利用文件及影像檢索建立胃癌診斷與治療的案例式推理

詹皇謙[a] 楊超然[a] 孫雷銘[a] 劉立[ab]

[a]台北醫學大學醫學資訊研究所 [b]台北醫學大學附設醫院

E-mail : jhcxxx@yahoo.com.tw

摘要

隨著資訊科技的快速發展，資訊科技也逐漸成為了醫療照護及生物醫學研究的關鍵工具。所以如何運用數位化的醫學資訊來幫助臨床工作者解決病患的問題，是一項重要的課題。

案例式推理(Case-Based Reasoning, CBR)是一種藉由以前的經驗，來解決目前所遭遇到的問題。這正如同人類在解決問題時，往往會師法前人十分類似，且可應用的領域相當的廣泛。其優點是可不用花太多功夫去做複雜的法則式推理(Rule-Based Reasoning, RBR)，而能快速的產生結果。在目前的 CBR 系統中，大部份都只有針對文件資料的部份進行。然而在醫學的領域中，影像資料所能提供的價值，遠超過其他的專業領域，若使文件與影像資料可合而為一，則對整體診斷成效會更有所助益。故本篇論文欲利用文件檢索結合影像檢索，增進案例式推理系統的能力，並且實際運用在胃癌的診斷上。

關鍵字：案例式推理、文件檢索、影像檢索、胃癌

前言

這些年來，雖然醫療科技與生物資訊有相當大的突破與進步，但在目前，癌症仍然是對人類健康與生命的最大威脅。而根據行政院衛生署癌症統計報告顯示，民國 91 年全國新發生癌症個案共 56,323 人，其中胃癌死亡人數為 2,446 人，其發生率排名雖有下降的趨勢，但仍居十大癌症死亡原因第五位。

早期的胃癌治愈率相當高，五年存活率可達 95%。但是一旦在胃癌的晚期才發現，其治癒率幾乎等於零。所以胃癌的防治，如能早期發現，早期治療，其痊癒率幾乎可與正常人無異。是以如何建立一個能早期診斷胃癌的機制，實為一重要的課題。

在人工智慧的領域裡，當問題領域中有清楚、簡明的知識表達；或案例的內容複雜、不易分割，及與經驗有關、重複性高的情況，案例式推理特別能發揮其功效。

而在目前的案例式推理系統中，大部份都只有針對文件資料的部份進行。然而在醫學的領域中，影像資料也提供了相當重要的資訊。胃癌的診斷，除了從臨床症候來評估之外，病患的胃鏡檢查影像的判讀，也具有相當大的重要性。故本研究計劃除了利用病患的病歷文件資料作為索引之外，也嘗試著將胃鏡的影像資料納入索引之中，進而使用替換式或轉換式案例改編法來解決案例式推理中的案例改編問題。而在案例擷取與案例相似度的計算上，則使用 CBR tool 的內建工具，另外，利用案例庫與訓練集（Training set）之間計算相似度的結果來推演文件與影像所共同構成的索引之權重，找出最適合且最能充分表現屬性的權重。

一. 案例式推理

CBR 係由 Schank & Abelson 在 1977 年從人工智慧領域中所分支出來的一套新理論與研究方法，是屬於一套依據先前經驗推論現況以處理問題的方法論，而經驗則為儲存於案例資料庫

