慈濟大學 醫學院 醫學資訊學系碩士班

碩士論文

**Department of Medical Informatics**
**College of Medicine**
**Tzu Chi University**
**Master Thesis**

# 以特定生醫物件為研究目標之生醫文獻檢索

# Identification of Articles with Focuses on Specific Biomedical Entities


郭哲廷

**Zhe-Ting Guo**

指導教授：劉 瑞 瓏 博士

**Rey-Long Liu, Ph.D.**


中 華 民 國 106 年 7 月

**July, 2017**

# 致　謝

本研究承蒙科技部研究計畫補助(計畫編號：MOST 105-2221-E-320-004)，謹此誌謝。2017 年 7 月 13 日口試結束，得知通過的那一刻，代表著我要跟研究所說再見的時候，從大學專題到現在，一切都烙印在我的心中。在學習的過程中感謝有指導老師劉瑞瓏的指導與陪伴，無論是研究或是做人處事都獲益良多，老師常說「過程比結果要緊」，在過程中能感覺到自己並不是能靜下心把事情解決的人，所以我真的幸運能遇到這麼好的指導教授，而老師給我一句話要我一輩子記下來，就是「先做人，再做事」，世間上一切的事情往往都與這句話脫離不了，感謝老師願意花時間與精神，耐心陪伴我這位學生。

感謝兩位口試委員，李官陵老師與林紋正老師百忙之中抽空審閱我的論文，且給予 TEE 指正與建議。也感謝我的家人對我的信任與關懷，給我心靈上很大的照顧，並特地從台南到台東來看研討會，在我遇到困難時能聆聽我的煩惱並給我意見。

最後，感恩在慈濟教育體系中，感謝所有醫學資訊學系的所有師長與同學們以及所有讓我成長的一切。即將邁入人生的另一個階段，期許自己能將所學的回饋於社會。

# 摘要

現今生醫領域已有許多生醫資料庫被建立及維護，內含生醫文獻中已發表之疾病、基因與化合物間之關聯。然而生醫研究的蓬勃發展產出了巨量且異動頻繁的生醫文獻，使得生醫學家很難從這些文獻中完整及時地找出生醫物件間之關聯。故本文提出一套新的資訊科技 TEE（Target Entity Extractor），依照生醫物件出現在文獻中不同段落給予不同比重之分數，找出各生醫文獻中之目標生醫物件（Target Entity），進而將各文獻依照其與特定生醫物件之關聯度，進行排名並推薦予使用者。我們並以實際生醫資料庫來驗證 TEE 之效能。結果顯示 TEE 顯著優於現今較佳之文獻排名技術。本研究成果可以有效輔助生醫學家在巨量多變的生醫文獻中及時完整地找出疾病、基因與化合物間之關聯，促進相關研究之進展。

# 第一章 簡介

人類基因組計畫公開了人類基因體序列，開啟了後基因體時代。生醫學家嘗試探索基因對於人體健康與疾病的影響，以瞭解基因與疾病之關聯，進而發展診斷、預防及治療方式，幫助更多遺傳疾病患者。

這些探索及研究產生大量生醫文獻。PubMed[1]為現今常被生醫學家使用之文獻檢索系統，收集了巨量生醫文獻摘要及連結。為獲取最新的知識，生醫學家及文獻探勘系統常需要花不少時間在搜尋疾病、基因與化合物的關聯文獻。由於文獻成長速度太快，及時完整的審閱所有文獻有一定的困難，故如何從大量文獻中找到與需求高相關之生醫文件是的一個重要挑戰。

## 1.1 問題定義

本研究提出一個名為 TEE (Target Entity Extractor)之技術，針對給定的一份生醫文獻，自動擷取出其主要討論之生醫物件(Target Entity)。藉由 TEE 之輔助，生醫學家可以找到與一組特定生醫物件(ex：特定疾病、基因及化合物)高度相關之文獻，促進生醫研究之進展。

