考量樣板品質與自動擴張技術之醫學資訊擷取

何承威
成功大學資訊管理研究所
r76984015@mail.ncku.edu.tw

王惠嘉
成功大學資訊管理研究所副教授
hcwang@mail.ncku.edu.tw

摘要

人類基因表現主要控管於轉錄步驟，而轉錄調控是一個非常重要的起始步驟。它實質在於蛋白質與DNA、蛋白質與蛋白質之間的相互作用。透過轉錄因子(Transcription Factor, TF)與目標基因(Target Gene, TGene)間的調控，決定了基因的表現以及後續蛋白質的形成。目前很多相關研究都已記錄在生物醫學文獻中，以資料庫的方式儲存。但也因文獻成長的速度驚人，生物學家越不易從如此大量的文獻中獲取所需資訊，必須耗費大量的時間與人力來進行資料過濾。如何找出 TF 和 TGene 之間的調控關係是很重要的議題，因此我們利用資訊技術協助生物醫療研究人員從大量的文獻中找出有用資訊。

為了從生醫文獻中擷取出所需的資訊，有許多專家學者提出方法來改善成效不彰的文件搜尋方式，但這些方法仍有潛在的缺點，例如人工產生樣板(Manual Generate Pattern)，雖然資料的 Precision 較高，但 Recall 低；若統計生物醫學常見的關鍵字(Keyword)，則會有 Precision 低但 Recall 高的情況。目前的樣板學習方法很多，其中 bootstrapping 是一個可以自動學習的方法，藉由 bootstrapping 技術不斷擴張學習可以改善過去需要人工檢視的問題，但因學習到的新樣板並未經過評估，導致最後的回傳結果不一定能代表兩個實體間的交互作用或關係。

為了能讓 bootstrapping 技術自動產生的樣板比對到的句子更能代表兩個實體間的交互作用或關係，本研究在 bootstrapping 產生新樣板後增加了樣版品質的評估，目的是要確保樣板在多次運行的過程中，並不會因為擴張而失去其重要性。最後再透過樣板比對(Pattern Matching)，擷取出 TF 和 TGene 之間的調控關係。希望藉由本研究能讓生醫研究人員快速又準確地獲得所需資訊，並達到節省人力及時間成本的目的。

關鍵詞：基因調控、文字探勘、Bootstrapping、樣板排序

1 緒論

1.1 研究背景

隨著資訊快速發展，其研究成果數量與日俱增，相關的文獻記錄也逐日豐富，對於日後深究探討而言是十分重要而珍貴的基石。許多文獻目前收藏於資料庫中，並提供了資料庫搜尋引擎方便使用者快速的搜尋。在如此大量的文獻中要找出相關的文獻，資料的搜集和前處理顯得格外重要。因此要於多如星辰的文獻中找出有興趣實體(Entity)間的關聯，除了下適當的關鍵字外，還是要人工一一去確認回傳的文獻，取得文獻背後所隱藏的知識以達到知識發現(Knowledge Discovery)是資料探勘的最終目的(Feelders et al., 2000)。或許透過文獻摘要能很快的知道這篇文獻是否可能為所需要的，但回傳結果通常過多，以致於要閱讀並過濾很多可能相關但非真正所需的文獻摘要。故資訊人員利用資料探勘技術，從大量原始資料或資料庫中萃取出有用資訊，根據機器學習及統計等方式，例如文章中字詞的前後文關係或是共同出現的次數，便可以花較少時間從大量的文獻中擷取出所需的資訊。

為了達成此目的，各不同資訊領域的專家學者紛紛提出如人工標定樣本產生樣板(Khoo et al., 2000)、利用本體論概念計算

回傳文件的相關程度(Ahmad, 2010)、利用本體論(Ontology)連結跨領域關係(Snehasis et al., 2010)或是透過章節分類來歸納出該篇所討論的重點(Tsai et al., 2009)，進而增加搜尋的精確率。

在眾多的資料探勘技術中，利用樣板比對找出實體間的關聯是常見的資料關係擷取方法(Brin, 1998; Zeng et al., 2004; Li et al., 2008)。回傳結果的準確度決定於樣板的好壞，樣板又可透過人工標定(桂卓慶, 2008)或自動擷取(Yangarber et al., 2002; Etzioni et al., 2004; Surdeanu et al., 2006)而得。人工標定樣板的其回傳結果一般而言較好，但其缺點是要耗費大量人力來進行標定的動作，且擴充性差，無法取得人工標定以外的新樣板；自動擷取樣板雖然回傳結果有時非使用者所預期，但該方法擴充性佳，能耗費少量甚至不用人工標定就能擷取出新樣板。

1.2 研究動機與目的

雖然自動擷取樣板只要耗費少量甚至不用人工標定就能擷取出新樣板，但回傳結果的準確度決定於所自動擷取出的樣板是否能代表實體間的關聯，如何產生適合的樣板及評估樣板品質是自動擷取樣板所要深入探討的議題。

在自動擷取樣板上，學者們各自提出了一些方法來擷取樣板，找出實體間的關係(Riloff, 1996; Xia, 2006, Chang & Choi, 2006)。而在評估樣板品質方面，有學者利用樣板相似度(Agichtein & Gravano, 2000)或樣板排序 (Yangarber, 2000; Yangarber, 2003; Stevenson & Greenwood, 2005; Greenwood & Stevenson, 2006) 的方法來評估樣板品質。但利用樣板相似度的方法對於找出新樣板並無太大幫助，因為新樣板間相似度不一定高，但其樣板比對到的句子卻可能是使用者感興趣的；而樣板排序方法若採用以文件為基礎的假設，即假設相關的樣板在相關的文件中應該更常出現的話，樣本裡相關跟非相關文件的比例或數量會對其表現結果造成很大的影響。目前許多研究藉由資料探勘技術，找出 TF 與 TGene 之間的調控關係，取代了傳統以搜尋引擎搜尋相關文獻並經由人力閱讀、篩選、整合出關聯資訊，過程耗時費力的方式。透過不斷的改良技術，希望能在醫療資訊龐大的文獻中，幫助生醫研究人員花較少的時間發現了如基因與基因(Chiang et al, 2007)、基因與蛋白質(Yu & Agichtein, 2003)、蛋白質與蛋白質(Marcotte et al., 2001; Ono et al., 2001)、基因與藥物(Garten & Altman, 2009)、基因與疾病(Cano et al., 2009)之間的交互作用或關聯，這對生物醫療資訊領域來說，無疑是一項重大的突破。

本研究主要目的在於如何產生適合且高品質的樣板來找出兩個有興趣實體間的關聯。本篇以 TF 和 TGene 為例，當今生醫研究人員在針對一個 TF 尋找其與相關 TGene 之間的調控關係時，仍需要透過搜尋引擎從大量的文獻中來搜索與此 TF 相關的文獻，回傳的結果過多且很多是非相關文獻，顯然不是個有效率的方法。例如在下關鍵字尋找文獻時使用的是一個 TF 的名稱(如 HIF-1)，實際上卻找到一篇有提到該 TF，但主要並非探討該 TF 的文獻。因此為了希望更準確的找出 TF 和 TGene 之間的調控關係，我們利用資料探勘技術解決目前知識存放在不同文獻造成擷取不便的困擾(Lin et al., 2008)。

本研究希望建立一資訊系統，利用 bootstrapping 的方式自動產生樣板，並對 bootstrapping 所得到的樣板做排序，期待能讓適合的樣板能適當的反應其影響力。透過樣板分數和文件與樣板的相關程度分數來排序樣板，只有排序分數高於門檻值的樣板才能進入 bootstrapping 程序進行新實體和關係擷取。由於多了評估樣板的品質的步驟，被實體跟樣板比對到的句子也更接近使用者的需求，其包含該句子的文獻也較有可能是使用者所期望的，增加了準確性。因本研究以 TF 和 TGene 為例，盼能從文獻中找出與某 TF 和相關 TGene 之間的調控關係，提供生醫人員相關的文獻連結，幫助他們更快速準確的找到他們要的資訊。

透過本系統我們希望達成以下目的：

(1) 由本系統自動擷取出從醫療檢索系統中透過輸入的 TF 名稱所回傳的文獻中相關 TGene 之間的調控關係，幫助生醫人員減少花費在搜尋及過濾非相關文獻的時間及人力成本。

(2) 透過統一醫學語言系統(Human Genome Organization, HUGO)將 TF 或 TGene 的別名(如 ACBD6 的別名為 MGC2404)、簡寫(如 HIF-1 為 hypoxia-inducible factor 1 的簡寫)列入考慮，以找出更多相關資訊來 bootstrapping。

(3) 提供生物研究人員正確的轉錄因子與目標基因間的調控關係，進而降低研究的成本。最終希望經由更了解 TF 與 TGene 之間的調控關係後，能幫助人類基因的治療及改善。

