從生醫資料庫中探勘基因表達證據之研究

Mining Evidences of Gene Expressions from Biomedical Database

李御璽	顏秀珍	陳信希	劉又誠
銘傳大學資工系	銘傳大學資工系	國立台灣大學資工系	國立成功大學資工所
leeys@mcu.edu.tw	sjyen@mcu.edu.tw	hh_chen@csie.ntu.edu.tw	r1750032@ss23.mcu.edu.tw

摘要

從上個世紀開始，有許多的專家投入生物醫學 (Biomedical) 領域的研究，其成果也不斷地被發表出來。相對的，也有越來越多的研究者投入在生物醫學文獻的資訊擷取上。一般而言，生物醫學資料庫必須經常藉由具專業知識的生物醫學專家來作管理及更新。然而，由於生物醫學文獻的發表速度成幾何成長，生物醫學資料庫的管理和更新已遠非單純人力所能負擔。因此，本研究擬建立一個不需依賴生物醫學領域專家的電腦輔助系統，以靴帶式向量空間模型 (Bootstrapping Vector Space Model)，自動判斷在生物醫學文獻之中，是否擁有任何基因產物的實驗結果或證據。實驗的結果顯示，我們的模型得到 F-Measure 0.502 不錯的判別率，優於單純的向量空間模型及支持向量機 (Support Vector Machine) 模型。在人力無法負荷的生物醫學文獻發表速度下，本論文所提出的模型能有效的節省更多的人力、時間及資源。

關鍵詞：生物醫學、靴帶式向量空間模型、資料探勘、基因產物

Abstract

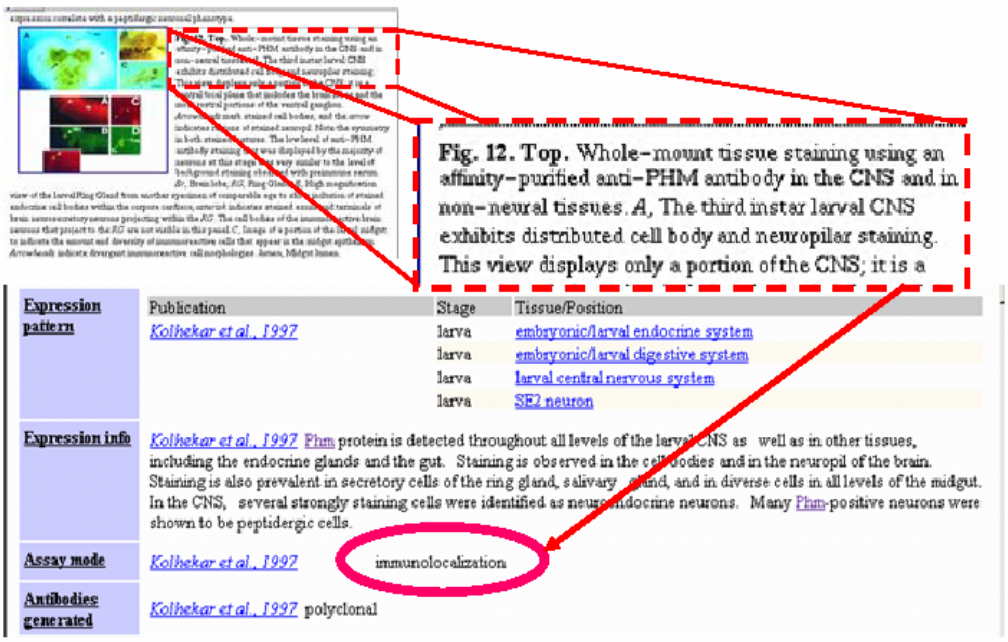
It has a large number of experts to participate in the biomedical research and result was published unceasingly from the last century. In this domain, biomedical database is the most important information storage. Frequently, the researchers want to study a kind of gene. They can search these databases to find some experimental results proposed by the previous researchers about this gene. However, a keyword-based search interface is usually provided for searching literatures in these databases. It can not satisfy the user needs, because the users will waste a lot of time to filter out some undesired results. Thus, it is very important to propose a model to automatically analyze the keyword-based search results for finding the literatures that have evidences of gene expressions. This paper proposes a classification model to generate four feature vectors based on biomedical training data. Two of these feature vectors are used to identify whether one literature has the gene product experimental results or evidences or not. The other two are used for protein identification. Besides, a bootstrapping mechanism is also adopted in this paper to filter out some noises and increase classification accuracy, to get the F-measure 0.502. The experimental results show that our model outperforms Vector Space Model.

Keywords: Biomedical, Bootstrapping Vector Space Model, Data Mining, Gene Product

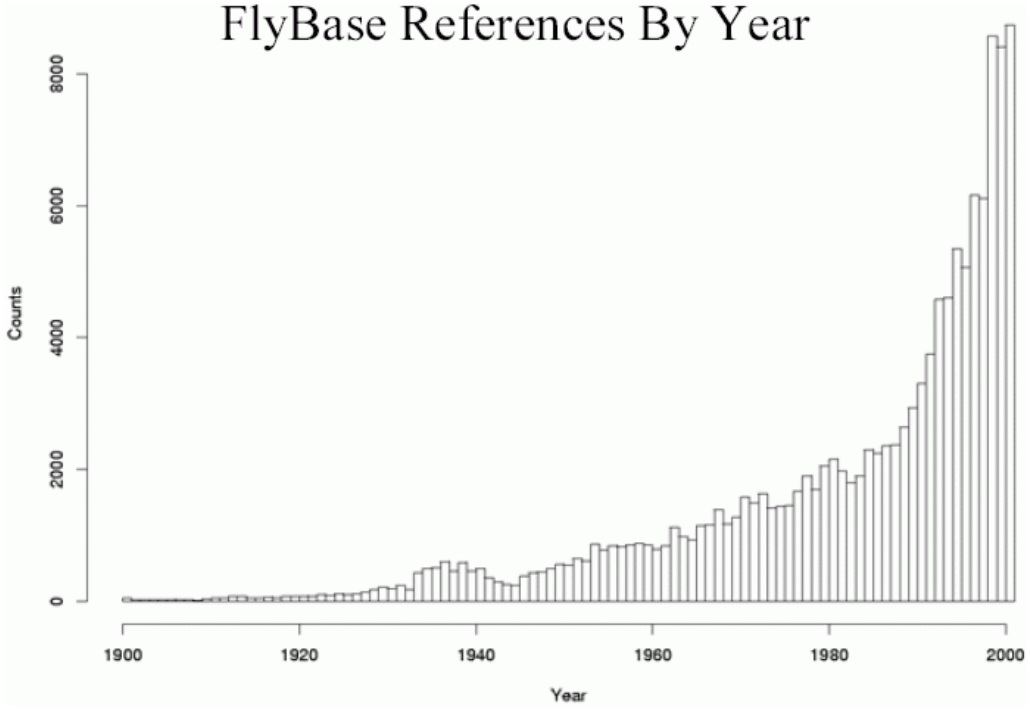
一、緒論

近幾年來，由於大量的研究資源及人力投入生物科技的領域，使得生物科技發展迅速，而其研究成果也形成了大量的生物醫學文獻，並放置於生物醫學資料庫中。以生物醫學界知名的 MEDLINE (MEDLARS ON LINE) 生物醫學資料庫 [1] 為例，該資料庫每年固定收錄約 4,600 個全世界著名且權威之期刊 (若累積歷年停止收錄、停刊、更名等期刊，期刊總數則超過 10,000 個)。目前資料庫中約有 880 萬筆紀錄，並以每月 15,000 ~ 20,000 筆記錄的速度增加中。