## 1.2 研究動機及主要貢獻

為獲取最新知識，生醫學家、遺傳疾病與基因編撰人員(例如：Genetics Home Reference, GHR[2]、Online Mendelian Inheritance in Man, OMIM[3]的編撰人員)、疾病基因文獻探勘系統等，每日需花不少時間在搜尋生醫物件之間的關聯，且每一筆關聯都需要多篇文獻的驗證才能確立。但文獻量龐大，且更新速度快，導致生

[1] PubMed: https://www.ncbi.nlm.nih.gov/pubmed/
[2] GHR: https://ghr.nlm.nih.gov/gene/GHR
[3] OMIM: https://www.omim.org/

醫學家無法及時掌握，且須花費許多人力與時間進行基因、疾病與化合物的更新與維護。編撰人員因此常來不及審閱文獻，不能及時完整地收集這些已於文獻中發表之關聯。TEE 之研發可輔助此類高相關文獻之收集。

然而，一篇生醫文獻提及特定生醫物件並不代表該生醫物件是該文獻之主要研究目標物件，故 TEE 依生醫物件出現在文章中之位置來減少誤判的情形。舉例而言，針對高血壓(hypertension)及鎂(magnesium)之配對來說，圖 1 為 CTD[4](Comparative Toxicogenomics Database)所選目標文獻，其內容為高血壓(Hypertension)與鎂(magnesium)的關係；圖 2 為非 CTD 所選文獻，內容主要為肌肉病變的婦女之研究，但在文章中也提及高血壓及鎂。兩者相比較，圖 1 實際談論到查詢配對之關聯，而圖 2 雖然有提到查詢目標，但文章內容並非主要討論該關聯。CTD 所選文章(圖 1)之配對均出現在標題，且出現在接近結論的段落。利用此位置資訊吾人應可更成功辨識出圖 1 文獻比圖 2 文獻更與目標配對<高血壓, 鎂>相關。

基於以上之分析，TEE 針對文獻與摘要進行分數計算，包含了以下二個層面：(1)研究目標(Goal)：當字詞出現於標題時，表示文獻可能以該字詞為研究目標；(2)研究結論(Conclusion)：在最後段落裡出現字詞，可能與該文獻中的總結有高度相關。

我們並設計一系列之實驗來證實 TEE 之效能。我們實際使用 CTD 中有人工審閱之<疾病, 基因>、<化合物, 基因>、<化合物, 疾病>關聯( Singhal 2001)。以 CTD 中各關聯配對之參考文獻作為目標文獻(高相關文獻)，非該關聯配對之文獻為非高相關文獻。依 TEE 計算分數，與其他關聯性排序之對照組進行實驗比較，進而驗證 TEE 之實際成效。

[4] CTD: http://ctdbase.org/

Title: Tissue and blood levels of zinc, copper, and **magnesium** in nitric oxide synthase blockade-induced **hypertension**.

Abstract: The aim of this study was to determine the levels of tissue and blood zinc (Zn), copper (Cu), **magnesium** (Mg) in nitric oxide (NO) blockade-induced . …. There were no differences between the…… and renal vein (22.1%). There were no significant Mg concentration changes in the hypertensive group's plasma, cerebellum, liver, duodenum, or aortal tissue. These findings suggest that **magnesium** depletion may play a role in the blood pressure rise that occurs in the model of chronic NO synthase inhibition-induced **hypertension**.

(節錄來源：https://www.ncbi.nlm.nih.gov/pubmed/?term= 11101042)

□：目標化合物； ▒：目標疾病； ＿＿：其他生醫物件

圖 1 CTD 所選之參考文獻：< magnesium, hypertension >

Title: Opioid-related narcosis in a woman with myopathy receiving **magnesium**.