2 文獻探討

本章節會先介紹與醫療資訊領域相關會用到的資源，接著介紹如何利用這些資源進行文字探勘(Text Mining)、自然語言處理(Natural Language Processing, NLP)、機器學習、Bootstrapping、樣板排序及先前研究，以了解目前資訊人員是如何利用這些技術來幫助生物醫療研究人員更快速準確的找出他們所需的資訊。

2.1 數位化醫療資訊相關資源

2.1.1 PubMed (文獻資源)

PubMed 為美國國家醫學圖書館的美國國家生技資訊中心(NCBI)所製作的生物醫學相關文獻的書目索引摘要搜尋引擎，並提供部分免費及付費全文連結服務。其核心主題為醫學，但亦包括其他與醫學相關的領域，像是護理學、生命科學、化學或其他健康學科。它同時也提供對於相關生物醫學資訊上相當全面的支援，像是生化學與細胞生物學。其資料庫收錄了自西元 1950 年至今從 Medline、生命科學期刊及在線圖書所收集的文獻資料，目前已超過二千萬筆，藏量相當豐富。

要使用 PubMed 搜尋引擎，使用者可透過 http://www.ncbi.nlm.nih.gov/pubmed 連至其官網，輸入欲查詢的關鍵字、作者、期刊名稱或 PMID 後，透過 Entrez 檢索系統，可檢索如下圖所示各資料庫，只要從其中一個資料庫出發，在檢索結果中即會進行相關資料庫的檢索，最後回傳與該關鍵字相關之生物醫學文件。

回傳的結果也可依使用者需求以不同的方式展現，主要方式有一般格式(Summary)、含摘要格式(Abstract)、XML 格式、MEDLINE 格式或只顯示其 PMID，點選進入某篇文獻後還能更進一步了解該文獻更詳細的資訊，如全文或相關引用等。

另外 PubMed 也提供 E-Utilities 的功能，讓開發者能自行開發程式去抓取 PubMed 裡的文獻資訊，因此本研究將利用此功能取得醫療相關文獻資訊，當成本研究之實驗樣本來源。

2.1.2 序列搜索系統

序列搜索系統(Sequence Retrieval System, SRS)是歐洲各國主要生物訊息中心必備的資料庫查詢系統。其目的是整合各個不同的資料庫，包括生物基因字典、核甘酸序列、蛋白質序列、蛋白質結構、轉錄因子及生物文獻，以提供使用者一個統一的搜尋入口。本研究利用其中的轉錄因子資料庫來識別文獻中的轉錄因子。

2.1.3 HUGO

人類基因組織(Human Genome Organization, HUGO)是為了調查基因的結構、功能及交互作用等所成立的組織，其建置的 HUGO 基因命名委員會(HUGO Gene Nomenclature Committee, HGNC)會記錄每個已知的人類基因的基因符號(Approved Symbol)、基因名稱(Approved Name)、基因別名(Aliases)、基因前符號(Previous Symbols)、基因前名字(Previous Symbols)等資訊，其中基因名稱跟符號都

是獨一無二。而基因別名是根據研究人員的不同，在記錄上可能會給予此基因不同的別名，以利其辨別，但也間接造成基因名稱混淆。
故本研究將利用此基因資料庫來識別文獻中出現基因的文字，以分辨別名及同義詞。

2.2 文字探勘

文字探勘(Text Mining)，也被稱為文本挖掘、文字採礦、智慧型文字分析、文字資料探勘或文字知識發現，通常是指從非結構化的文字中，透過資訊擷取、資料探勘、機械學習、統計學等技術萃取出有用的資訊的一門科學，其大部分的資訊(超過80%)都是以文字儲存(維基百科編者, 2010)。目前已有許多生物醫學相關領域的研究，透過文字探勘從許多文獻裡找出生物醫學中的關聯資訊，如蛋白質與蛋白質間的關聯(Marcotte et al., 2001; Ono et al., 2001)、蛋白質與基因間(Yu & Agichtein, 2003)、藥物跟基因(Garten & Altman, 2009)的關聯等。本研究將利用文字探勘技術從 PubMed 眾多生物醫學文獻摘要中找出 TF 與其相關 TGene 之間的調控關係。

2.2.1 自然語言處理

自然語言處理(Natural Language Process, NLP)旨在使電腦能夠「懂」人類語言所要表達的涵意，是屬於人工智慧及語言學領域的分支學科(維基百科, 2010)。目前自然語言處理的主要範疇如斷詞技術(Word Segmentation)、字根還原(Stemming)、資訊截取(Information Extraction, IE)、資訊檢索(Information Retrieval, IR)、詞性標註(Part-of-speech tagging)、句法分析(Parsing)等。

以下將介紹本篇主要會用到的三項技術，分別為斷詞技術、詞性標註和字根還原。

2.1.1.1 斷詞技術

斷詞技術是根據某一語言的句法規則，判斷文字分界點的位置，將每個單字切割出來，以供其他自然語言處理如詞性標註或字根還原等繼續使用。不同的語言有其適合的斷詞工具，本篇的文獻以英文為主，由於英文的詞與詞之間是以空白作為間隔，但有兩個例外狀況，第一個是標點符號通常是緊接在詞之後而無空白間隔；另一個是縮寫詞像"HIF"與"is"縮寫成"HIF's"，"do"與"not"縮寫成"don't"。所以斷詞的工作是將與標點符號相鄰的字以空隔分離，並將縮寫詞拆成兩個單詞。本篇使用的英文斷詞工具因其整合了詞性標註的功能，在下一小節再做說明。

2.2.2 詞性標註

詞性標註(Part-of-Speech Tagging, POS)是對一個句子當中的每一個單字進行詞性的判斷，並且加以標記。透過這些標記，電腦便可以分析出這個句子的架構及文法，由此便能更準確的解讀該句所要表達的涵意。

目前已有很多廣泛被使用的詞性標註工具，如 Genia-tagger、CLAWS、和 CRFTagger 等等，本研究選擇 Genia-tagger 來當作標註工具，因其擅長於標記生物醫學方面的文獻如 Medline 摘要(Shah et al., 2003)。我們以 PubMed 裡的其中一篇文獻中的一小段文字「DEC1 and DEC2 are directly inducible by HIF-1.」當例子，進行詞性標註，則會得到表 1 結果。

表 1 Genia-tagger tagging 範例句的結果

Word	Base	POS Tag	Chunk Tag	NE Tag
DEC1	DEC1	NN	B-NP	B-protein
and	and	CC	I-NP	O
DEC2	DEC2	NN	I-NP	B-protein
are	be	VBP	B-VP	O
directly	directly	RB	B-ADJP	O
inducible	inducible	JJ	I-ADJP	O
by	by	IN	B-PP	O
HIF-1	HIF-1	NN	B-NP	B-protein
.	.	.	O	O

2.2.3 字根還原

字根還原(Stemming)的目的在於將各單字中不同時態或是複數單字還原成原本最基本的型態。由於一個句子中的單字可能會帶有各式各樣的動詞時態或是複數等的情況，如「sign」、「signing」以及「signed」三者在意義上是相同的，但這些單字對未經字根還原處理前的電腦來說是三個不一樣的單字。若不事先做字根還原處理，將會增加電腦的資料處理量。因此透過字根還原，便可將上例三個單字轉化為「sign」的原型，視為同一個單字，以增進比對準確率。

2.3 機器學習

機器學習(Machine Learning, ML)是幫助資訊擷取系統利用經驗來自動發掘擷取的規則(Mitchell, 1997)。目前機器學習的方法根據要拿來學習時人類監督的量可被概分為三大類，分別為監督式機器學習法(Supervised Machine Learning)、非監督式機器學習法(Non-supervised Machine Learning)以及半監督式機器學習法(Semi-supervised Machine Learning)(Xia, 2006)，以下一一介紹。

2.3.1 監督式機器學習法

監督式機器學習法需要一組事先人工標定的訓練資料讓學習演算法能藉此建立其擷取規則。但因為人工標定需要耗費大量人力資源及時間，且標定的結果只能用在該特定領域，轉移到其他領域的成本過高。故雖然在一特定領域時其表現不錯，但因其人力標定和轉移成本太高，不適合用在幫助生物醫療人員從文獻中快速發現 TF 與其相關 TGene 之間的調控關係。

2.3.2 非監督式機器學習法

與監督式機器學習法剛好相反，它不需要事先標定好的訓練樣本就能利用更紮實的學習演算法來自動擷取樣板。Yangarber et al. (2002)提出樣板擷取演算法開發一關係擷取系統，並聲稱不需人為介入。但也因為缺少人工標定的訓練樣本，使其結果通常比監督式機器學習法來得差。

2.3.3 半監督式機器學習法

半監督式機器學習法剛好在監督式和非監督式學習法之間取得平衡。它利用少量的人工標定種子樣板(Seed Patterns)，搭配主動式學習來達到監督式學習的成果又不像監督式學習需要大量人力成本和時間。本研究因考慮到查準率(Precision)及查全率(Recall)的問題，故決定採用半監督式機器學習法進行實驗。

