。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% [19]。而 SYNDIKATE 是自然語言分析系統，是特別為了醫療文章的文字結構所開發的。

文本資料進行學習和分類前，會將它表示成 *tfidf*（term frequency times inverse document frequency）向量形式。tfidf(i)定義如下：

tfidf(i) = TF(Wi,dj) * IDF(Wi) = TF(Wi,dj) * log(D / DF(Wi))

其中 TF(Wi,dj)表示詞 Wi 在文件 dj 中的出現頻率；D 為總文件數；DF(Wi)表示包含詞 Wi 的文件數。[22]

對所有訓練文件進行分詞處理，統計每個詞的文件出現頻率等資訊；然後根據獲取的 DF 資訊構造每篇文章的 *tfidf* 向量。

■ 字詞關聯性

我們利用名詞的關聯性進行非結構化資源分類。醫師將會同時出現的字詞，歸類在同一個類目（category），所定義出的每一類目組合成一個有意義的觀念（concept）；每一個觀念和類目之間存在著關係。這些字詞關聯性（term associations）因著不同的分類定義而不同。

傳統的文件探勘技術多以抽取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[12]，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。

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[20]，形成一個具客制化的分類類目。

■ 類目架構

文件分類自動化協助實習醫師、住院醫師和醫師管理大量的醫學資訊。所有的文件都會依據字詞關聯性被分類到適合的類目。因應醫學知識快速的擴展，系統必需動態地收集和分類來自網際網路和線上數位圖書館的資訊。

因此，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[21]，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，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

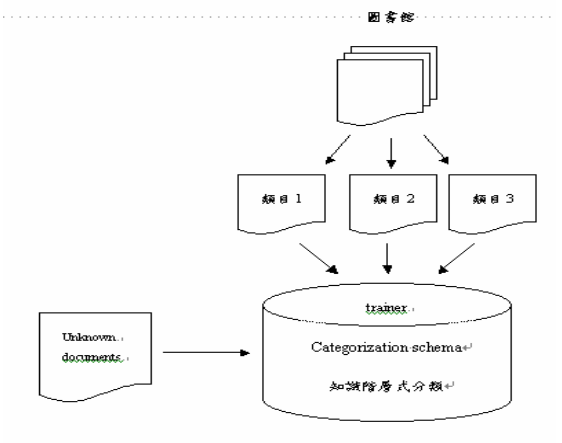
本文所採用的知識管理系統 Clever Craft，是一套完全以 Java 開發、專為從事知識管理與分析等領域的專業人員所設計的專業知識發掘工具，主要在對非結構性之文字資料進行分析。Clever Craft 對擁有大量文字資料的使用者、提供利用文件探勘之貝氏網路演算法進行分析，對於文獻中概念字詞的關聯性進行分析，並以語義網路圖呈現所隱藏的知識。其功能簡介如下：

1. 全文資訊檢索（Information retrieval）
2. 文件概念分群（Conceptual clustering）
3. 多國語言(Multi-linguistics)
4. 詞庫系統(Dictionary)
5. 統計斷詞 (Statistical NGram)
6. 分類學習(Document classifications)
7. 自動分類(Automatic clustering)
8. 自動資料蒐集下載(Automatic downloads)
9. 建立推論規則(Deductive rules)

系統架構中有兩個重要模組：（1）為自動分類的訓練模組（2）為知識階層式分類模組。

（1）自動分類的訓練模組：

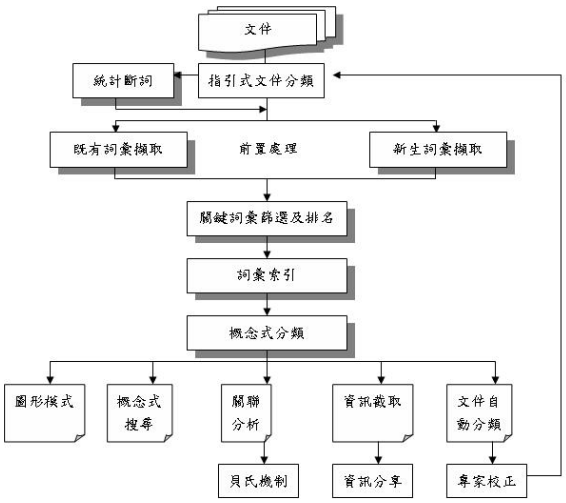
以 MeSH 作為基本主題架構類目，由臨床專科醫師與工作人員對於其主題概念進行增減，以符合臨床醫師所需的相關類目，再由專科醫師將訓練樣本（文獻）分別歸於該醫師對於每篇文章應隸屬的類目中，並讓每個類目都有一篇以上的訓練文章，使得該系統能就其字詞關聯性加以學習後，找出該領域的分類規則，另外找出測試樣本（文獻）讓系統進行自動分類，並由其他主治醫師對於系統經學習的分類規則所分出的類目結果進行評估。（見圖一）