中之所有案例(1)。其運作主要是模仿人面對問題時實際上的推理方法，由以前所遭遇過的經驗中找出最相似的案例，經由更改案例的內容以解決當前所面對的問題(2)。

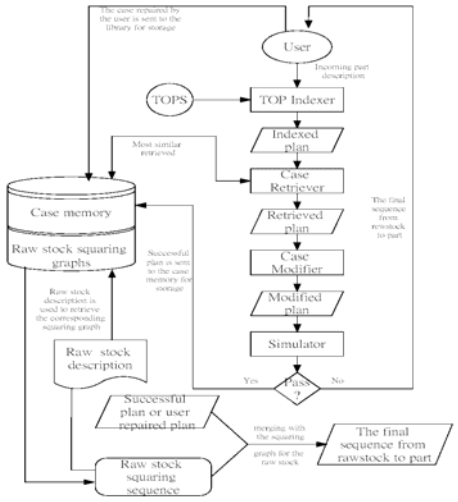

Figure CBR 架構示意圖

二. 文件檢索

文件檢索的對象是無結構化或是半結構化的文件，內容由單字與片語組成。在本研究中，對於文件檢索的定義，並非透過特殊的方式來分析或整合文件中的資訊，而是找出文件中在描述某些特定資訊時所隱含的規則，並將此規則應用於此相關特定資訊的找尋與分析。胃癌病徵間的關聯性預測，是透過文件探勘的方式尋找文件描述胃癌病徵間關聯性時所隱含之規則，並透過規則來預測關聯性(3)。

三. 影像檢索

Content-Based Image Retrieval（CBIR）是一種以影像內容為查詢對象的查詢方式。一般而言，人們在描述影像內容時，最常由顏色（color）、花紋（texture）以及形狀（shape）這三個角度來觀察。有別於傳統的文字查詢，CBIR 希望讓使用者透過影像本身的特徵來作查詢，而非單單藉由文字。

研究材料與方法

資料的前置處理

一. 資料收集

本研究收集的資料以北部某醫學中心近五年接受胃癌治療的病患， 經過篩選完整病歷記載與胃鏡檢查報告，五年內共有 340 位病患，其後排除不符合條件者共有 206 個完整的病例，並隨機選擇其中 150 位做為案例庫而另外 56 位則代入系統中作評估之用，以評估系統的準確性。

在病例經過篩選之後，醫學文件及醫學影像的資料分別建立檔案，並利用 Text mining tool 及 CBIR 之 GIFT system 為工具以建立文件與影像之資料庫，並在之後 CBR tool 建立索引檔及病例庫。

二. 文件檢索

在病例記錄內容的選擇上，本研究由於是採取手術前診斷的預測模式，故病例的內容上採用

1. chief complain, admission note, present illness, family history 及 past history
2. 實驗數據。
3. 數本醫學教科書有關胃癌部份的內容
4. 近數年有關胃癌的醫學期刊文獻約兩百篇，一同作為文件檢索的內容。
5. 建立常用的醫用文字辭庫，盡量取用與消化系統與癌症有關的醫用文字來建立辭庫，增加檢索速率。
6. 將辭庫代入 Text mining tool 中進行檢索，將病患的文件資料輸入後計算相關文字或片語出現的頻率，並以出現頻率的百分比高低做為索引的選擇與權重值的評估。
7. 參考教科書及期刊文獻等來來進行索引及權重之建立（如性別、血型、年齡……等）。並將之代入 CBR tool 中建立完整的案例庫，以供比對之用。

三. 影像檢索

本研究計劃嘗試著將胃鏡的影像資料納入索引之中，影像部分則儘量選取以病灶為中心的影像，並盡量避免選擇太小及含有內視鏡鏡身的影像，以免影響檢索的準確度。

影像以現有 open source 的 CBIR 工具為評估方式，建立影像之案例庫，並以之計算相似度。影像案例庫建立後，我們便將欲查詢之影像輸入，經過計算後新影像與影像案例庫之間影像的相似度傳回 CBR tool，並提供為 CBR system 的一個索引，進而提昇整個 CBR system 的準確度，以增進本研究在未來發展的可擴充性。

研究方法

一. 胃癌病患資料的收集及資料庫的建立

本研究胃癌病患資料的收集，即直接由院內的資訊系統內獲取。病患的收集以在醫院入院，接受胃鏡篩檢，並接受外科手術切除治療有確定病理報告者為建立病患資料的標準。並依照胃癌分期將病患分類，分為早期胃癌第Ⅰ型，第 Ⅱa 型，第 Ⅱb 型，第 Ⅱc 型及第 Ⅲ 型，進行期胃癌第一型、第二型、第三型、第四型以及惡性淋巴癌等十種類型。並利用 text mining tool 建立病患的資料庫。