2.4 Bootstrapping

Bootstrapping 是一個可自我擴充學習的方法。先讓學習器(Learner)從為數不多的種子資訊(Seed Information)如種子詞(Seed Words)、種子樣板或種子組合(Seed Tuple)中學習，再利用已知的種子資訊找出更多的資訊，如此反覆的學習以達成自我擴充的目的。圖 1 即為一範例，利用 Seed Tuple 來進行 Bootstrapping 程序，一輪完成後得到新的樣板跟新的 $Tuple_1$ 和 $Tuple_2$。接著再拿新的 $Tuple_1$ 和 $Tuple_2$ 到下一輪去學習新的樣板跟 Tuple，如圖 2 所示。

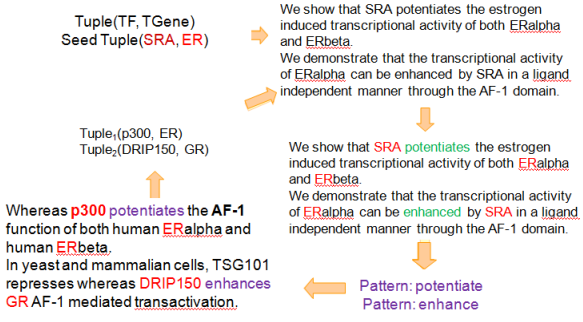

圖 1 Bootstrapping 第一輪

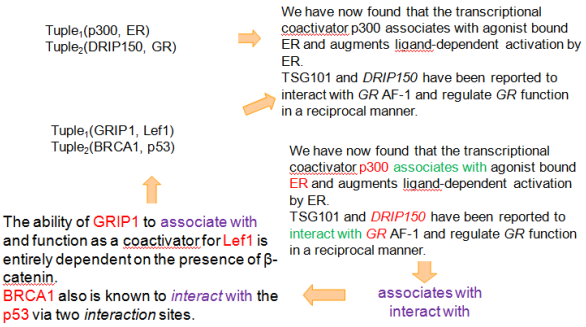

圖 2 Bootstrapping 第一輪

2.5 樣板排序

樣板排序(Pattern Ranking)的目的是要找出相關性的大小順序。在半監督式學習法中常會基於兩種假設來排序樣板(Liao & Grishman, 2010)： (1)以文件為基礎的(Document-Centric)方法，即假設相關的樣板在相關的文件中應該更常出現(Yangarber, 2000; Yangarber, 2003)。(2)以相似度為基礎的(similarity-centric)方法，即假設相關的樣板間在字詞上相似度應該較高(Stevenson & Greenwood, 2005; Greenwood & Stevenson, 2006)。

2.5.1 以文件為基礎

Riloff (1996)首先提出以文件為基礎的方法，宣稱只要樣本可以被分成相關跟非相關文件的話，樣板就能根據其在相關跟非相關文件內出現的次數來評估。(Yangarber et al., 2000)將 Riloff 的方法整合進 **b**ootstrapping 的程序中，以少量的種子樣板(seed patterns)開始，且不需要人工分類文件或是標定樣本。Surdeanu et al. (2006)更進一步先利用兩個分類器將文件根據需求來分類後再做 bootstrapping。但以文件為基礎的方法有個問題，樣本裡相關跟非相關文件的比例或數量會對其表現結果造成很大的影響。

2.5.2 以相似度為基礎

Stevenson & Greenwood (2005)提出一個排序候選樣板的方法，他們假設有用的樣板會跟已經被接受的樣板有相似的字詞。故利用 WordNet 來計算字詞的相似度來決定兩兩樣板間的相似度。但以相似度為基礎的方法也有個問題，若有一詞多義的情形，樣板相似度高就不一定代表兩個樣板是相似的，結果會導致 bootstrapping 效率下降。

因本研究只針對摘要做處理，其所對應到的樣板數已經不多，若採用以相似度為基礎的方法，即假設相關的樣板間在字詞上相似度應該較高，則對於找出新樣板並無太大幫助，因為新樣板間相似度不一定高，但其樣板比對到的句子卻可能都是使用者感興趣的。故本研究採用以文件為基礎的方法之概念，用以改善樣板評估的方式。

2.6 小結

綜合以上，本研究將以 PubMed 作為主要的資料來源，並且透過 UMLS 來辨識出文獻裡的相關實體，以協助標定作業的進行。在機器學習部分，本研究考慮到查準率(Precision)及查全率(Recall)的問題，希望利用少量初始樣板來學習新的實體跟關係，故採用半監督式機器學習法進行 bootstrapping，並利用 Wang et al. (2011)所採用之 Bootstrapping 的方法加入樣板排序，假設相關的樣板在相關的文件中應該更常出現，透過樣板分數和文件與樣板的相關程度分數來排序樣板，只有排序分數高於門檻值的樣板才能進入 Bootstrapping 程序進行新實體和關係擷取。由於多了評估樣板的品質的步驟，故回傳的句子也更接近使用者的需求，增加了準確性。

3 研究方法

本研究主要是導入半監督式學習法的 bootstrapping 技術並結合樣板排序，透過較可靠的樣板以更準確的找出 TF 與相關 TGene 之間的調控關係。前處理階段(Pre-processing Phase)使用者利用 TF 名稱作為種子詞利用 PubMed 的 E-Utilities 取得

與該 TF 名稱相關的文獻摘要集當成總樣本數，並開始進行斷詞、詞性標註和字根還原等前處理(Pre-processing)的動作。接著進入初始訓練階段(Initial Training Phase)，將有出現 Tuple(Disease, Drug-related)的句子訓練成正確樣板後，進入樣板排序階段(Pattern Ranking Phase)。首先給初始的樣板一個分數，利用此分數來計算文件與樣板的相關程度分數，最後透過樣板所對應到的文件與樣板的相關程度分數來排序樣板的分數，接著進入 Bootstrapping 階段(Bootstrapping Phase)。利用分數高於門檻值的樣板找出新的 Tuple(Disease, Drug-related)，新 Tuple 再從樣本中比對，將比對到的句子訓練成新樣板，新樣板再度回到樣板排序階段去排序。如此不斷循環直到達到終止條件。最後在樣板比對階段(Pattern Matching Phase)利用初始樣板和 bootstrapping 學習來的樣板透過樣板比對來擷取出 TF 和其相關 TGene 之間有調控關係的句子，呈現給使用者，並提供原文連結。

3.1 研究架構

因 Wang et al. (2011)所提出的系統架構是把學習到的所有新樣板都當成正確樣板下去找 Tuple，但其實並不是所有的樣板都能代表 TF 和相關 TGene 之間的調控關聯。當我們把這些樣板也訓練成正確樣板後，其所比對到的文獻就很有可能不是使用者所需要的。因此本研究將之改良後，其新的系統架構圖如圖 3 所示，希望透過文件排序和樣板排序的方式，對擷取出來的新樣板做進一步的限制，目的是為了讓回傳的結果更接近使用者的需求。本論文方法主要分成五個區塊，分別為前處理階段(Preprocessing Phase)、初始訓練階段(Initial Training Phase)、樣板排序階段(Pattern Ranking Phase)、Bootstrapping 階段(Bootstrapping Phase)和樣板比對階段(Pattern Matching Phase)。整個系統流程如下：

(1) 前處理階段：首先使用者輸入一 TF 名稱當關鍵字透過本系統連至 PubMed 搜尋符合的摘要內容，將之下載下來至本機端後，利用 TF 和 TGene 字典進行 TF 和 TGene 名稱之實體標定，只針對同時有 TF 和 TGene 的句子做處理，透過斷詞、詞性標註、字根還原、去除冗字(Stop-word)及考慮同義字等前處理後，將處理完的結果存進資料庫。

(2) 初始訓練階段：從上一階段 PubMed 搜尋到的文獻摘要中，使用者選定幾篇感興趣的或由系統隨機選取幾篇後，將其中有包含 Tuple(TF, TGene)的句子進行初始正確樣板訓練(Positive Pattern Training)，所得的正確樣板稱之為種子樣板。

(3) 樣板排序階段：從初始訓練階段所訓練而來的正確樣板會給予一個分數，藉由此分數可算出被該樣板比對到的文件與樣板的相關分數。再利用此文件與樣板的相關分數來排序樣板，故排序高的樣板所比對到的文件與樣板之相關程度應較接近使用者的需求，最後將排序前幾名的樣板加入到種子樣板中，進入 Bootstrapping 階段。而排序低的樣板就不會被丟進下一階段來尋找 Tuple，減少非相關文獻回傳給使用者的機率，進而達成回傳準確率提升。

(4) Bootstrapping 階段：利用樣板排序階段所得到排序較高的樣板到前處理階段所得來的文獻摘要中去找新的 Tuple(TF, TGene)，新 Tuple 再去比對句子後，去除符合錯誤樣板的句子，過濾掉不相關的句子後，再去除在初始訓練階段已經訓練成正確樣板的句子，減少產生重複樣板的機率，剩下的句子將訓練成新樣板，最後丟回樣板排序階段來評估樣板。樣板排序階段又會再產生排序分數高於門檻值的樣板回到本階段繼續尋找新樣板，如此不斷循環直到達到終止條件。