生物醫學資料庫提供大量的生物醫學文獻。然而，除了將文獻放置於生物醫學資料庫外，生物醫學的領域專家還需將所有的文獻內容整理，擷取



圖一、果蠅文獻資料庫 FlyBase 之文獻資訊擷取範例



圖二、果蠅文獻資料庫 FlyBase 的歷年文獻發表數量

出有用的資訊，並更新 (Curated, Updated) 至生物醫學資料庫中。圖一說明一位生物醫學的領域專家如何將一篇有關於果蠅 (Drosophila, Fruit Flies) 基因實驗的文獻，擷取出有用的資訊，並更新至果蠅文獻資料庫 FlyBase 中 [2, 3]。這位領域專家認定這篇文章中的某個段落，是在敘述使用一個抗體 (Anti-Body) 作用於一個組織上，然後觀察其變化，符合生物醫學實驗方法中 Immunolocalization 檢驗的步驟。故將此篇文獻的實驗方式 (Assay Mode)，在實驗方式的欄位中記載為

Immunolocalization。

由於生物醫學界的蓬勃發展，相關研究文獻的發表速度也已經超越了人力所能閱讀、整理的範圍。圖二顯示果蠅文獻資料庫 FlyBase 所搜集到的文獻，逐年以指數方式成長 [2, 3]。傳統人工的處理方式已不敷使用。本研究的目的是希望能夠自動從生物醫學文獻中，擷取出重要的相關資訊，以協助生物醫學領域專家擷取有用的資訊。在本論文中，我們將以 KDD Challenge Cup 2002 Task 1 [4, 5] 所提供的生物醫學文獻為實驗的對象，發展一套分類系統 (Classification System) 以判別在待測的文件中，是否包含某種特定基因轉錄物 (Transcript) 或基因蛋白質 (Protein) 的實驗結果或證據。

KDD Challenge Cup 2002 為 ACM SIGKDD 在 International Conference on Knowledge Discovery and Data Mining (KDD 2002) 中所舉辦的比賽。資料的來源是取自於果蠅文獻資料庫 FlyBase。大會會給予每位參賽者：(1)果蠅基因或分子生物學的文件。(2)那些文件有包含基因產物的實驗結果或證據。(3) 那些文件有包含某種特定基因產物的實驗結果或證據。

本篇論文的章節安排如下：在下一節中，我們將介紹相關的研究工作。第三節將說明向量空間模型分類法 (Classification Based on Vector Space Model)。第四節將介紹我們所提出的靴帶式向量空間模型分類法 (Classification Based on Bootstrapping Vector Space Model)。在做結論之前，我們將比較我們所提出的方法與現有方法的優劣。

二、相關研究工作

基本上，我們可以將 KDD Challenge Cup 2002 Task 1 的問題當作是一個分類問題。Regev, *et al.* [6] 提出一個以規則為基礎 (Rule-Based) 的資訊擷取系統 (Information Extraction System)。它結合了句子的比對 (Pattern Matching)、自然語言處理 (Natural Language Processing，NLP)，以及語意條件限制 (Semantic Constraints) 來擷取文件中的証據 (Evidence)，並依此判定此文件是否包含了基因的實驗結果或證據。於文件基因結果或證據的判斷上，它可達到 F-Measure 0.78 的判別率。而於基因產物 (Gene Product) 的基因實驗結果或證據的判斷上，它可達到 F-Measure 0.67 的判別率，明顯的優於 KDD Challenge Cup 2002 Task 1 的其它參賽隊伍。

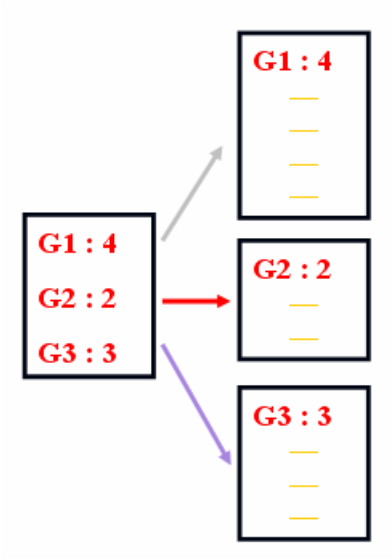
Keerthi, *et al.* [7] 提出一個以 Naive Bayes 分類器為基礎的文件分類系統 (Text Classification System)。它考慮特徵詞類別與目標基因的距離，來擷取分類系統所需的重要特徵 (Feature Extraction)。於文件的基因結果或證據的判斷上，它可達到 F-Measure 0.73 的判別率。

Ghanem, *et al.* [8] 提出以 SVM 分類器為基礎的文件分類系統 (Text Classification System)。他提出一個進階的特徵選取方法 (Advanced Feature Selection Method)。此方法擷取高頻率關鍵詞於同一句子或相鄰句子中結合的型樣 (Pattern) ，來建立特徵向量表 (Feature Vector Table)，並輸入 SVM 分類器中進行分類。於文件基因結果或證據的判斷上，它可達到 F-Measure 0.58 的判別率。而於基因產物(Gene Product)的基因實驗結果或證據的判斷上，它可達到 F-Measure 0.59 的判別率，比單純使用關鍵詞為基礎的特徵選取 (Keyword-Based Feature Selection) 方法，提供更高的分類正確率。

以上三篇是比賽中獲勝隊伍所發表的論文。在致勝的關鍵點上，他們都是大量依賴生物醫學領域專家 (Biomedical Domain Expert) 的協助。Regev, *et al.* [6] 除了使用一般的基因詞典之外，亦使用針對此問題所自行建立的關鍵正向詞詞典、反向詞詞典以輔助判斷。Keerthi, *et al.* [7] 亦從包含基因實驗結果或證據的檔案中，由專家擷取出部份關鍵詞以供分類時判斷。Ghanem, et al. [8] 則透過專家篩選可用以分類的型樣 (Pattern)，以提高分類的正確率。以上的工作，皆非一般傳統電腦資訊人員所能獨立完成。

蘇家玉 [9] 所提出的方法，是少數不依賴生物醫學專業人員的電腦輔助模型(Computer-Aided Model)。它從訓練文件中擷取實驗結果或證據的特徵，並依據這些特徵發展決策樹模型 (Decision Tree Model, DTM) 及向量空間模型 (Vector Space Model, VSM) 來判斷待測文件是否包含基因的實驗結果或證據。此方法的效能，在決策樹模型上，轉錄物的判斷可達到 F-Measure 0.3627 的判別率，蛋白質的判斷可達到 F-Measure 0.3788 的判別率。而在向量空間模型上，轉錄物的判斷可達到 F-Measure 0.3741 的判別率，蛋白質的判斷可達到 F-Measure 0.4274 的判別率，皆優於決策樹模型。