Abstract: An Asian multiparous woman weighing 47 kg, who suffered from a rare myopathy, congenital fibre type disproportion, was given morphine 10 mg intramuscularly for labour analgesia. After delivery, she had diastolic **hypertension** and proteinuria and was prescribed **magnesium** sulphate. S….. Blood gas analysis revealed a respiratory acidosis. Naloxone administration reversed the symptoms. Further doses were required as the respiratory depression recurred. Opioid-related narcosis is the most likely diagnosis in this case. Other possible differential diagnoses were **magnesium** overdose or a post-ictal state. The presence of a myopathy could render this patient susceptible to the respiratory effects of opioids. Other explanations for an exaggerated and delayed response to opioids include co-administration of other respiratory depressant drugs such as **magnesium** sulphate, co-morbidity such as renal impairment and genetic variability in the metabolism of morphine. …..

(節錄來源：https://www.ncbi.nlm.nih.gov/pubmed/?term= 17643282)

□：目標化合物； ▒：目標疾病； ＿＿：其他生醫物件

圖 2 非 CTD 所選之參考文獻：< magnesium, hypertension >

# 第二章 文獻探討

## 2.1 自生醫文獻中擷取資訊

以往已有擷取生醫文獻中資訊之技術，例如：尋找疾病與基因之關聯(Özgür et al. 2008)、疾病和基因與藥物之連結(Domedel-Puig and Wernisch 2005)、搜尋基因之證據句子(Kim et al. 2013)、尋找疾病可能之候選基因(Patnala et al. 2013)、擷取蛋白質或生物分子之間的交互作用(Ahmed et al. 2005; Baral et al. 2007)等。然而，這些技術之性能常常是有限的，有相當高比例之資訊無法被正確完整地擷取出來，尤其當文獻中的語法、語義、句子結構超出了系統的語言處理技術時，則準確性可能進一步下降。本研究則改以找出文獻中目標物件(target entity)為目標。研究成果可以成為上述研究之前處理器，用以提升生醫資訊擷取之效能。

## 2.2 布林模型

布林模型(Boolean Model)為早期之資訊檢索模型( Singhal 2001)。利用布林運算 AND、OR、NOT 設定查詢的條件，檢索出符合相關之文章。雖然簡單且直覺，但布林模型只分符合與不符合兩類，並無針對文章相關性排序。因此，許多檢索引擎會先以布林模型為基礎，對文章做基本的過濾，再用其他方式排序查詢結果(以下分本敘述之)。本文所提出之方法亦可用於排序文獻。

## 2.3 BM25

BM25(Robertson et al. 1998)為效果不錯的文章排序方法(Boyack et al. 2011)。現今仍有許多研究使用 BM25 進行實驗與技術的開發 BM25。公式如圖 3 所示，BM25 對於文章中所出現的查詢詞 t，利用 IDF 考慮 t 的獨特性，若 t 很少出現在其他文獻中，則 t 很可能是具有獨特性的字詞，則其 IDF 高。利用 t 在文章中的出現次數 TF，當文章 t 出現越多次，則代表此文章與 t 的關聯機率越高。BM25

認為文章之長度會影響 t 的出現次數，故利用文章長度來正規化進而調整相關性分數。除此之外，BM25 考慮了整體查詢詞 $a_1$ 出現於文章 $a_2$ 之情形，以避免 $a_1$ 中某個字詞的 TF 值過大而影響 $a_1$ 與 $a_2$ 之相似度，故有了除以 TF 的機制。本研究與 BM25 最大之差異是 TEE 考慮了生醫物件出現位置，當生醫物件出現在標題及摘要結論段落時，我們認為此文章與該物件有高度相關。

$$BM25(a_1, a_2)=\sum_{t\in a_1\cap a_2}\frac{TF(t,a_2)(k_1+1)}{TF(t,a_2)+k1\left(1-b+b\frac{|a_2|}{avgal}\right)}\log_2 IDF(t)$$

| | |
|---|---|
| $a_1$、$a_2$：文章 | N：訓練資料之總文章數 |
| n：訓練資料中出現$a_1$之文章數 | $\|a_2\|$：$a_2$之長度 |
| $TF(t,a_2)$：t 在 d 之出現次數 | $avgal$：訓練資料中之平均長度 |
| 參數：$k_1$、b 用語調控 TF | |