(5) 樣板比對階段：將前處理階段存進資料庫裡的句子，先利用錯誤樣板進行

過濾後，接著再把達到終止條件後所有從初始訓練階段和 Bootstrapping 階段訓練而成高於門檻值的樣板所比對到含有 Tuple(TF, TGene)的句子都抓出來，並將句子和原文連結呈現給使用者看。

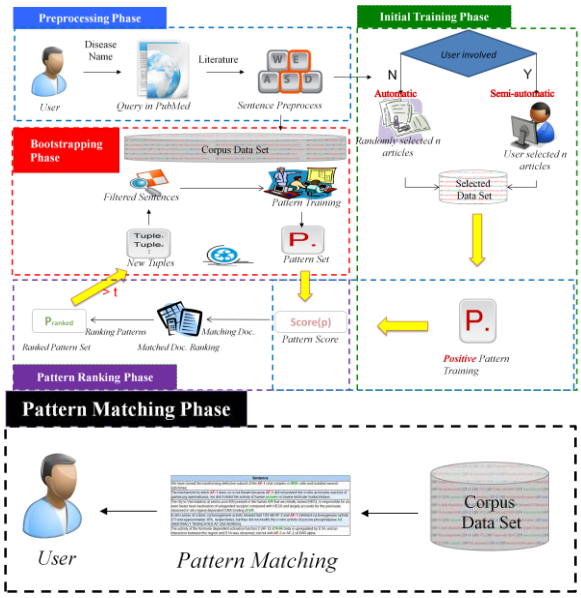

圖 3 系統架構圖

3.2 前處理階段

在前處理階段，首先由使用者輸入 TF 名稱做為關鍵字，系統自動從 PubMed 抓回相關文獻的資訊。接著對文獻裡的句子做前處理的動作，包含斷詞、詞性標註、字根還原及去除冗字等等。最後將處理的結果儲存至資料庫中，整個流程圖如圖 4 所示。

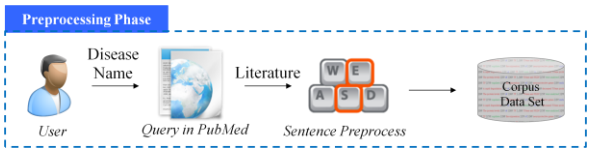

圖 4 前處理階段流程圖

3.2.1 PubMed 查詢

在前處理階段，首先由使用者輸入 TF 名稱做為關鍵字，由系統利用 NCBI 所提供的 E-Utilites 服務自動自 PubMed 資料庫抓取相關文獻摘要，其結果如圖 5 所示。

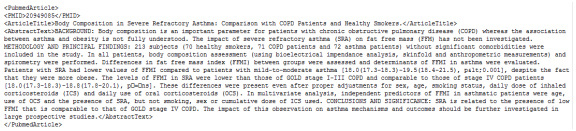

圖 5 利用 E-Utilites 服務取回 PubMed 摘要

本研究將所取回的文獻集合以 C(Corpus)表示，集合 C 裡的每份文獻 L_n(Literature)表示，$C=\{L_1,L_2,\cdots,L_n\}$。文獻 L 中的每個句子以 S_i 表示，$S_i \in L$。

3.2.2 句子的前處理

為了讓後續處理能更加順利，我們先將非結構化的文獻資料進行前處理的動作，其處理內容說明如下：

(1) TF 和 TGene 實體之標定：為了找出文獻中同時含有 TF 和 TGene 實體的句子，本研究利用 UMLS 來識別 TF 和 TGene 實體，其字典包含 TF 名稱、別名或同義字。若句子內同時出現 TF 和 TGene 名稱或是只出現 TF 名稱，則將之標記起來。若句子中只出現 TGene，甚至兩者都沒出現，則將該句丟棄，以下舉例說明：

會標記的句子：

PMID：12354771

(a)原句子：DEC1 and DEC2 are directly inducible by HIF-1.

(b)標記後的句子：TGene and TGene are directly inducible by TF.

丟棄的句子：

PMID：16964427

原句子：EGFR is involved in the UV signal transduction pathway leading to skin cancer.

此外，若 TF 或 TGene 有其他別名，則因作者習慣而可能有不同的寫法，例如 EGFR 的別名為" ERBB "，VEGF 的別名有”VEGFA”、”VEGF-A”，故在標記時這些名稱也要考慮進去。

(2) 斷詞技術與詞性標註：本研究使用

GeniaTagger 對句子做斷詞和詞性標註的動作，先將句子切成一個個單字後，對每個單字進行詞性標註。

(3) 字根還原：將字詞還原成字根，如「reduced」、「reducing」的字根都是「reduce」，則當句子出現「reduced」、「reducing」時都將之還原成字根「reduce」，以便後續進行比對。

(4) 去除冗字：本研究利用 Wang et al. (2011)所採用 Fox (1992)提供的 425 個刪除字，來去除冗字，如 the、a 等字。

文獻中有同時出現 TF 跟 TGene 實體的句子，透過上述前處理的步驟後，供往後各個階段使用。

3.3 初始訓練階段

將前處理階段所取得的文獻，先決定是否要使用者參與，若使用者對某幾篇文獻特別有興趣可將之挑選出來，否則由系統自動抓取幾篇當輸入資訊(Input)。接著將其中有包含 Tuple(TF, TGene)的句子進行初始正確樣板訓練(Positive Pattern Training)，所得的正確樣板稱之為種子樣板，流程如圖 6 所示。

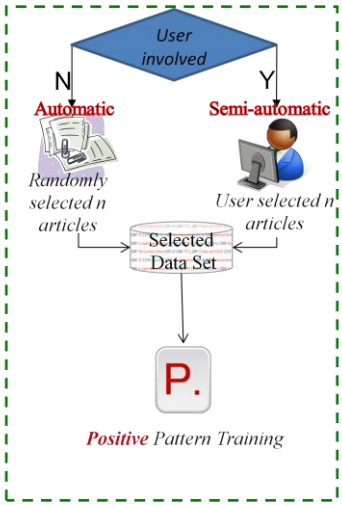

圖 6 初始訓練階段流程圖

3.3.1 正確樣板訓練

本研究採用 Wang et al. (2011)的方法將句子以向量(Vector)表示。其將句子結構分成五個部分，$Entity_1$ 和 $Entity_2$ 代表 TF 或 TGene 的實體，但兩者不能同為 TF 或 TGene。<Before>、<Middle>、<After>分別為出現在 $Entity_1$ 左邊的向量、$Entity_1$ 和 $Entity_2$ 中間的向量以及 $Entity_2$ 右邊的向量。

依桂卓慶(2008)所提出以片語為基礎之句子結構進行分析，片語分成名詞片語(Noun Phrase, NP)、介系詞片語(Prepositional Phrase, PP)及動詞片語(Verb Phrase, VP)，藉由界定片語的視窗大小會比單字的視窗大小有效率。故本研究以片語的視窗大小為考量，我們以一句子為例：
PMID：12354771
前處理後的句子：[SBAR Because] [NP DEC1 and DEC2] [VP are] [ADVP directly] [ADJP inducible] [PP by] [NP HIF-1], [NP these transcription factors] [VP may be] [ADJP crucial] [PP for] [NP the adaptation] [PP to] [NP hypoxia].

其中<Before>的視窗大小為 1，<Middle>的視窗大小為 4，<After>的視窗大小為 7。

利用上述結構化分析的結果開始進行正確樣板訓練。若將<Before>、<Middle>、<After>等三個向量進行組合，會得到<Before>、<Middle>、<After>、<Before+Middle>、<Before+After>、<Middle+After>及<Before+Middle+After>共 7 種不同的組合，產生不同的樣板。

初始正確樣板訓練階段主要是將醫療研究人員判定為正確句且同時含有 TF 和 TGene 實體的句子進行正確樣板訓練或是由系統自動抓取同時含有 TF 和 TGene 實體句子直接進行正確樣板訓練。每個樣板只要能成功比對一個以上的句子就能進入樣板排序階段進行排序，所產生的正確樣板定義為一個正確樣板集合(Positive Pattern Set, *PSet*)，$PSet=\{Pos_1,Pos_2,\ldots,Pos_n\}$，將每個樣板 Pos_i, $i\in[1,n]$儲存至資料庫。而該句所出現的 TF 和 TGene 實體也定義一個集合(Tuple Set, *TSet*)儲存之。