然而這兩個模型都有兩個共通的缺點，就是其訓練文件中的句數太少且包含太多的雜訊在其中。本論文所提出的方法 (靴帶式向量空間模型) 是想在不依賴生醫專業人員與知識的基礎上，改進蘇家玉所提出的向量空間模型，擴大其訓練文件的大小，且自動過濾對分類結果沒有幫助的雜訊。期望在效能上超越蘇家玉所提出的方法，並接近前 3 名所提出的方法。



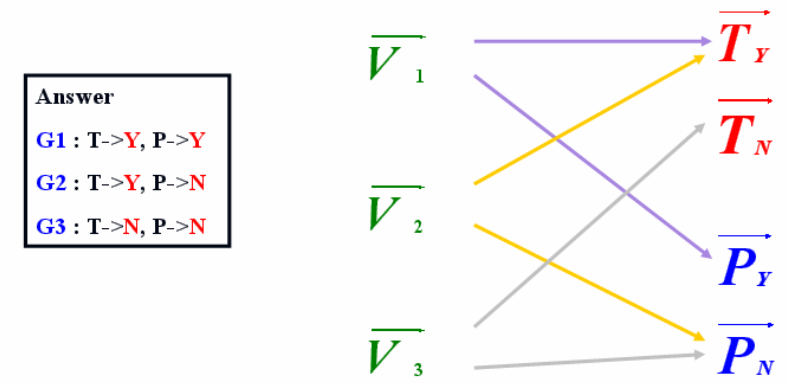
圖三、建立虛擬文件

G1	Sentence 1	t1 , t4 , t5
	Sentence 2	t2 , t4 , t7
	Sentence 3	t3 , t4
	Sentence 4	t1 , t9
G2	Sentence 1	t2 , t4 , t7, t9
	Sentence 2	t2 , t3
G3	Sentence 1	t3 , t4
	Sentence 2	t4 , t5 , t6
	Sentence 3	t4 , t7

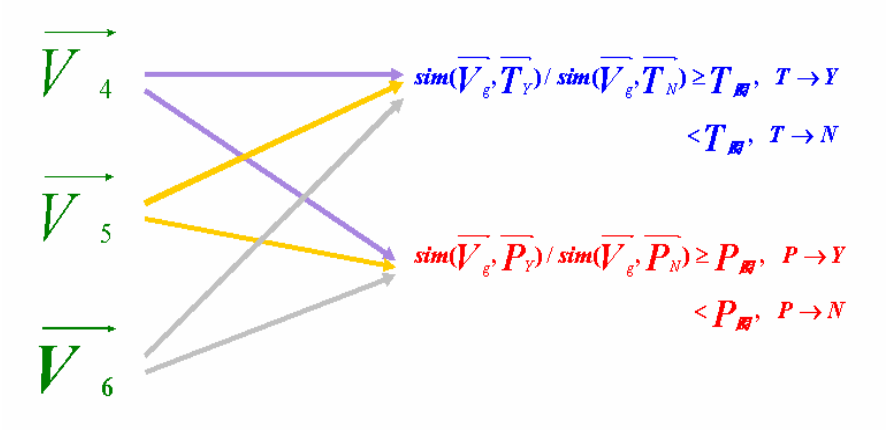
Feature Vector Table

	t1	t2	t3	t4	t5	t6	t7	t8	t9
d1	2	1	1	3	1	0	1	0	1
d2	0	2	1	1	0	0	1	0	1
d3	0	0	1	3	1	1	1	0	0

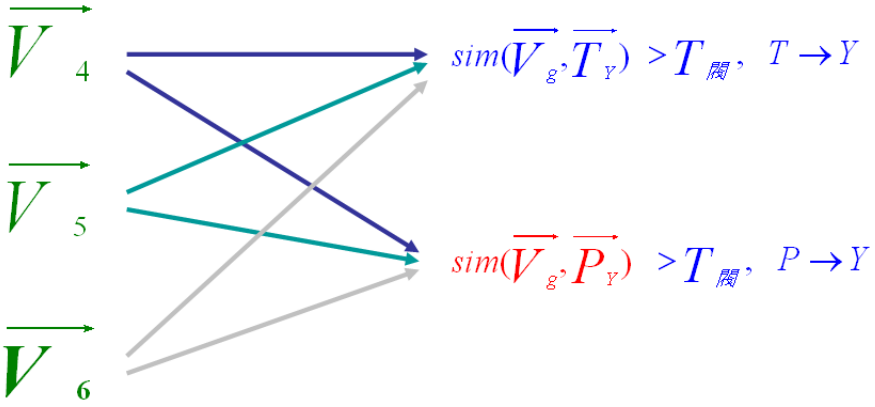
圖四、建立虛擬文件之特徵向量表



圖五、參考標準答案建立四個特徵向量



圖六、比較測試文件向量之相似度是否高於門檻值



圖七、另一種比較方式

三、向量空間模型分類法

在這一節中，我們將說明向量空間模型分類法。整個方法分成訓練及測試階段。在訓練階段，針對每個訓練文件及文件中特定的基因 (假設有 K 個)，會有 K 個虛擬文件 (Virtual Documents) 被建立出來。一個虛擬文件代表一個特定的基因。其內容就是此特定基因出現在訓練文件中所在句子的集合。圖三是一個例子。圖三的左邊有一個訓練文件，且這個文件有 3 個特定的基因。因此，有 3 個虛擬文件會被建構出來。由於第 1 個特定的基因 (G1) 出現在此訓練文件中的 4 個句子，因此虛擬文件 1 的內容就是此 4 個句子的集合，其餘類推。

接下來針對每一份虛擬文件，重要的關鍵詞 (Keyword) 會被擷取及建構出來，並建立其特徵向量表 (Feature Vector Table)。圖四是延續圖三所建立出來的關鍵詞及特徵向量表。圖四中 t1, t2, …, t9 代表 9 個關鍵詞，d1, d2, d3 代表對應 3 個特定基因的虛擬文件。特徵向量表中的數字則代表 9 個關鍵詞在 3 個虛擬文件中出現的次數。例如關鍵詞 t2 在虛擬文件 d1 中出現 1 次，d2 中出現 2 次。

特徵向量表建立完成後，每一個虛擬文件會以一個向量來表示。以圖四來說d1, d2, d3 會建立出 3 個向量V1, V2, V3，且每個向量的維度是 9。每個向量i中的每一個維度j的值$w_{i,j}$是以下公式計算得出 [10, 11]：