圖一、　系統自動分類架構

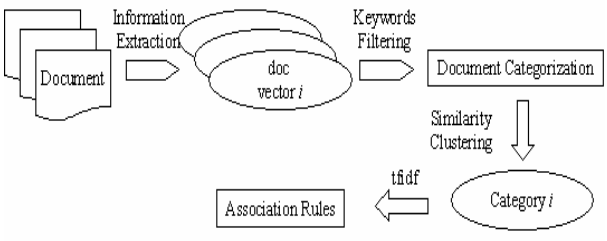
（2）知識階層式分類模組：

在此分類模組中，主要涵蓋領域專家定詞以及詞庫斷詞二種模式，再依其關鍵詞彙進行排序及篩選。在詞庫斷詞中運用自然語言處理，依詞彙所出現的頻率作為自動建構的關鍵詞，由文獻中統計配對出現的字彙，出現頻率較高的字組即視為具意義的詞彙，並以這些詞彙作為之後文章斷詞的依據，常常可透露出文獻中所隱藏的而不易被發現的知識。但是因為系統所處理的文獻為醫學研究文獻，並非一般性用語，文獻中常出現一些特定的醫學專業術語以及各種血品名稱，所以需要具有領域知識作為輔助，所以在前置作業中所使用的詞庫需有專家詞庫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，以進入指導式學習分類，依其文獻相似度（similarity）進行分類，使其歸屬於適合的類目中，見圖二。



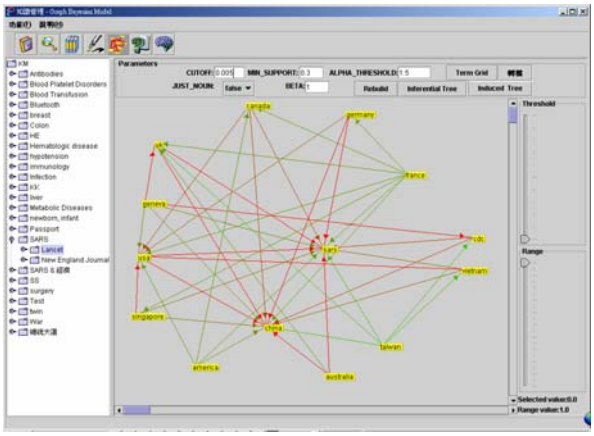
圖二、　知識階層式分類模組

系統流程簡述如圖三。所有的文件藉由去掉一些常用詞（stop word）以及 *tfidf* 值小於門檻值的詞的方式被轉換成向量。關鍵詞過濾器（*keyword filtering*）可以保留有意義的關鍵詞的字幹。文件分類會將相似度高的類目歸類到同一個類目。而每個類目都會有其關聯原則。

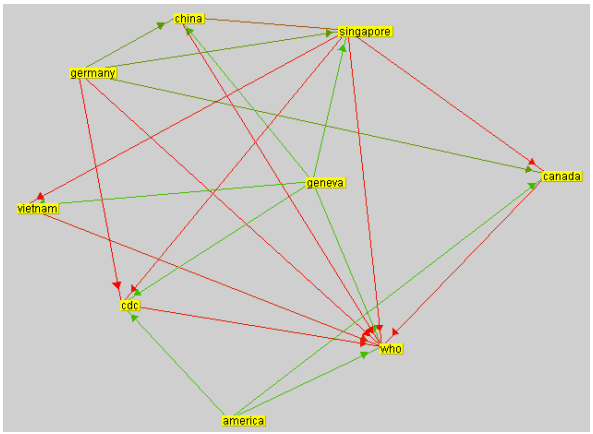


圖三、　系統流程圖

例如：我們在 *New England Journal of Medicine* and *the Lancet* 蒐集有關 SARS 的文獻資料。我們只想要看它影響的地區，因此只考慮”countries”,” C.D.C.”,” W.H.O.”和”SARS”這些重要字詞。其它概念相關性不高的字詞都被淘汰。於是在 the Lancet 期刊中，依關聯性規則結果如圖四所示，而在 New England Journal of Medicine 期刊中，所得結果如圖五所示。