二. 利用文件檢索工具建立病患文字資料庫之索引

建立病患文字資料庫之索引　　在 CBR 中，索引的建立是非常重要的，因為索引的選擇直接影響推論結果的好壞，且好的索引對於案例改編的品質也有很大的影響，而好的索引必須具有預測性(predictive)、可用性(usefulness)、具體性(concreteness)及助益性(Usefulness)。

特定字詞庫　　特定字詞庫在醫學文件檢索中佔著舉足輕重的地位，由於系統中欲處理的文件著重在於醫學研究的文獻報告，所以在文件中常會出現特定的生物醫學用語，或是基因名稱等，是故必須要有一背景知識來做為輔助，以利於往後的分析。特定字詞庫共分為六個部份，分別為疾病名稱、臨床症候與病徵、過去病史、關聯性字詞、無義關鍵字與反義關鍵字。

三. 利用影像檢索工具建立病患胃鏡影像資料庫之索引

資料索引主要的目的是要增進執行的效率。在分析一般查詢系統的執行效率時，通常可分為兩個方面：一個是『資料建構時』的執行效率（Off- line efficiency）， 另一個為『使用者查詢時』的執行效率（On- line efficiency）。『資料建構時』的工作包含了影像切割、特徵擷取，而『使用者查詢時』的工作則是尋找符合某些特徵的影像（for query by feature）、或是尋找與某個指定影像相似的影像（for query by example）。

影像查詢的資料索引技術需要解決兩個大問題：

(1) 特徵空間的維度（dimension of feature space）通常很大：所謂特徵空間，所指的是描述資料庫中影像的所有特徵表示法（影像描述）所構成的空間。通常影像查詢系統的特徵空間的維度，大致都是以百來計算。

(2) 在特徵空間中的距離度量通常不是歐式空間的距離度量法（Euclidean distance measure i.e. L2 metric）：對於大部份的特徵表示法而言，其最適合的距離度量並非 L2 metric。

為解決以上兩個問題，目前常用的方法是先對影像描述個別進行維度縮減（Dimension Reduction），再應用支援非歐式空間距離度量的多維度索引技術建立索引（Multidimensional Indexing Techniques that support non-Euclidean similarity measure）。

四. 利用替換式或轉換式案例改編法建立病患案例式推理之案例改編

在 CBR 實際應用中，或因案例庫中案例不全(incomplete case)，或因解答空間(solution space)太大，無法將典型的案例全放入案例庫。此時，如出現未曾遇過的問題，就必須藉由案例改編，以使取回的案例能有效的解決問題。除了案例擷取外，案例改編為案例式推理中另一項最重要的組件，以提供系統更周全的問題解答。部分的案例推理系統甚至只分為案例擷取及案例改編兩大區塊，兩邊可獨立運作，互不干擾。這也是為什麼有些案例改編系統不提及案例擷取，卻可正常工作的原因。

一般而言，有以下四類改編方法：

替換式改編法(Substitutional adaptation)是針對單一個特徵以替代、調整方式改變其值，不牽涉加入、減少、或重組特徵的工作。此為最基本的改編方法，大部分的案例式推理系統多少皆有使用。在問題和取回的案例十分近似時，可以發揮相當大的功用。轉換式改編法(Transformational adaptation)的作法則是以加入、刪除、或重組某些特徵來達成改編動作。當上述兩類方法皆不適用時，表示問題相當複雜，需要創新的改編方法，產生改編法(Derivational adaptation, or Generative adaptation)即是其一。本法源自推導式類推法，系統會參考以前類似的的改編痕跡(trace)，重演(replay)其改編步驟於新的問題上。一個較複雜的案例式推理系統，多半擁有數個改編方式，因此，混和上述三類技巧而成的組合式改編法(Compositional adaptation)即是常見的作法。