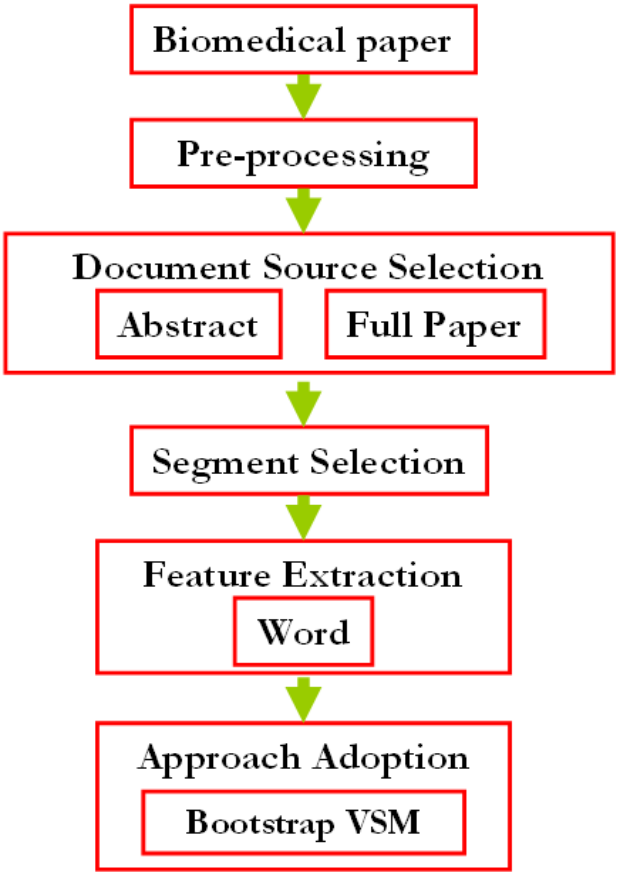
$$w_{i,j} = \frac{freq_{i,j}}{\max_{j=1,2,\ldots,n} freq_{i,j}} \times \log \frac{N}{N_j}$$

$freq_{i,j}$: 虛擬文件d_i，第j個關鍵詞，在特徵向量表中的頻率

n: 關鍵詞的總數

N: 虛擬文件的總數

N_j: 第j個關鍵詞，所出現的虛擬文件總數



圖八、系統架構圖

3 個向量建立完成後，假設G1 根據參考答案是有基因轉錄物 (T->Y) 及基因蛋白質 (P->Y) 的實驗結果，因此V_1會被加入T_Y及P_Y向量中 (T_Y, T_N, P_Y, P_N起始為空向量)。由於G2 只有基因轉錄物(T->Y)的實驗結果而沒有及基因蛋白質(P->N)，因此V_2會被加入T_Y及P_N向量中。如圖五所示。

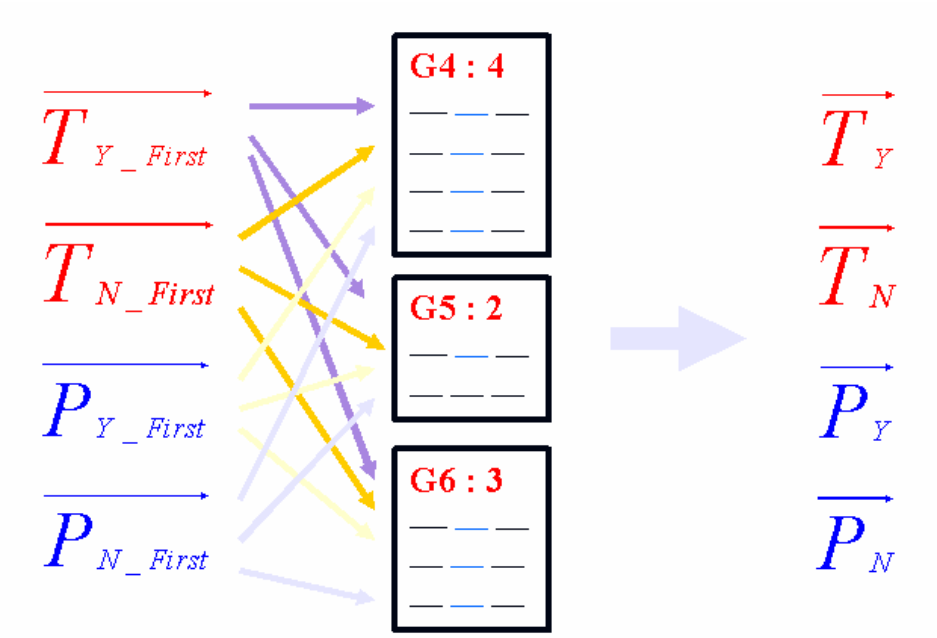
所有的虛擬文件均依這些步驟來進行。所有的虛擬文件處理完成後，T_Y及P_Y分別代表有基因轉錄物及基因蛋白質實驗結果的向量，T_N及P_N則代表沒有實驗結果的向量。

在測試階段，也是運用同樣的方法將待測文件中特定的基因(假設也有 3 個)，表示成 3 個向量V_4, V_5, V_6。唯一不同的是計算每個向量i中的每一個維度j的值$w_{i,j}$是以下公式計算得出 [10, 11]：

$$w_{i,j} = \frac{freq_{i,j}}{\max_{j=1,2,\ldots,n} freq_{i,j}}$$

然後再將V_4, V_5, V_6與T_Y及T_N分別計算其相似度(餘弦值(CosΘ)) [10, 11]，並將這兩個相似度相除。相除的結果如果大於一個門檻值，則預測為有基因轉錄物的實驗結果，否則預測為沒有。基因蛋白質的部分也是相同處理，如圖六所示 (其中V_g'代表V_4, V_5或V_6)。圖七為另一種比較的方式。它僅採用T_Y及P_Y有實驗結果或證據的向量。只要相似度大於其門檻值，則預測為有實驗結果或證據，否則預測為沒有。

在向量空間模型中，目標基因名稱所出現的句子，占整篇實驗文獻中實為少數。因此，擷取目標基因名稱所出現的句子，有可能因此失去了一些其它重要的特徵訊息。若能擴大其虛擬文件的大小，且自動過濾對分類結果沒有幫助的虛擬文件，則模型的準確度必然會大幅提昇。這也是我們要提出靴帶式向量空間模型分類法的原始動機。



圖九、捨去未通過門檻值的虛擬文件，並重新建立 4 個向量

四、靴帶式向量空間模型分類法

本系統所提出之靴帶式向量空間模型分類法之整體架構，如圖八所示。本系統將使用不同的資料來源：1、僅使用文件中的摘要 (Abstract)，2、使用文件的全文 (Full Paper)，以用來比較不同的資料來源，是否對於特徵的擷取、以及模型建立後的預測準確度有明顯的影響。本系統主要分為二大部份：1、資料的前處理 (Data Pre-processing)，2、分類模型的建立 (Model Building)，分別敘述如後。

4.1、資料前置處理

原始的實驗文件為純文字 (Plane Text) 格式。由於未經過加工處理，文件中充斥著妨礙建立模型的雜訊。因此在訓練階段，建立模型之前，必需先將文獻整理成所需要的型式，以利特徵的擷取以及模型建立的正確性。主要需處理的部分有三：1、同義字 (Synonym) 的取代，2、停用詞 (Stop Word) 的去除，3、字根 (Stem) 的還原。

在生物醫學領域中，由於部份基因名稱並未完全統一。在不同的文件中，相同的基因可能會有不同的引用名稱。為避免將相同基因所帶來的特徵視為不同，造成模型訓練上的不精確，故必需先將實驗文件中，相同意義但名稱不同的基因，統一以同一個名稱出現，再進行特徵擷取。我們在這個階段所使用的同義詞詞典為 KDD Challenge Cup Task 1 所提供的同義詞詞典。