圖四、　the Lancet.中 SARS 之字詞關聯性



圖五、　New England Journal of Medicine 中 SARS 之字詞關聯性

資料來源

我們從不同資料庫蒐集了許多醫學文獻，並且由小兒科醫師訓練這些文獻的分類作業。而這些技術性術語（technical term）的歸類依據是根據 MeSH 分類法；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，如圖六所示，以符合專業醫師對樣本文獻的分類類目。

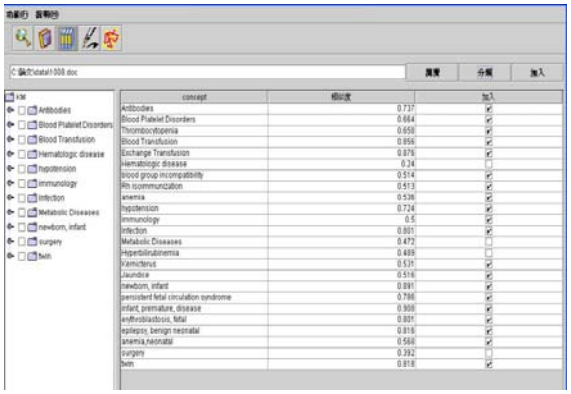
Antibodies
Blood Platelet Disorders
Thrombocytopenia
Blood Transfusion
Exchange Transfusion
Hematologic Disease
Blood Group Incompatibility
Rh Isoimmunization
Anemia
Hypotension
Immunology
Infection
Infant, Newborn Disease
Persistent Fetal Circulation Syndrome
Infant, Premature Disease
Erythroblastosis,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

圖六、　階層式類目

醫學文獻的來源資料庫有 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。在歸類方面，依據由 NCBI 所使用的 MeSH 為基本架構做分類。我們選了十個類目的二十一個專業術語。首先，我們選出 305 個訓練樣本讓醫師做分類。測試樣本（108 篇文獻）讓系統自動做分類。訓練樣本中不同的類目（表一）進行類目歸類，其系統介面如圖七所示。

表一、　臨床小兒科醫師所定義的類目

cateolory	expert judgment	classifer judgment/similarity ratio>0.5	precision
Infant, Newborn, Diseases	30	28	0.933
Anemia, Neonatal	8	7	0.875
Epilepsy, Benign Neonatal	0	0	1
Erythroblastosis, Fetal	4	4	1
Infant, Premature, Diseases	12	12	1
Persistent Fetal Circulation Syndrome	1	1	1
Blood Platelet Disorders	5	5	1
Thrombocytopenia	4	3	0.75
Hematologic Diseases	30	1	0.03
Blood Group Incompatibility	6	3	0.5
Rh Isoimmunization	6	3	0.5
Anemia	0	0	1
Antibodies	6	5	0.833
Blood Transfusion	17	17	1
Exchange Transfusion	10	10	1
Metabolic Diseases	0	0	1
Hyperbilirubinemia	2	2	1
Jaundice	1	0	0
Kernicterus	2	1	0.5
Infection	14	11	0.785
Immunology	6	5	0.833
Surgery	5	4	0.8
Twin	3	3	1



圖七、　自動分類介面

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，如表二。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。

表二、　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，其所得結果如表三所示。

表三、 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，結果如表四所示，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。

表四、 類目精確度評估

cateolory	expert judgment	classifer judgment(similarity ratio>0.5)	precision
Infant, Newborn, Diseases	30	28	0.933
Anemia, Neonatal	8	7	0.875
Epilepsy, Benign Neonatal	0	0	1
Erythroblastosis, Fetal	4	4	1
Infant, Premature, Diseases	12	12	1
Persistent Fetal Circulation Syndrome	1	1	1
Blood Platelet Disorders	5	5	1
Thrombocytopenia	4	3	0.75
Hematologic Diseases	30	1	0.03
Blood Group Incompatibility	6	3	0.5
Rh Isoimmunization	6	3	0.5
Anemia	0	0	1
Antibodies	6	5	0.833
Blood Transfusion	17	17	1
Exchange Transfusion	10	10	1
Metabolic Diseases	0	0	1
Hyperbilirubinemia	2	2	1
Jaundice	1	0	0
Kernicterus	2	1	0.5
Infection	14	11	0.785
Immunology	6	5	0.833
Surgery	5	4	0.8
Twin	3	3	1