五. 利用 K-Nearest-neighbor 方式計算案例相似度

K-Nearest-neighbor　k 個最近相鄰法(K-Nearest Neighbor, K-NN)是距離為基礎，使用距離矩陣經排序後，用來取回(Retrieve)被預測出來的 k 個鄰近案例值，評估案例庫中的每個問題案例屬性變項的相似度，用多樣的權重因子。計算相似度的總和之計算公式表示如下；

$$Similarity(T,S)=\sum_{i=1}^{n} f(T_i,S_i)*W_i \quad (1)$$

T 是目標案例，S 是案例庫案例，n 是每個案例的所有屬性，i 是每個屬性，f 是案例庫案例中的目標案例第 i 個屬性的相似度函數，W 是第 i 個屬性的重要性權重。相似度的正規化後其值降至包含於 0 與 1 之間。0 是完全不相似，1 表示完全 100%的相似。

許多案例庫推理校正案例使用 K-NN 分類法，所有的相似函數其靈敏度是受分離的、互相的、雜訊屬性等因素所影響。所以相似度在案例庫推理中是相當重要的。而 K-NN 的演算法會因為在模擬中遭遇到相同的雜訊，這便需藉助多次嘗試做批次最佳化，使在同一組分類內其組內同質性最大，並且組與組間的變異也最大。也同時給予屬性變項權重與改善正確性。

屬性權重　各種不同的 K-NN 分類常使用於案例庫系統在取回案例時，K-NN 假設每個案例 X={X1,X2,… … .Xn }是被定義成有 n 個屬性的資料集，屬性可能是數值或是分類符號屬性，當 Xc 是 X's 中的某一種分類值。假設給定一個搜尋 q 和案例庫 L，K-NN 從案例庫 L 中取回 q's 中的 k 個最相近的案例並預測 q 的主要權重分類，且 K 值要大於等於 1，被定義公式如下；

$$\text{Distance}(x,q)=\sqrt{\sum_{f=1}^{n} W_f * difference(x_f,q_f)^2} \quad (2)$$

且 $W_f \geqq 0$ 對所有的 f

$$\text{difference}(x_f, q_f) = \begin{cases} |x_f - q_f| & \text{若屬性 } f \text{ 是數值} \\ 0 & \text{若屬性 } f \text{ 是類別且 } x_f = q_f \\ 1 & \text{若屬性 } f \text{ 是類別且 } x_f \neq q_f \end{cases} \quad (3)$$

歐幾理德距離是能夠使用在連續值與象徵值的資訊，如公式(3)對連續值與象徵值的處理。公式(1)當所有權重都是 1 則會允許重複、不適當、不好的屬性直接影響距離的計算結果而成為 K-NN 的缺失，當這樣的屬性出現時 K-NN 的效能就會變差。在公式(2)中 K-NN 是可以滿足案例庫的取回案例。

實驗結果

精確度評估

CBR 系統的評估方式一般以 P 值(Precision Value）為主：

$$Precision = \frac{number\ of\ relevant\ items\ retrieved}{number\ of\ all\ items\ retrieved}$$

If P(10)>90%, Result is Highly perfect

If 80%<P(10)<90%, Result is very good

If 70%<P(10)<80%, Result is good

在本研究中，將案例相似度的評估以胃癌之分類（Classification）, 分期（Stage）以及兩者合併（Classification +Stage）分別評估，而在案例庫的尋找中，以尋找相似度最高之前 20 名依序排列。

醫學文件查詢結果

在分類（Classification）方面，有 46.43%案例可在第一相似案例中找到相同分類的案例，而 98.21%的案例在前十個相似案例中皆可找到最相似者，P(10)=98.21。也就是說，有九成的案例可在前十個查詢案例中找到結果。而在第十七位才找到相似案例的案例屬於早期胃癌 E2a，由於此種分類的病例較少見，案例庫中的案例也比較少，可能在搜尋上會有部分誤差，值得探討。