3.4 樣板排序階段

根據 Yangarber et al. (2000)所提出的以文件相關性為基礎的排序方法，其透過 Bootstrapping 樣板比對將未標註和未分類文件 U 分成相關 R 和非相關 $\overline{R}$ $=U-R$，並藉由樣板的 $Prec^{i}(p)$算出第 i 輪的文件的相關程度分數 $Rel^{i}(d)$，實驗結果證明此方法的確能提升精確率。故本研究利用文件的相關程度分數來排序樣板，目的在找出更適合進入 Bootstrapping 階段的樣板，進而改善回傳給使用者的結果，流程如圖 7 所示。

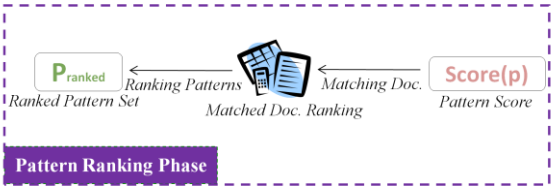

圖 7 樣板排序階段流程圖

3.4.1 樣板分數

計算樣板分數是為了得知該樣板與相關文件的關係程度，程度範圍介於 0 到 1。本研究將從初始正確樣板訓練階段所訓練出來的種子樣板 Pos_i 其樣板分數 $Prec^{0}(Pos_i)$ 設為 1，代表被初始正確樣板比對到的句子都先假設其為使用者想看的句子，其每輪的樣板分數透過 Yangarber et al. (2000)所提出的以下公式逐次更新：

$$Prec^{i+1}(p_n) = \frac{1}{|D(p_n)|} * \sum_{d_k \in D(p_n)} Rel^{i}(d_k)$$

其中 i 代表 Bootstrapping 階段第 i 輪；p_n 為目前所有樣板中的第 n 個樣板；$|D(p_n)|$ 為被該樣板所比對成功的文件個數，只要文件的句子中同時出現一 TF 實體及該樣板就算比對成功；d_k 為 p_n 所比對到的第 k 份文件；$Rel^{i}(d_k)$為被該樣板 p 所比對成功的文件 d 之相關程度分數，其計算方式在下一小節會詳細說明。

此公式的目的在於取得下一輪樣板重新計算後的分數。此新分數透過被該樣板 p 所比對成功的文件 d 之相關程度分數之總合，再除以比對成功的文件個數而得。因為若該樣板非有用或相關的樣板，卻因該樣板所比對到的文件數過多而造成樣板分數增加，其結果會造成各個文件與樣板的相關程度分數計算錯誤。而將比對成功的條件設為只要文件的句子中同時出現一 TF 實體及該樣板就算比對成功是因為若依照原本要句子中同時出現 TF 和相關 TGene 實體及樣板才算比對成功的話，摘要中所能獲得的樣板數量不多，故很難計算各個文件與樣板的相關程度分數。

3.4.2 文件與樣板之相關程度分數

本研究修改 Liao & Grishman (2010)的方法，利用樣板分數來計算文件與樣板的相關程度分數，其分數介於 0 到 1。文件 d 在第 i 輪與樣板的相關程度分數以 $Rel^{i}(d)$ 表示。被種子樣板所比對到的文件其與樣板的相關程度設為 1，而沒被比對到的文件其初始的文件與樣板的相關程度分數都是 0。其每輪的文件與樣板的相關程度分數透過以下公式逐次更新：

$$Rel^{i}(d) = \max\left(1 - \prod_{p_n \in P(d)} \left(1 - Prec^{i}(p_n)\right), \max\left(Prec^{i}(p_n)\right)\right)$$

其中 $Rel^{i}(d)$ = Bootstrapping 第 i 輪的文件 d 與樣板之相關程度分數；p_n 為目前所有樣板中的第 n 個樣板；$P(d)$為在文件 d 中出現的樣板集合。

此公式的目的在於藉由每一輪新的樣板分數，重新評估文件與樣板的相關程度分數。在 max 函數中左邊的情況是只要出現分數高的樣板數越多，其文件與樣板的相關程度越高，反之亦然；而因本研究是採用文獻摘要為樣本，摘要中本身含有的樣板數已不多，故右邊的情況是只要摘要中有出現一個分數高的樣板，就代表該份文件很有可能是使用者想要看的。

3.4.3 樣板排序

每一輪在算出文件相關程度分數後，利用各個文件相關程度分數對樣板做排序的動作。排序的目的是要決定哪些樣板較適合

進入 Bootstrapping 階段繼續後續處理。我們利用以下公式來排序樣板：

$$RankP(p_n) = \frac{\sum_{d_k \in D(p_n)} Rel(d_k)}{|D(p_n)|}$$

公式的目的是將樣板所比對到的所有文件相關分數累加後，除以比對到的文件個數，以避免有一分數低的樣板若比對到多份文件時，其累加後所得的樣板排序分數會增加的情形。最後會產生一排序後的樣板清單，去除掉已使用過的樣板後，挑選排序分數大於樣板分數門檻值 t 的新樣板進入 Bootstrapping 階段開始擴充樣板，t 由使用者自行決定。

3.5 Bootstrapping 階段

進入本階段後利用 *TSet* 裡的 $Tuple_i$ 到集合 C 的每篇文獻 L_i 中比對含有此 $Tuple_i$ 的句子 S_i，先利用桂卓慶(2008)所產生的錯誤樣板集合 *NSet*，$NSet=\{Neg_1, Neg_2, \ldots, Neg_n\}$ 來過濾句子，再利用已訓練成正確樣板的樣板 Pos_i 來過濾句子，剩下的句子將會進行正確樣板訓練，產生新的候選樣板，再回到樣板排序階段做排序。排序分數高於門檻值的樣板再到集合 C 進行比對，找出符合樣板的句子 S_j。對句子進行結構化分析後，取出句子中的 Tuple(TF, TGene)，將未存在 TSet 中的 $Tuple_n$ 新增至 *TSet* 裡，並到集合 C 的每篇文獻 L_i 中尋找含有此 $Tuple_n$ 的子子 S_n，如此不斷循環直達終止條件，其流程如圖 8 所示。

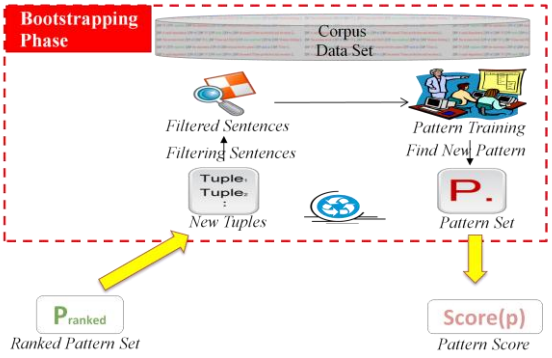

圖 8 Bootstrapping 階段流程圖

3.5.1 句子過濾

將 TSet 裡的 $Tuple_i$ 到集合 C 的每篇文獻中 L_i 比對含有此 $Tuple_i$ 的句子 S_i，先進行前處理和結構化分析後，利用錯誤樣板 Neg_i 比對 S_i，去除含有負面字或模糊字的句子，再利用已存在 *PSet* 中的樣板來過濾句子，其流程如圖 9 所示。剩下的句子進行正確樣板訓練之程序，產生新的候選樣板。

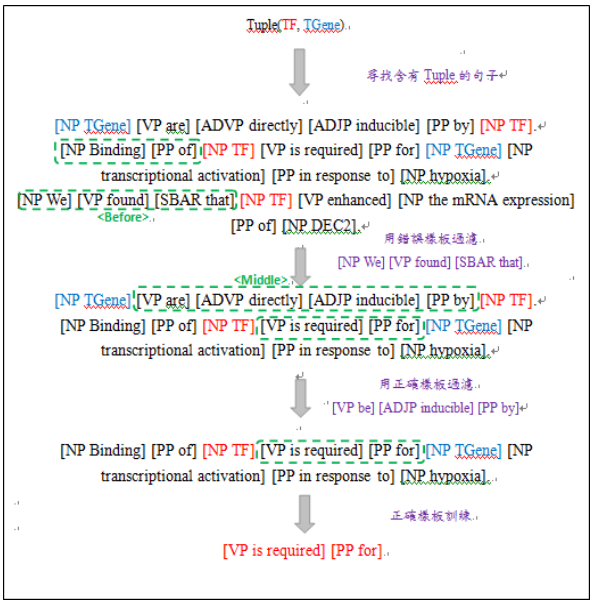

圖 9 句子過濾流程圖

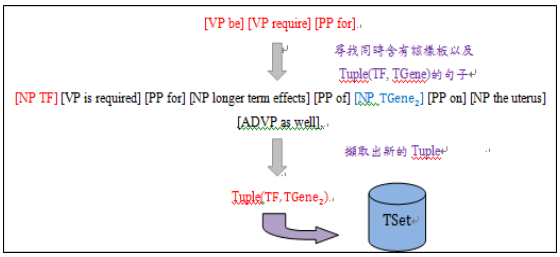

圖 10 尋找新的 Tuple 的流程圖

新的候選樣板加入種子樣板後，回到樣板排序階段重新排序樣板。接著利用樣板排序階段所挑選出來的樣板至集合 C 的每篇文獻中 L_i 比對，比對的方式分以下兩種：

(1) 將同時出現這些樣板以及 TF 和 TGene 實體的句子 S_i，進行結構化分析，擷取出句子中的 TF 和 TGene 實體，並檢查獲得的 Tuple 是否在 *TSet* 中，若