圖 3 BM25 計算公式

## 2.4 Lucene

Lucene 是一個快速且具彈性的全文搜尋引擎，能有效率地為每篇文章進行索引(indexing)，並提供依相關性排序之排名檢索(Hirsch et al. 2007; Michael and Otis 2010)。Lucene 為開放軟體，可免費下載供使用[5]。 Lucene 進行查詢詞(query)與文章(document)相關性分數計算時，考慮了許多計分條件：(1)字詞頻率(TF)、(2)逆向文件頻率(IDF)、(3)協調因子(coordination factor，當查詢詞完整出現於文章中時使其加分)；(4)正規化查詢詞權重(讓不同查詢詞所得之結果間可以相互比較)、(5)搜尋時查詢詞的加分(boost 針對特定字詞 t 進行加分)；(6)文章長度正規化(norm 為了使不同查詢詞之間可以相互比較)。相對於 Lucene，本研究所提出之 TEE 自動依生醫物件在文章中之位置來計算其與該文章之相似度。

## 2.5 依字詞位置檢索文獻

[5] Lucene: http://lucene.apache.org/core/

以往亦有方法考慮了字詞出現位置來檢索文獻。例如，eGRAB (Extractor of Gene-Relevent Abstracts)認為若基因出現於文獻中之特定位置即代表該文獻是以該基因為主要研究目標(Domedel-Puig and Wernisch 2005; Tudor et al. 2010)。考量之字詞位置可包括：(1)基因是否出現在文獻標題中(2)基因是否出現在摘要第一句，(3)基因是否出現在文獻最後一句，(4)基因是否在文章中出現至少三次。然而出現於第一句可能敘述背景描述或是目前已知的研究成果概述，不一定能代表此文章必定是在談論該主題。此外，最後一句雖然談結論的可能性高，但文章結論不一定只局限於最後一句。本研究所提出之 TEE 可以針對多個生醫物件進行處理，且可更具彈性地依生醫物件出現位置來衡量其為目標物件之可能性，不受限於上述考量句子之作法的缺點。

## 2.6 字詞鄰近度

字詞鄰近度(Term Proximity)是考慮文章中查詢詞之間出現位置之相近程度。當查詢詞中的字詞在文章中之位置越鄰近，則該文章與該查詢詞之相關度越高。許多研究認為加入查詢詞間之鄰近度能提高檢索效果(Liu and Huang 2011; Zhao and Yun 2009)。TEE 與此類鄰近度檢索技術之主要差異是：TEE 找出個別文獻中的目標物件（target entity），且不限於目標物件的個數。鄰近度的文獻檢索技術則是在計算文獻與查詢目標間的相似度（非擷取文獻中的目標物件），且需運作於查詢目標為多個的情況下，(當目標為一個物件時，即無鄰近度可用以改善檢索效能)。

# 第三章 TEE 之研發

TEE 之作業流程如圖 4 所示，以下分述相關步驟之作法。收集疾病-基因、化合物-疾病、化合物-基因配對相關之文獻，並將生醫學家所認定真正談論配對關聯之文獻列為目標文獻。針對一組配對和其對應之文獻。經過前處理以及 TEE 的處理，透過排名的程序進行排序。