在分期（Stage）方面，有 32.14%案例可在第一相似案例中找到相同分類的案例，而 100%的案例在前九個相似案例中皆可找到最相似者，P(10)=100%。

在合併查詢（Classification +Stage）方面，有 23.21%案例可在第一相似案例中找到相同分類的案例，而 98.21%的案例在前十個相似案例中皆可找到最相似者，P(10)=98.21% 。

以此觀之，對將來新查詢的案例，可由查詢前十個案例即可，如此可以加快查詢的速度並增加系統的效法。

醫學影像查詢結果

A. 在分類（Classification）方面：一般而言，分類大致以胃癌本身的外觀為主，雖然影像檢索之正確性略遜於文件檢索之正確性，但仍有 37.50%案例可在第一相似案例中找到相同分類的案例，而 92.85%的案例在前十個相似案例中皆可找到最相似者，P(10)=92.85%；也就是說，有九成的案例可在前十個查詢案例中找到結果，也有偏誤值，case158 在第 125 位才找到相似案例的案例，屬於胃淋巴癌，由於此種分類的病例案例庫中的案例也比較少，可能在搜尋上會有部分誤差，可利用 CBR 中的案例改編法則將此案例經改編後加入案例庫中。

B. 在分期（Stage）方面：由於在分期上，外觀反而成為較不重要的表徵，主要是以腫瘤侵犯之深度而定，故在影像檢索的難度更高，準確性也相形下降，但仍有 25%案例可在第一相似案例中找到相同分類的案例，而 82.14%的案例在前十個相似案例中皆可找到最相似者，P(10)=82.14；也就是說，有八成的案例可在前十個查詢案例中找到結果。

C. 在合併查詢（Classification+Stage）方面，只有 16.07%案例可在第一相似案例中找到相同分類的案例，而 67.86%的案例在前十個相似案例中皆可找到最相似者，P(10)=67.86%；有接近七成的案例可在前十個查詢案例中找到結果。

醫學文件與醫學影像的綜合查詢結果

由以上文件檢索與影像檢索的查詢結果來看，文件檢索的相似度查詢較影像檢索為高，本研究在醫學的範疇中將文件檢索（T）與影像檢索（G）兩者的結果分別設以不同的權重，再代入 CBR system 中求取最佳解。

本研究接下來將兩者的權重值分設定分為九組：T1G9,T2G8…到 T9G1，一一代入 CBR system 中求解，結果如下表：.