無則新增至 *TSet*，流程如圖 10 所示。

(2) 只要句子 S_j 裡有同時出現這些樣板及 TF 的實體，不一定要有 TGene 實體出現，進行結構化分析，利用樣板擷取出句子中的 TF 和另一實體，並將所獲得的 Tuple 新增至 $TSet_2$，其目的是要方便在樣板排序階段計算文件與樣板之相關程度分數用，並不利用這些 Tuple 去尋找新樣板，流程如圖 11 所示。

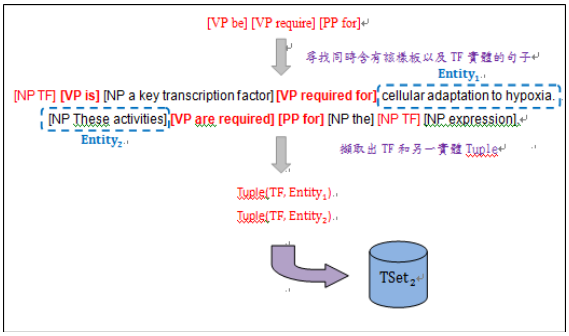

圖 11 記錄 $TSet_2$ 之流程圖

3.5.2 Bootstrapping

上一小節所找到的新的 Tuple 又會再次進行句子過濾，重覆執行 Bootstrapping 階段的各個程序，直達終止條件為止。根據 Li et al. (2008)指出 Bootstrapping 的終止條件可分為兩種：

(1) 給定一個最大的反覆運行次數(Maximum Iteration Times)，達到最大反覆運行次數時，就終止訓練。

(2) 當沒有新樣板或新 Tuple 產生時，就終止訓練。

由於本研究在 Bootstrapping 階段多了樣板排序的方法，挑選排序分數高於樣板分數門檻值 *t* 的新樣板進入 Bootstrapping 階段開始擴充樣板，若已無高於 t 的新樣板，則代表也無新 Tuple 產生，即和 Li et al. (2008)所指出的第二種終止條件相同。

3.6 樣板比對階段

本階段即將存入資料庫中的正確樣板及錯誤樣板對句子進行樣板比對以找出 TF 和相關 TGene 的調控資訊，流程圖如圖 12 所示，每個步驟說明如下：

(1) 將句子經過前處理及結構化分析後，產生 S_{11},S_{21},S_{33},...,S_{ij} 等句子，接著先進行錯誤樣板比對。

(2) 若句子的<Before>向量被錯誤樣板 Neg_i 比對到，則將該句丟棄。

(3) 剩下的句子的<Before+Middle>或<Middle>向量若再被排序分數高於門檻值 *t* 的正確樣板比對到，則將該句擷取出來給使用者看。

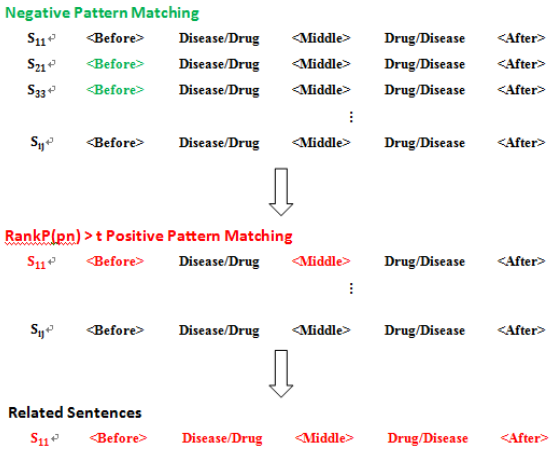

圖 12 樣板比對流程圖

4 系統建置

4.1 實驗環境

本系統的實作環境是架構在 Linux 系統上的作業平台，作業系統版本為 Ubuntu 8.10，資料庫管理系統為 MySQL 5.0.67，系統開發的程式語言為 Perl 5.8.7 及 PHP 5.2.4。

4.2 資料來源

本研究使用「HIF-1」做為 TF 關鍵字，至 NCBI 的 PubMed 上搜尋文章，並下載文獻做為資料集，共取得標題含摘要之文獻 602 篇及句子 5651 句後，請專家進行人工標定，得知其所含之正確句有 476 句，錯誤句有 5,179 句(包含未提及 TF 和 TGene 的句子)。接著將句子分成主動(TF-TGene)

與被動(TGene-TF)型態分別進行各階段的實驗，其中 TF-TGene 句子共 847 句，正確句有 316 句，錯誤句有 531 句；TGene-TF 句子共 731 句，正確句有 215 句，錯誤句有 516 句。

4.3 實驗結果與分析

4.3.1 實驗一：系統挑選之初始資料集

本實驗的目的是想得知初始資料集需要多少篇，才能比不評估樣板達到更好的成效。因此本研究以隨機的方式給定 100,200,…,600 篇的初始摘要篇數來觀察，並針對這 6 種數量的篇數做 10 次實驗後取平均，探討使用者需初始多少篇摘要，所設定之樣板分數的門檻值在多少時較為適當。

我們將 Precision、Recall 和 F-measure 以折線圖的方式呈現，分別如圖 12、13、14 所示。

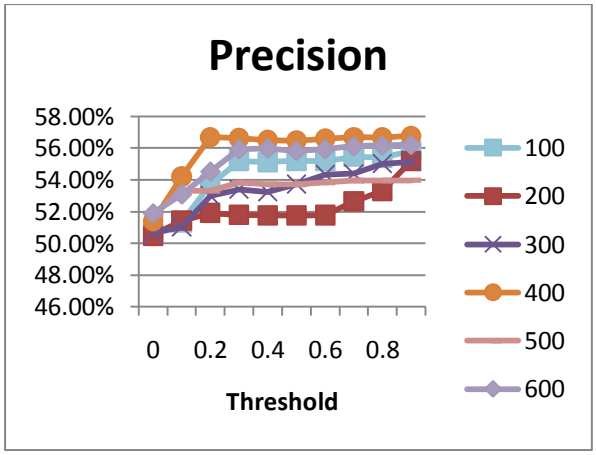

圖 12　初始篇數為 100,200,…,600 篇時各個樣板分數的門檻值所得到的 Precision

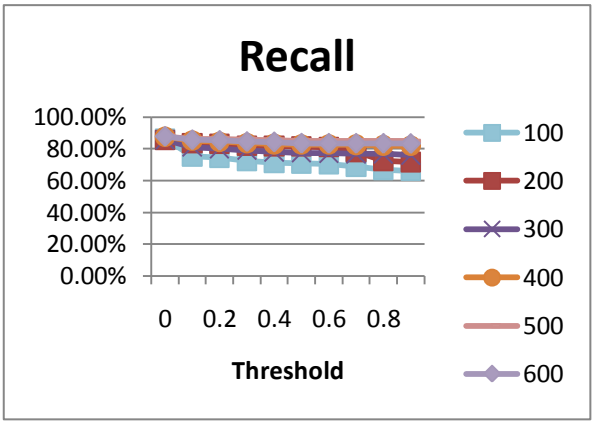

圖 13　初始篇數為 100,200,…,600 篇時各個樣板分數的門檻值所得到的 Recall

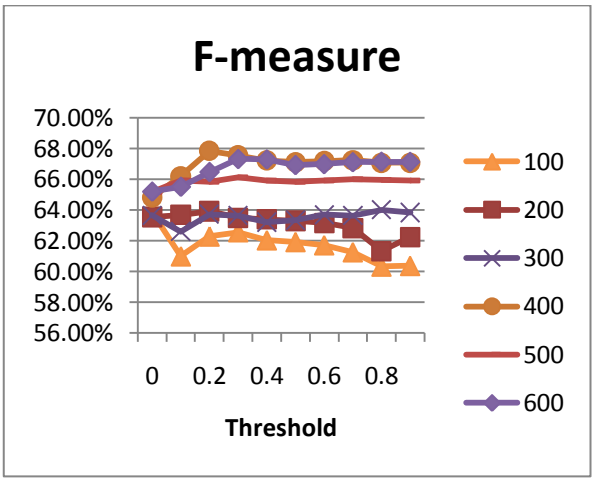

圖 14　初始篇數為 100,200,…,600 篇時各個樣板分數的門檻值所得到的 F-measure

由圖 12 我們可以看出，若不評估樣板品質時，其樣板分數門檻值為 0。當樣板分數門檻值逐漸提升時，其 precision 有逐漸上升的趨勢，而在 400 篇時，precision 的成效最好。因此我們固定篇數為 400 篇，將 Precision 在不同樣板分數門檻值時與不設門檻值（即門檻值為 0）所做的成對樣本 t 檢定之 p-value 整理成表 2。在 95%的信心水準下，不管樣板分數門檻值設多少，皆顯著優於不設門檻值。這代表多了樣板品質的評估，的確可以增加回傳結果的準確率。