此外，在英文的使用中，會大量的出現像 a、in、the 等停用詞 (Stop Words)。然而停用詞對於關鍵詞的擷取及辨識上，如果不將其排除，不但無法正面的提昇辨識率，甚至會變成辨識上的錯誤引導。故在進行文件關鍵詞的擷取前，必需先過濾停用詞，以提高模型建立後的正確率。本實驗所使用的停用詞詞典為 Stop Word Lists of Fox [12]。

同時，在英文的語法結構中，由於名詞的複數、動詞的時態、詞性的變化等等，相同的詞會有不同的呈現方式。然而其所代表的意義大致是相同的。如果將其視為不同的關鍵詞，則相對稀釋了其重要性。故在英文的關鍵詞擷取之前，必需先經過字根還原 (Stemming) 的處理。本實驗以 Porter Stemming Algorithm [13] 做為字根還原的依據。

4.2、模型的建立

我們所提出的方法，主要也是由訓練階段及測試階段所組成。在訓練階段，由於基因產物的實驗結果或證據未必會單純的存在於目標基因所出現在的句子裏，因此若考慮較小的範圍做為特徵擷取的背景(Context)，有可能因此失去了一些其它重要的特徵訊息。但是若考慮較大的範圍的話，雖然較可以避免失去了一些重要訊息，但是相對的也會因此而包含了較多不必要的雜訊，形成影響模型建

立其準確度的干擾。

由於在目標基因名稱所出現在的句子，其前後句包含有基因產物的實驗結果或證據的比例也相當高，因此本實驗選擇目標基因g_x所出現的句子再加上其前後各一個句子，做為建立其虛擬文件的背景。接下來我們也是依照向量空間模型的方法，建立了T_{Y_First}, P_{Y_First}, T_{N_First}及P_{N_First} 4 個向量。T_{Y_First}及P_{Y_First}分別代表有基因轉錄物 (mRNA, Transcript) 及基因蛋白質 (Polypeptide) 實驗結果或證據的向量，T_{N_First}及P_{N_First}則代表沒有實驗結果的向量。

然而如同前述，有些虛擬文件本身並不帶有實驗結果或證據的訊息。因此，本模型採用靴帶式 (Bootstrapping) 的方式，逐步刪除無預測能力的虛擬文件。其詳細步驟如下所示：

1. 我們逐一重新檢驗每一個虛擬文件。若虛擬文件能利用先前已建成之T_{Y_First}, P_{Y_First}, T_{N_First}及P_{N_First} 4 個向量，正確的分類，則保留之，否則捨去，如圖九所示。以此法過濾掉與整體特徵訓練向量差異性特別大的虛擬文件，避免不帶有實驗結果或證據的虛擬文件，干擾整體特徵訓練向量模型的建立。

2. 利用保留下來的虛擬文件，重新建構T_Y, P_Y, T_N及P_N 4 個向量。

3. 重複執行步驟 1 和 2 直到所有的虛擬文件均能分類正確。

五、實驗結果

本研究的實驗資料來源為 KDD Challenge Cup 2002 Task 1，ACM SIGKDD 2002 International Conference on Knowledge Discovery and Data Mining (KDD 2002) 所舉辦比賽的競賽資料。總共有 861 篇有關果蠅基因或分子生物學的文件以及每一篇文件的想查詢的基因名稱名單及其正確解答。我們將這些文件分成訓練和測試資料。訓練和測試資料之篇數比是 3:1，所以訓練資料有 647 篇，測試資料有 215 篇。

在實驗結果的評估上，大會是以精確率 (Precision Rate)、召回率 (Recall Rate)、與 F-Measure 來評估一個系統的好壞。精確率與回覆率在數學上的定義主要就是分母的不同。精確率代表所有預測有基因實驗證據的基因中，真正有基因實驗證據的比率；召回率則代表所有有基因實驗證據的基因，有被預測出來的比率。如下所示：

Precision Rate = 預測有基因實驗證據且真正有基因實驗證據的數量 / 預測有基因實驗證據的數量

Recall Rate = 預測有基因實驗證據且真正有基因實驗證據的數量 /有基因實驗證據的數量

一般來說 Recall 與 Precision 越高，代表系統的效能越好。F-Measure 則是結合了以上二種評估指標，當 Precision 與 Recall 的值皆高時，F-Measure 之值亦會很高。其定義如下所示 [10, 11]：

F-Measure = 2 * Precision Rate * Recall Rate / (Precision + Recall)

實驗的結果如表一所示。其中Abs.代表資料來源為文件中的摘要 (Abstract)，Full代表資料來源為文件的全文 (Full Paper)。Transcript Y代表只使用T_Y；Protein Y代表只使用P_Y；Transcript YN代表使用了T_Y及T_N；Protein YN代表使用了P_Y及P_N。在表一中我們比較向量空間模型(VSM) 與我們所提出的模型 (BVSM) 的差異。

由表一可以看出，靴帶式向量空間模型所得的 F-Measure 值均高(優)於向量空間模型。其中，我們所提出的方法 F-Measure 最高為 0.502，而向量空間模型 F-Measure 最高為 0.358。與向量空間模型法做比較，由於靴帶式向量空間模型法是利用回饋的方式，它能純化基因產物的實驗結果或證據特徵向量，提高分類效率。在同時考慮有產物特徵與無產物特徵作為分類的依據時，亦能純化不包含基因產物的實驗結果或證據特徵向量，進而得到更佳的分類效率。圖十至圖二十一為在不同的相似度參數 (Similarity Threshold) 下的實驗結果。

表二為支持向量機 (Support Vector Machine, SVM) [14] 在此議題下的實驗結果。其中 S1 代表只選基因出現的句子當做訓練及測試資料；S3 則代表選擇基因出現的句子以及前後兩句當做訓練及測試資料。雖然加入前後兩句可擴大特徵擷取的範圍，然而也加入了太多的雜訊。因此，S3 的實驗結果普遍比 S1 差。由表二的結果顯示，靴帶式向量空間模型法的效能 (F-Measure 最高為 0.502) 亦比支持向量機 (F-Measure 最高為 0.494) 為佳。

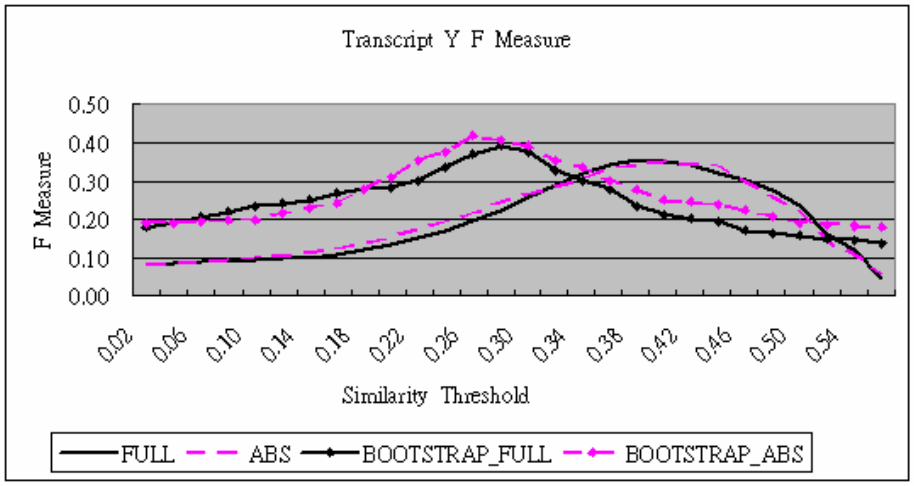
由於參賽隊伍第 3 名 [8] 的 F-Measure 為 0.59，我們的模型為 0.502。因此，我們所提出的方法已大幅超越不依賴生醫專業人員與知識的系統，並已接近前 3 名所提出的方法。