	T1G9			T2G8			T2G7			T4G6		
	C.%	S.%	C.+S.%	C.%	S.%	C.+S.%	C.%	S.%	C.+S.%	C.%	S.%	C.+S.%
1	32.14%	19.64%	12.50%	35.71%	21.43%	12.50%	41.07%	26.79%	17.86%	48.21%	28.57%	23.21%
2	44.64%	33.93%	25.00%	52.83%	39.29%	28.57%	53.57%	41.07%	30.36%	64.29%	44.64%	35.71%
3	51.79%	46.43%	32.14%	58.93%	46.42%	35.71%	66.07%	53.57%	37.50%	69.64%	55.36%	42.86%
4	60.71%	57.14%	41.07%	66.07%	55.36%	37.50%	71.43%	58.93%	41.07%	82.14%	66.08%	50.00%
5	73.21%	66.07%	46.43%	75.00%	69.64%	44.64%	76.79%	66.07%	44.64%	83.93%	78.57%	64.29%
6~10	85.71%	83.93%	55.36%	89.29%	85.71%	50.00%	87.50%	89.29%	69.64%	94.64%	87.50%	75.00%
11~20	98.21%	92.86%	78.57%	98.21%	96.43%	60.71%	100%	96.43%	83.93%	96.43%	96.43%	85.71%
21~end	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	T5G5			T6G4			T7G3			T8G2		
	C.%	S.%	C.+S.%	C.%	S.%	C.+S.%	C.%	S.%	C.+S.%	C.%	S.%	C.+S.%
1	53.57%	30.36%	21.42%	62.50%	33.93%	30.36%	55.35%	37.50%	28.57%	50.00%	37.50%	25.00%
2	60.71%	44.64%	25.00%	69.64%	48.21%	37.50%	73.21%	55.37%	42.85%	69.64%	60.71%	46.43%
3	69.64%	60.71%	35.71%	80.36%	57.14%	42.86%	76.79%	62.50%	48.21%	76.79%	67.56%	55.36%
4	80.36%	67.86%	51.79%	82.14%	60.71%	48.21%	80.36%	66.07%	51.79%	83.93%	73.21%	62.50%
5	87.50%	76.79%	58.93%	85.71%	73.21%	58.93%	85.71%	71.43%	60.71%	89.26%	82.14%	69.64%
6~10	96.43%	91.07%	80.36%	87.50%	92.86%	87.50%	96.43%	96.43%	87.50%	96.43%	89.29%	91.07%
11~20	98.21%	96.43%	89.29%	100.00%	100.00%	100.00%	100.00%	100.00%	96.43%	100.00%	100.00%	100.00%
21~end	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	T9G1											
	C.%	S.%	C.+S.%									
1	51.79%	30.36%	21.43%									
2	73.21%	53.57%	44.64%									
3	82.14%	71.43%	60.71%									
4	85.71%	78.57%	64.29%									
5	89.29%	82.14%	67.86%	11~20	100.00%	100.00%	98.21%					
6~10	98.21%	98.21%	94.64%	21~end	100.00%	100.00%	100.00%					

Table 文件檢索與影像檢索合併後所有結果

參酌上述的結果，當文件與影像檢索權重設為 6:4 時（ 6T4G ），檢索所得之結果最佳。

1. 分類（Classification）方面結果:P(10)=87.50%。
2. 分期（Stage）方面結果:P(10)=92.86%。
3. 合併查詢（Classification+Stage）結果:P(10)=87.50%。

而在將文件與影像檢索合一後，第一案例準確性 P(1) 在正確率上皆有明顯之提升

1. 於分類（Classification）方面結果：P(1)= 62.50%（Text 46.43%, Image 37.50%）。
2. 於分期（Stage）方面：P(1)= 33.93%（Text 32.14%, Image 25.00%）。
3. 於合併查詢（Classification+Stage）方面：P(1)= 30.36%（Text 23.21%, Image 16.07%）。

由此可見文件與影像合併檢索的結果較分別檢索為佳。

討論與結論

在目前已進行的案例式推理系統中，經驗的選擇大部份皆採用單純的文字經驗，即是採取前人經驗之文書資料，然後再藉這些資料分析後建立索引與權重來建立案例庫。而影像方面的案例式推理仍然少見。然而在臨床醫學的診斷與治療上，文件資料與影像相互間之關聯性與重要性是其他領域所無法比擬的。故將文件與影像資料相互整合，是建立醫學案例式推理相當重要的工作。

醫院的 PACS 系統內含豐富的醫療影像，若能適當應用本研究來輔助醫生做胃癌的早期診斷，則可以增加胃癌早期症狀發現機率，將有助於醫療品質之提升。本研究後續除累積更多的案例庫資料，亦可朝其他替代 K-NN 相似度演算法之方向提升系統之推理相似度，以增進系統品質。

參考資料

英文文獻

1. Snell: Clinical Anatomy for Medical Student; 5th edition, 1995.
2. Jeng, B.C. and T.P. Liang (1995), “Fuzzy Indexing and Retrieval in Case-Based Systems,” Expert Systems with Application, Vol. 8, No. 1, pp.135-142.
3. Aha, D.W. (1998), “The Omnipresence of Case-Based Reasoning in Science and