由圖 13 我們可以明顯看出，當初始的篇數提高時，其整體 recall 值也漸漸提高；而當樣板分數的門檻值上升時，因所回傳的句子數也越少，故 recall 值有逐漸降低的

趨勢。

由圖 14 我們可以看出，圖中以 400 篇的成效最好，因此我們固定篇數為 400 篇，分別探討在樣板分數為 0,0.1…,0.9 時的效益，發現樣板分數門檻值設在 0.2 時最好。為了證明樣板分數門檻值設在 0.2 時優於其他樣板分數門檻值，我們將 F-measure 在不同樣板分數門檻值時所做的成對樣本 t 檢定之 p-value 值整理成表 3，在 95%的信心水準下，樣板分數門檻值設為 0.2 時皆顯著優於其他門檻值。

表 2 各個樣板分數門檻值與不設門檻值之 Precision 的成對樣本 t 檢定 p-value

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
p-value	0.0026	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**

表 3 在各個樣板分數門檻值時 F-measure 之成對樣本 t 檢定之 p-value 表

Threshold	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	-									
0.1	0.0150	-								
0.2	0.0000	0.0097	-							
0.3	0.0000	0.0162	0.0058	-						
0.4	0.0000	0.0598	0.0003	0.0116	-					
0.5	0.0000	0.0677	0.0001	0.0042	0.0927	-				
0.6	0.0000	0.0576	0.0002	0.0103	0.2715	0.0016	-			
0.7	0.0000	0.0542	0.0006	0.0317	0.4288	0.0060	0.0821	-		
0.8	0.0001	0.1078	0.0025	0.0223	0.1817	0.3678	0.2448	0.1234	-	
0.9	0.0001	0.1039	0.0041	0.0275	0.2345	0.4076	0.3000	0.1953	0.4477	-

4.3.2 實驗二：從給定篇數中只挑選人工標定之正確句當初始資料集

根據本研究的方法，我們將初始正確樣板的分數設為 1，因此初始樣板是否從正確句訓練而得可能會造成實驗結果有很大的不同。故本實驗的目的在於探討若初始資料集皆為人工標定之正確句，是否可以增加回傳結果的成效。由於需人工標定，希望初始越少篇即能達到較好的成效越好，因此我們先隨機給定 100,200,…,600 篇的摘要中只挑出正確句的部分當成初始資料集，並針對這 10 種數量的篇數做 10 次實驗後取平均。

我們將 Precision、Recall 和 F-measure 以折線圖的方式呈現，分別如圖 12、13、14 所示。

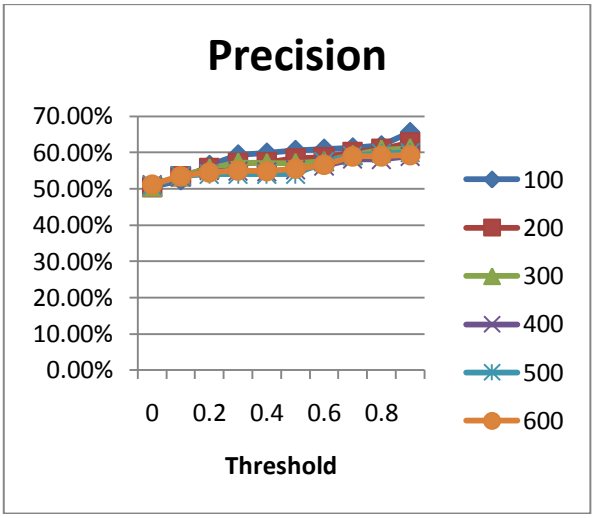

圖 15　初始隨機 100,200,…,600 篇中只挑選人工標定之正確句時各個樣板分數的門檻值所到的 Precision

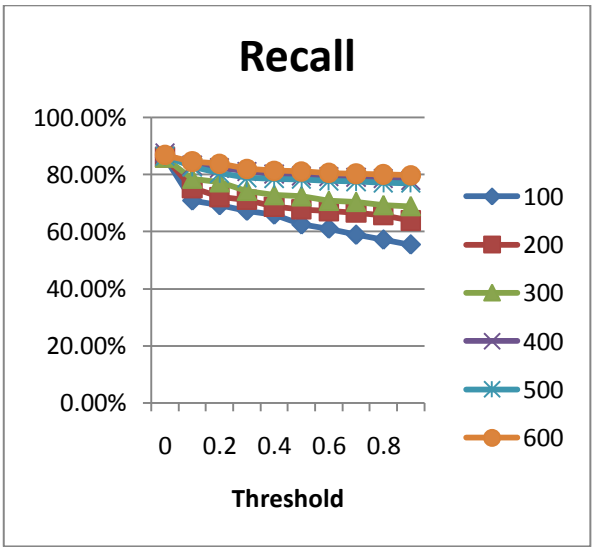

圖 16　初始隨機 100,200,…,600 篇中只挑選人工標定之正確句時各個樣板分數的門檻值所得到的 Recall

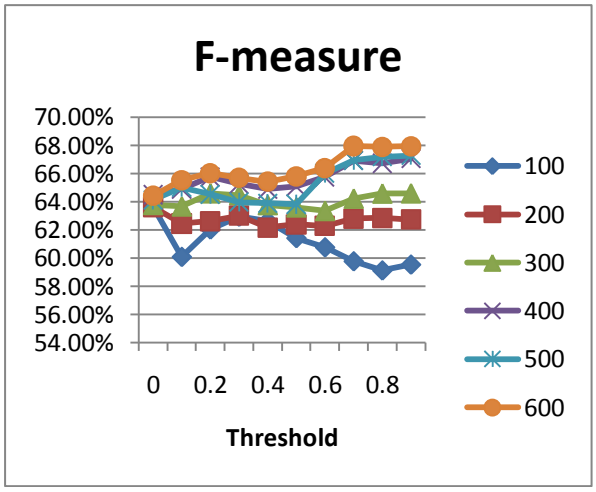

圖 17　初始隨機 100,200,…,600 篇中只挑選人工標定之正確句時各個樣板分數的門檻值所得到的 F-measure

由圖 15 我們可以看出，其 Precision 上升的幅度比隨機初始篇數還大。而從圖 16 我們也可以看出，初始資料集越多，其 recall 值也越高。但隨著樣板分數門檻值的上升，因所回傳的句子數也越少，故 recall 值有降低的趨勢。

由圖 17 我們可以看出，隨著初始資料集的篇數上升，其 F-measure 也跟著上升。因此初始正確句數的越多，其成效越好。

最後我們將其結果跟<實驗一>來做成對樣本 t 檢定比較，如表 4 所示。發現同樣初始 400 篇，只挑正確句的樣板分數門檻值設為 0.9 和隨機篇數時樣板分數門檻值設為 0.2，在 95%信心水準下，其只挑正確句的 Precision 優於<實驗一>隨機初始篇數的結果，但 F-measure 卻無顯著優於隨機初始篇數的結果。這代表雖然在初始資料集時增加正確句確實可以提升 Precision，但對於整體的表現而言並無明顯較好。

表 4 隨機初始 400 篇和隨機初始 400 篇中只取正確句之成對樣本 t 檢定結果

	Precision		F-measure	
隨機	-		-	
只挑正確句	0.0387	-	0.0623	-

5　結論

由於生物資訊的興盛，促使生物醫學領域的文獻大幅增加，接踵而至的問題是如何從龐大的生醫文獻中找出所需的資訊。雖然目前有許多文字探勘相關技術的應用，能讓生物學家快速的從大量的文獻中取得所需資訊，但所取得的資訊還是過多，且其中還是有很多非相關的資訊。

因此，本研究利用樣板排序的方式增加回傳的準確性，希望藉由 Bootstrapping 方法來自動產生樣板的同時，也不因為自動擴張樣板而降低樣板品質，使得回傳的結果非使用者所預期。其最終目的是期望能花最少的人力與時間成本，找出更多 TF 與 TGene 的調控資訊，以供生物學家進行分析與研究。

5.1 研究成果

本研究提出了一個樣板排序的方法，改善了 bootstrapping 因不斷擴張而使樣板品質降低的情形。藉著樣板品質的提升，能讓使用者過濾更少的文件就能得到所要的資訊。

以下我們根據每個實驗的結果進行分析與建議：

(1)實驗一：本實驗的目的是想得知初始資料集需要多少篇，才能比不評估樣板達到更好的成效。因此以隨機的方式給定 100,200,…,600 篇的初始摘要篇數來觀察，並針對這 6 種數量的篇數做 10 次實驗後取平均，探討使用者需初始多少篇摘要，所設定之樣板分數的門檻值在多少時較為適當。根據其實驗結果顯示，隨機初始資料集在超過 400 篇後皆比不評估樣板的 Precision 和 F-measure 來得好，而樣板分數門檻值設在 0.2 時整體成效最好。這說明了增加樣板品質評估的步驟的確可以讓使用者過濾較少的文件即能快速得到 TF 和 TGene 之間的調控關係，減少其人力及時間成本。