表一、Bootstrap VSM 與 VSM 效率比較表

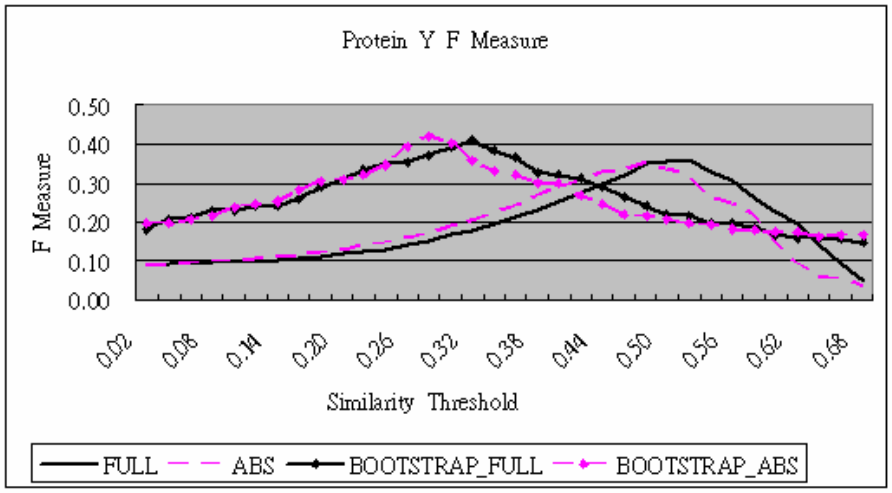
		F-Measure		Recall		Precision	
		BVSM	VSM	BVSM	VSM	BVSM	VSM
Transcript Y	Full	0.392	0.352	0.559	0.380	0.302	0.329
	Abs.	0.418	0.352	0.639	0.327	0.311	0.379
Protein Y	Full	0.410	0.358	0.517	0.416	0.340	0.318
	Abs.	0.423	0.356	0.681	0.479	0.307	0.286
Transcript YN	Full	0.477	0.291	0.742	0.393	0.351	0.231
	Abs.	0.486	0.315	0.512	0.391	0.463	0.264
Protein YN	Full	0.480	0.250	0.802	0.253	0.342	0.248
	Abs.	0.502	0.276	0.702	0.187	0.391	0.527

表二、支持向量機的實驗結果

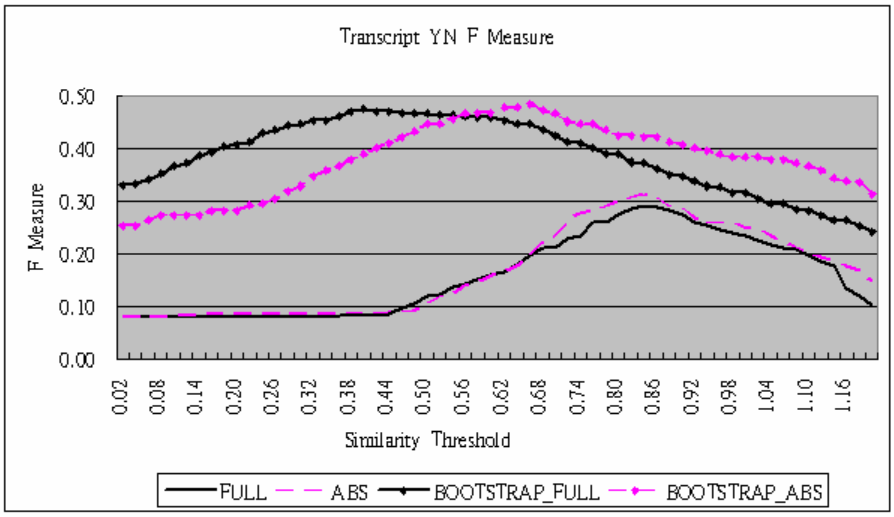
SVM		F-Measure		Recall		Precision	
		S1	S3	S1	S3	S1	S3
Transcript	Full	0.478	0.361	0.343	0.250	0.786	0.652
	Abs.	0.380	0.319	0.250	0.214	0.790	0.624
Protein	Full	0.494	0.411	0.364	0.289	0.769	0.714
	Abs.	0.333	0.294	0.255	0.193	0.482	0.617