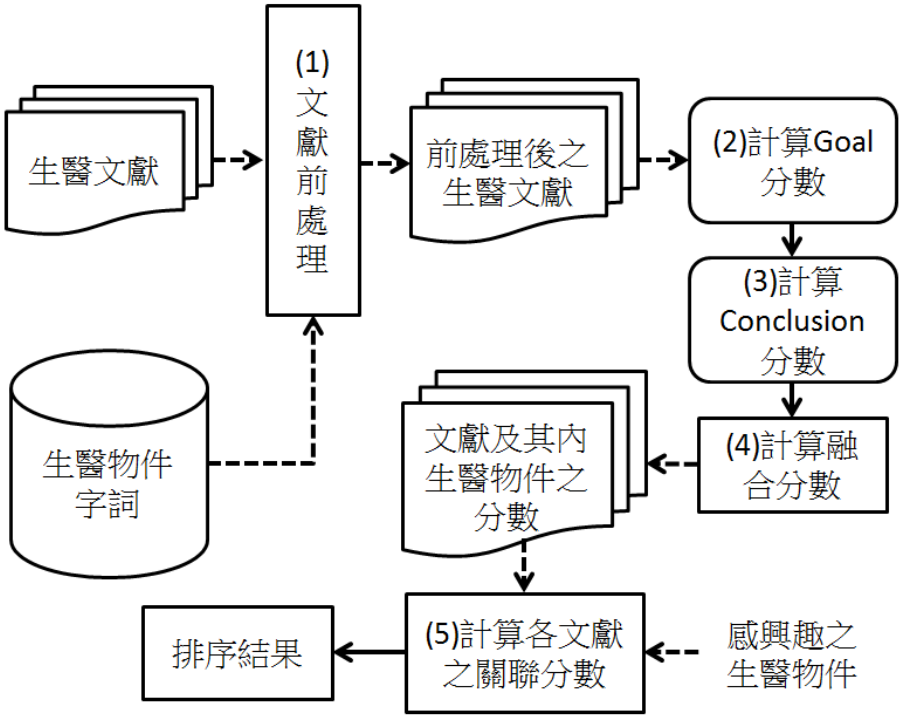


圖 4 TEE 系統流程圖

## 3.1 文獻前處理

為有效辨識文獻中的生醫物件(含疾病、基因與化合物)，我們用生醫字詞詞庫進行長詞判斷，將文獻中的每個字詞切割成獨立字串，以長詞優先的方式判斷出在文獻中的生醫字詞。長詞判斷是將文獻中的每個字詞切割成獨立字串，以長詞優先的方式判斷出在文獻中的生醫字詞。例如文獻中生醫字詞為:

“Dyskeratosis congenita”，長詞判斷會先擷取 Dyskeratosis 到資料庫比對，將會判斷為部分比對到資料庫之字詞，系統會在把下一個字加進來， “Dyskeratosis congenita”進去資料庫比對到一個完全符合的字詞，之後再將文獻中的下一個字加進來 “Dyskeratosis congenital is”，進入資料庫發現搜尋不到，故能判定“Dyskeratosis congenita”是生醫字詞。

我們並利用 PubMed 所定義的停詞(stopword)來濾除文獻中的贅詞[6]。接下來處理字串中之大小寫字元，當遇到首字大寫和單獨出現的大寫字母時轉換成小寫。由於生醫物件之名稱可能會是全部大寫字母(如：RHO)或字母與數字混合(如：BRCA1)，故當遇到全部大寫或是字母混合數字時則將其保留不改變。

此外同義詞處理亦相當重要。由於文章中可能同時會出現全名、別名與縮寫，雖然字詞不同但卻代表著相同的意思，例如「Breast Cancer」、「Tumors, Breast」、「Malignant Neoplasm of Breast」這三個詞代表相同的疾病。CTD 中每個物件包括了該物件之代號(symbol)、原名(name)、編號(ID)及同義詞(synonym)。故我們採用 CTD 之疾病、基因與化合物的字詞庫來協助同義詞處理，把同義詞一律取代為其統一的編號。

對於文獻中字詞之詞型變化，我們亦對文獻及字詞庫中字詞一併進行前處理。例如，cancers 轉換成 cancer，如此一來，可以讓字詞更正確地比對。我們所採用的是 Porter 的演算法[7]。