(2)實驗二：本實驗的目的在於探討若初始資料集皆為人工標定之正確句，是否比隨機初始資料集而不進行標定有較好的成效。根據其實驗結果顯示，初始資料集皆為人工標定之正確句只能讓 Precision 增加，而

對於 F-measure 並無顯著較好，而樣板分數門檻值設在 0.9 時整體成效最好。因此若使用者希望增加回傳結果的 Precision，則輸入初始資料集需為由使用者自行判斷為正確句的句子。而更進一步推究 F-measure 並無顯著較好的原因，可能是那些不是正確句的句子所訓練而得的樣板也有可能出現在正確句中，因此若在初始時將其拿掉，必會影響最後各個樣板的分數，造成最後部分應為正確樣板的樣板因分數不高而未將其判斷為正確樣板。

5.2 未來研究方向

本研究尚有許多值得探討與改善的地方，以下是針對後續研究方向提出幾點建議：

(1)初始資料集的挑選：本研究是將初始階段所得到的樣板分數皆設為 1 後，往後幾輪 bootstrapping 都是由這些初始樣板分數與新樣板分數不斷重複計算。。故初始資料集對本研究而言是非常重要的，建議由使用者挑選認為有興趣的文獻摘要當成初始資料集，其最後回傳的結果是使用者所預期的機率較高。

(2)對正確句、錯誤句或模糊句做更細的評分：本研究目前是將正確句標定為 1，錯誤句為 0，模糊句為-1。若不對句子做更細的評分，就無法得知回傳的結果與初始資料集的相關程度高低，也就無法確定回傳的結果就是使用者所期望看到的。

(3)補齊字典：在本研究的實驗中發現，有些沒被比對到的正確句有部分原因是因為該句並沒有出現完全符合TF字典裡的字，造成系統比對不到該句。

參考文獻

[1]桂卓慶，*利用文字探勘技術萃取轉錄因子與目標基因調控資訊*，國立成功大學資訊管理研究所，2008。

[2]Agichtein, E., & Gravano, L. , "Snowball:Extracting relations from large plain-text collectins," *In Proceedings of the 5th ACM International Conference on Digital Libraries*, San Antonio, Texas, United States, 2000.

[3]Ahmad, K., Eyas, E., & Zakariya, Q., "Ranking web sites using domain ontology concepts," *Information & Management*, Vol 47, Aug. 2010, pp. 350-355.

[4]Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O, "Open Information Extraction from the Web," *In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, Hyderabad, India, 2007.

[5]Brin, S., "Extracting Patterns and Relations from the World Wide Web," *In Proceedings of the International Workshop on The World Wide Web and Databases*, Valencia, Spain, 1998.

[6]Bui, Q. C., Nuallain, B. O., Boucher, C. A., & Sloot, P. M, "Extracting causal relations on HIV drug resistance from literature," *Bmc Bioinformatics*, Vol. 11, 2010.

[7]Cano, C., Monaghan, T., Blanco, A., Wall, D. P., & Peshkin, L., "Collaborative text-annotation resource for disease-centered relation extraction from biomedical text," *Journal of Biomedical Informatics*, Vol. 42, Feb. 2009, pp.967-977.

[8]Chen, E. S., Hripcsak, G., Xu, H., Markatou, M., & Friedman, C., "Automated acquisition of disease-drug knowledge from biomedical and clinical documents: An initial study," *Journal of the American Medical Informatics Association*, Vol. 15, Feb. 2008, pp. 87-98.

[9]Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, S., & Yates, A., "Web-Scale Information Extraction in KnowItAll(Preliminary Results)," *In Proceedings of the 13th international conference on World Wide Web New York*, USA, 2004.

[10] Feelders, A., Daniels, H., & Holsheimer, M, "Methodological and practical aspects of data mining," *Information &*

Management, Vol. 37, Sep. 2000, pp. 271-281.

[11] Fox, C., "Lexical analysis and stoplists: Prentice-Hall," 1992.

[12] Garten, Y., & Altman, R. B., "Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text," *BMC Bioinformatics*, Vol. 10, Feb. 2009.

[13] Greenwood, M., & Stevenson, M., "Improving semi-supervised acquisition of relation extraction patterns," *In Proceedings of the Workshop on information Extraction Beyond the Document*, Sydney, Australia, 2006.

[14] Khoo, C.,, Chan, S., Yun, N., "Extracting causal knowledge from a medical database using graphical patterns," *In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Hongkong, China, 2000.

[15] Li, W., Liu, T., & Li, S., "Bootstrapping for extracting relations from large corpora," *Journal f electronics*, Vol. 25, Jan. 2008, pp. 89-96.

[16] Liao, S. & Grishman, R., "Filtered Ranking for Bootstrapping in Event Extraction," *In Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, 2010.

[17] Lin, C., Tan, B., & Chang, S., "An exploratory model of knowledge flow barriers within healthcare organizations," *Information & Management*, Vol. 45, May 2008, pp. 331-339.

[18] Marcotte, E. M., Xenarios, L., & Eisenberg, D., "Mining literature for protein-protein interactions," *Bioinformatics*, Vol. 17, Nov. 2000, pp. 359-363.

[19] Mitchell, T., *Machine Learning: The McGraw-Hill*, 1997.

[20] Ono, T., Hishigaki, H., Tanigami, A., & Takagi, T., "Automated extraction of information on protein-protein interactions from the biological literature," *Bioinformatics*, Vol. 17, Sep. 2000, pp. 155-161.

[21] Qu, X. A., Gudivada, R. C., Jegga, A. G., Neumann, E. K., & Aronow, B. J., "Inferring novel disease indications for known drugs by semantically linking drug action and disease mechanism relationships," *BMC Bioinformatics*, Vol. 10, May 2009.

[22] Riloff, E., "Automatically Generating Extraction Patterns from Untagged Text," *In Proceedings of the Thirteenth National Conference on Artificial Intelligence*, Portland, Oregon, 1996.

[23] Shah, P. K., Perez-Iratxeta, C., Bork, P., & Andrade, M. A., "Information extraction from full text scientific articles: Where are the keywords?" *BMC Bioinformatics*, Vol 4, May 2003.

[24] Snehasis, M., Mathew, P., & Kalyan, M., "Multi-way association extration and visualization from biological text documents using hyper-graphs," *Artificial Intelligence in Medicine*, Vol. 49, Mar. 2010, pp. 145-154.

[25] Stevenson, M., & Greenwood, M., "A Semantic Approach to IE Pattern Induction," *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Ann Arbor, MI, 2005.

[26] Surdeanu, M., Turmo, J., & Ageno, A., "A Hybrid Approach for the Acquisition of Information Extraction Patterns," *In Proceedings of the EACL 2006 Workshop on Adaptive Text Extraction and Mining*, Trento, Italy, 2006.

[27] Tsai, R. T. H., Lai, P.-T., Dai, H. J., Huang, C. H., Bow, Y. Y., Chang, Y. C., Pan, W.H., & Hsu, W.L., "HypertenGene: extracting key hypertension genes from biomedical literature with position and automatically-generated template features," *BMC Bioinformatics*, Vol. 10, Dec. 2009.

[28] Wang, H.C., Chen, Y.S., Kao, H.Y., & Tsai, S.J., "Inference of transcriptional regulatory network by bootstrapping patterns," B*ioinformatics*, 2011.

[29] Xia, L., *Adaptive Relationship*

Extraction by Machine Learning, University of Sheffield, 2006.

[30] Yangarber, R., "Counter-Training in Discovery of Semantic Patterns," *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, 2003.

[31] Yangarber, R., Grishman, R., Tapanainen, P., & Huttunen, S., "Automatic Acquisition of Domain Knowledge for Information Extraction," *In Proceedings of the 18th International Conference on Computational Linguistics*, SaarbrXucken, Germany, 2000.

[32] Yangarber, R., Grishman, R., Tapanainen, P., & Huttunen, S., "Unsupervised discovery of scenario-level patterns for information extraction," *In Proceedings of Conference on Applied Natural Language Processing ANLP-NAACL*, Seattle,WA, 2002.

[33] Yu, H., & Agichtein, E., "Extracting synonymous gene and protein terms from biological literature," *Bioinformatics*, Vol. 19, Feb. 2003, pp. 340-349.

[34] Zeng, X., Li, F., Zhang, D., & Vakali, A., "An XML-Based Bootstrapping Method for Pattern Acquisition," *In Proceedings of the 6th International Conference on Enterprise Information Systems*, Porto, Portugal, 2004.

[35] Zerhouni, & Elias, A., "US biomedical research: Basic, translational, and clinical science," *The Journal of the American Medical Association*, Vol. 11, Sep. 2005, pp. 1352-1358.