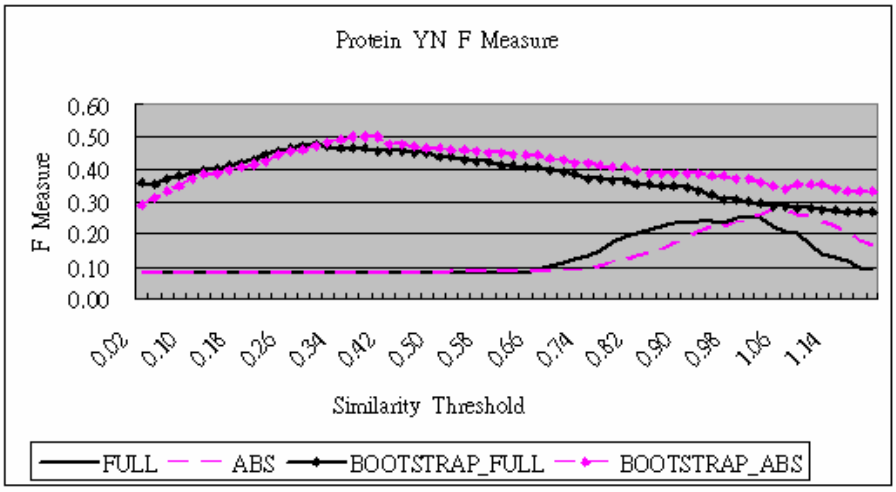
圖十、只使用T_Y的轉錄物的F-Measure分佈



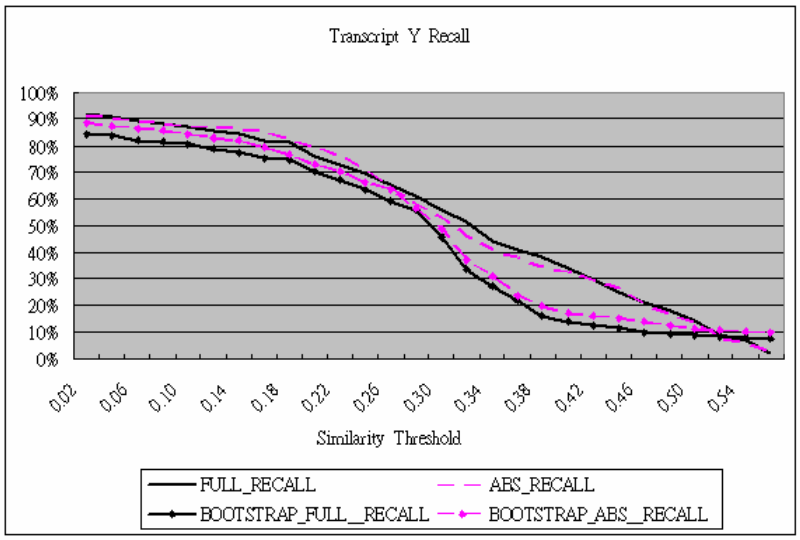
圖十一、只使用P_Y的蛋白質的F-Measure分佈



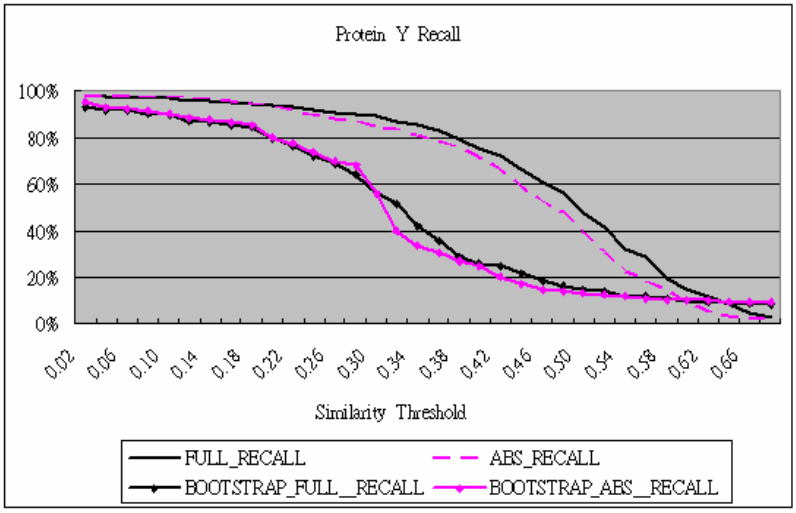
圖十二、同時考慮T_Y及T_N的轉錄物的F-Measure分佈



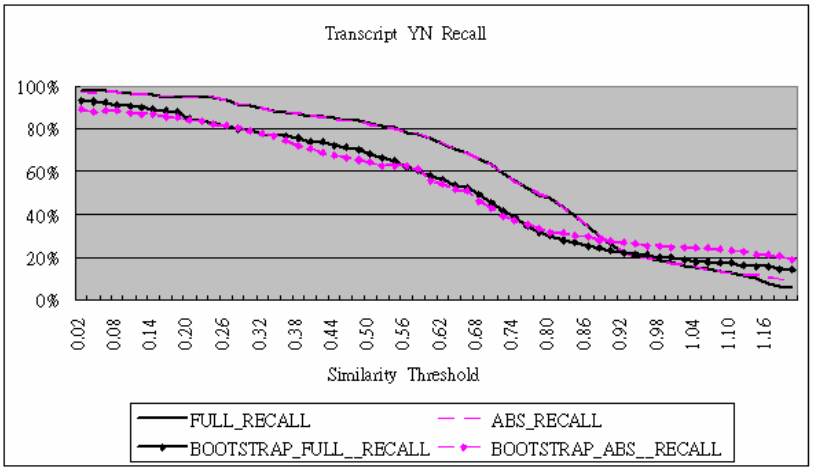
圖十三、同時考慮P_Y及P_N的蛋白質的F-Measure分佈



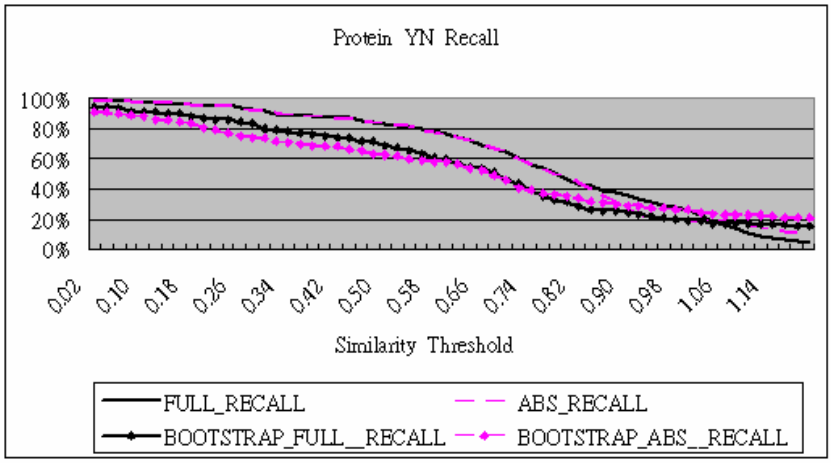
圖十四、只使用T_Y的轉錄物的Recall分佈



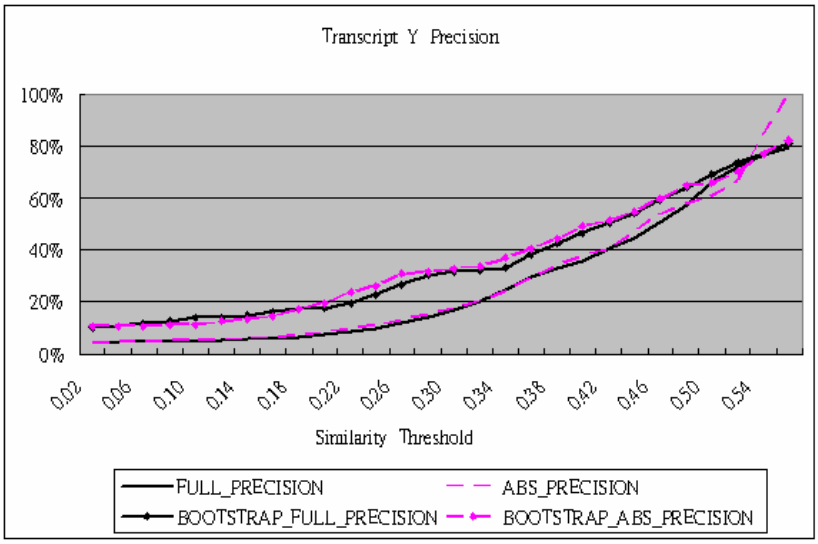
圖十五、只使用P_Y的蛋白質的Recall分佈



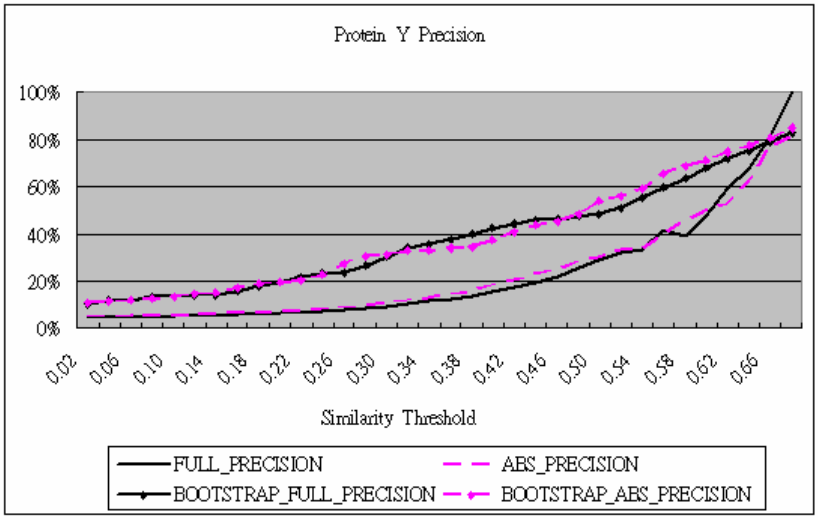
圖十六、同時考慮T_Y及T_N的轉錄物的Recall分佈



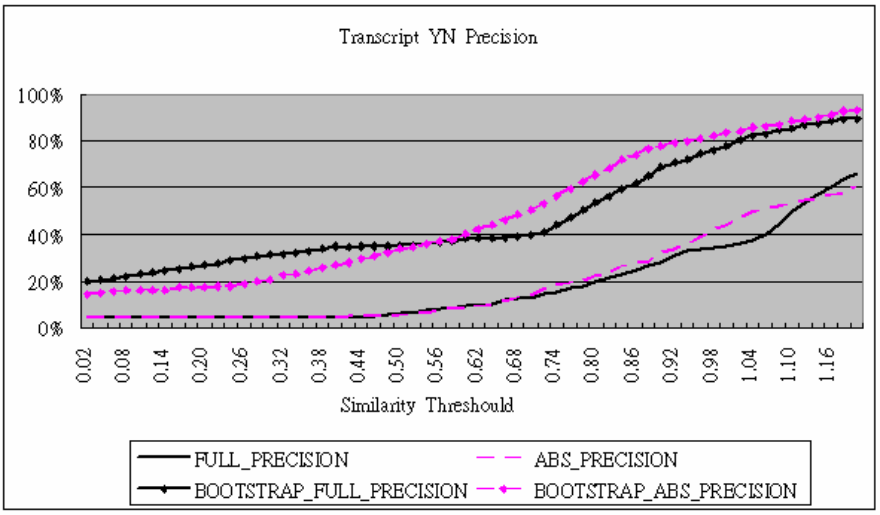
圖十七、同時考慮P_Y及P_N的蛋白質的Recall分佈



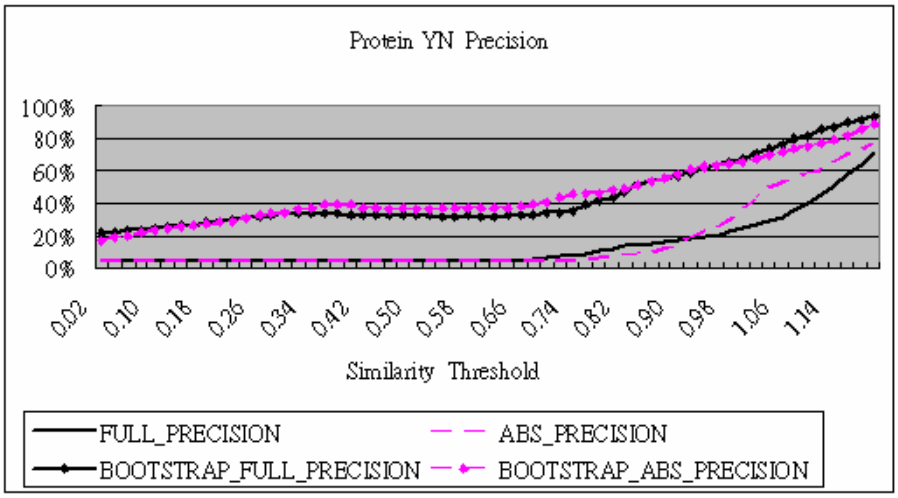
圖十八、只使用T_Y的轉錄物的Precision分佈



圖十九、只使用P_Y的蛋白質的Precision分佈



圖二十、同時考慮T_Y及T_N的轉錄物的Precision分佈



圖二十一、同時考慮P_Y及P_N的蛋白質的Precision分佈

六、結論

在本研究當中，我們建立一個不需依賴生物醫學領域專家及知識的電腦輔助系統，能夠自動判別在生物文獻之中是否擁有任何基因產物的實驗結果或證據，以利協助生物醫學資料庫的更新。實驗的結果顯示，我們所提出的方法已大幅超越不依賴生醫專業人員與知識的系統，並已接近前 3 名所提出的方法。在人力無法負荷的生物醫學文獻發表速度下，本論文所提出的模型能有效的節省更多的人力、時間及資源。

七、致謝

這篇論文是國科會計劃 (NSC 94-2622-E-130-001-CC3 & NSC 94-2213-E-130-004) 研究成果的一部份。我們在此感謝國科會經費支持這個計劃的研究。

八、參考文獻

[1] PubMed, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed.

[2] National Center for Bio-technology Information, http://www.ncbi.nlm.nih. giv/Entrez/. 1999.

[3] FlyBase, http://www.flybase.org/.

[4] A. Yeh, L. Hirschman, and A. Morgan, “Background and Overview for KDD Cup 2002 Task 1: Information Extraction from Biomedical Articles”, SIGKDD Explorations, Vol. 4, Issue 2, pp. 87-89, 2002.