討論及未來展望

隨著知識表現電子化的成長（高於 80%），如何找出其背後的知識是越顯重要。關聯原則是一種以文字表現知識的重要方法。根據關聯原則，我們不只能定義出文件集中同時出現的字詞，還能發掘文件中提到的事件之間的關連。

以往發掘以文字表示關聯原則的方法多用在給文件作標籤，或是從文件中抽取關鍵詞。本文則是利用文件索引來取代布林索引（Boolean indexing，單純藉由一個字辭是否出現在此文件來計算文件關聯原則的可靠度）。

此方法的好處在於，我們不需要人力去標示文件而且不需要事前準備就可以應用在不同的領域。我們的演算法可以輕易的與背景知識(background knowledge）或本體論（ontology）建立關聯性，並從特定的字詞中去發現關聯原則。解決了在 term level 關聯原則分析無法整背景知識的問題。

醫院中的檢驗、診斷與處方等為結構化的資料，而病歷報告、醫囑、及文獻等皆為文字，屬於非結構式資料，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，使得醫學院學生或住院醫師可很快速的建構自己所需的知識，進而在面對病患時能提供快速且精確地查檢資料與相關概念分析圖，藉由相關文獻的佐證以作出對病患最佳的處置。

透過自動化分類的文件探勘系統，將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。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是十分重要的，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。

參考文獻

[1] Huth E.J., The underused medical literature.

Ann Intern Med, 1989. 110(2): 99-100.
[2] Covell D.G., Uman G.C., Manning P.R., Information needs in office practice: are they being met? Ann Intern Med, 1985. 103(4): 596-599.
[3] Gorman P.N., Helfand M., 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.
[4] Smith R., What clinical information do doctors need? BMJ, 1996. 313(7064): p. 1062-1068.
[5] Godin P., Hubbs R., Woods B., Tsai M., Nag D., Rindfleisch T., Dev P., Melmon K. L., A New Instrument for Medical Decision Support and Education: The Stanford Health Information Network for Education, in Proceedings of the 32nd Hawaii International Conference on System Sciences, 1999.
[6] Huth E., Needed: an economics approach to systems for medical informaiton. Ann Intern Med, 1985. 103: p. 617-9.
[7] Gorman P.N., Ash J., Wykoff L., Can primary care physicians' questions be answered using the medical journal literature?. Bull Med Libr Assoc, 1994. 82(2):140-146.
[8] Hubbs P. R., Tsai M. C., et al. The Stanford Health Information Network for Education: integrated information for decision making and learning. In Proceedings of AMIA Annual Fall Symposium, 1997, 505-508.
[9] Barnes B.E., Creating the practice-learning environment: using information technology to support a new model of continuing medical education. Acad Med, 1998. 73(3): 278-281.
[10] Wyatt J. C., Knowledge management and innovation in medicine: how to go beyond practice guidelines? Advances in Clinical Knowledge Management, 2002; 5.
[11] Sebastiani F., Machine learning in automated text categorization. ACM Computer Survey; 2000.
[12] Feldman R., Aumann Y., Amir A., Kl´osgen W., Zilberstien A., Text mining at the term level. In Proceedings of 3rd International Conferenceon Knowledge Discovery, KDD-97, pages 167–172, Newport Beach, CA, 1998.
[13] Tierney W. M., Miller M. E., Overhage J. M., McDonald C. J., Physician order writing on microcomputer workstations. JAMA 1993; 269: 379-383
[14] Barnes B.E., Creating the practice-learning environment: using information technology to support a new model of continuing medical education, Acad Med, 1998. 73(3): 278-281.
[15] Frize M., Solven F. G., Stevenson M., Nickerson B. G., Buskard T., Taylor K., Computer-Assisted Decision-support Systems for Patient Management in an Intensive Care Unit. Proc. Medinfo '95 1995; Vancouver:1009-1012. 14.
[16] Shank R. C. , Case-based teaching: Four experiences in educational software design, (Technical support No. 7). Institute for Learning Sciences, Northwestern University, 1991.
[17] Borko H., and Bernick M.,. Automatic document classification. in ACM, 1963. 10(2):131-135.
[18] Fukunaga K., Introduction to statistical pattern recognition (2nd edition). (New York, 1990).
[19] Feldman R., Mining unstructured data in ACM SIGIR, pages 182-192, San Diageo, CA,1999
[20] 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.
[21] Sebastiani F., Machine Learning in Automated Text Categorization. ACM Computer Survey, 2002.34(1):12.
[22] Salton G., Buckley C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513-523.