[5] Knowledge Discovery and Data Mining (KDD) Challenge Cup 2002, http://www.biostat.wisc.edu/craven/kddcup/index.html.

[6] Y. Regev, M. Finkelstein-Landau, and R. Feldman. “Rule-Based Extraction of Experimental Evidence in the Biomedical Domain–KDD CUP 2002 (Task 1)”, SIGKDD Explorations, Vol. 4, Issue 2, pp. 90-92, 2002.

[7] S. S. Keerthi, et al. “A Machine Learning Approach for the Curation of Biomedical Literature–KDD CUP 2002 (Task 1)”, SIGKDD Explorations, Vol. 4, Issue 2, pp. 93-94, 2002.

[8] M. M. Ghanem, Y. Guo, H. Lodhi, and Y. Zhang. “Automatic Scientific Text Classification Using Local Patterns–KDD CUP 2002 (Task 1)”, SIGKDD Explorations, Vol. 4, Issue 2, pp. 95-96, 2002.

[9] C. Y. Su, Computer-Aided Construction of Curating Biomedical Databases: Extracting Evidences of Gene Expressions, Master Thesis, Department of Computer Science and Information Engineering, National Taiwan University, 2002.

[10] G. Salton. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley, 1989.

[11] B. Y. Ricardo and R. N. Berthier, Modern Information Retrieval, Addison-Wesley, 1999.

[12] C. Fox, Information Retrieval: Data Structures and Algorithms-Lexical Analysis and Stop-Lists, pp. 102-130, Prentice Hall, 1992.

[13] M. F. Porter, “An Algorithm for Suffix Stripping”, Program, Vol. 14, No. 3, pp. 130-137, 1980.

[14] B. Scholkopf and A. Smola, Learning With Kernels - Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, 2001.