## 3.2 物件分數之計算與融合

一份文獻中每一個生醫物件有二個分數：Goal 分數及 Conclusion 分數，而此兩個分數需進行融合，以便獲得該物件在該文獻中之得分。如圖 5，當一個生醫物件出現在文獻標題與摘要之中後段，我們認為此文章與該物件有高度相關。

[6] PubMed 之停詞表: https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/
[7] Poter Stemming: https://tartarus.org/martin/PorterStemmer/

故當一個物件出現在標題時 TEE 給該物件 Goal 分數滿分(1 分)，在摘要中越接近結論段落則給予相對高之 Conclusion 分數。

將此兩分數相加後進行正規化。當文獻中該生醫物件比其他生醫物件出現相對多次，即可能代表該生醫物件與文獻有關聯，方法為：該文獻之生醫字詞分數/該文獻所有生醫字詞分數總和。而候選文章中配對詞不一定會出現在文獻摘要中，真正提及配對詞的文獻在摘要中大部分會談論該配對的關聯，因此本研究將配對詞均有出現在文章中的文獻進行加分，公式如圖 4 之(5)。

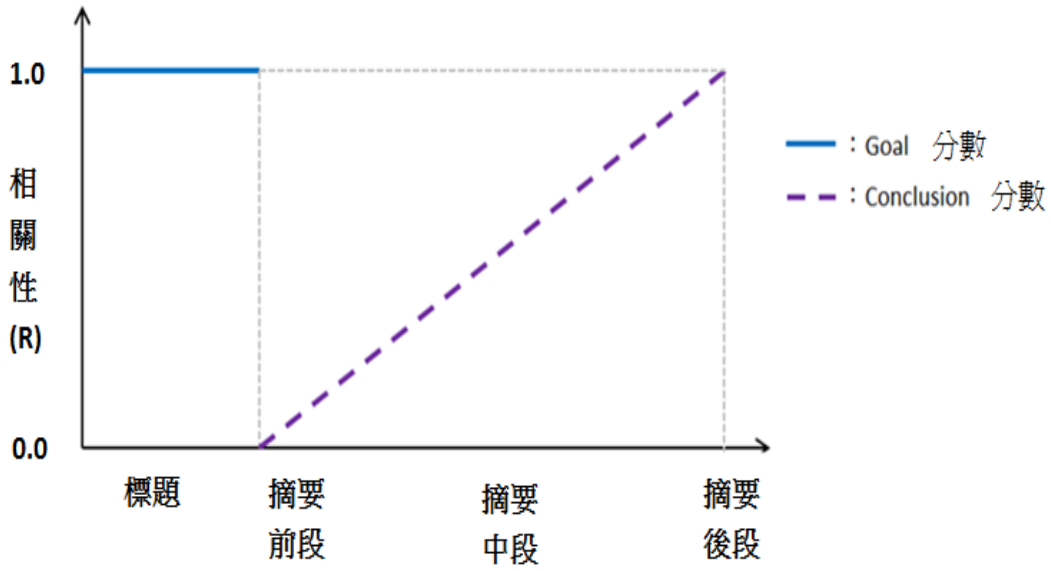


圖 5 Goal 及 Conclusion 關聯分數之計算方法

接下來 TEE 針對物件之 Goal 分數及 Conclusion 分數進行融合及正規化。參見公式(1) ，一個物件 x 在一篇文章 d 之分數為其 Goal 分數及 Conclusion 分數之和。如公式(2)，給定一組物件 p，則 p 在 d 之最後之分數需再經正規化始能獲得。

$$GC(d,x) = Goal(d,x)\,分數 + Conclusion(d,x)\,分數 \qquad 公式(1)$$

$$NGC(d,p) = \frac{\sum_{x \in p} GC(d,x)}{\sum_{y\ appear\ in\ d} GC(d,y)} \qquad 公式(2)$$

此外，因候選文章中配對詞不一定會出現在文獻摘要中，因此本研究將配對詞均有出現在文章中的文獻進行加分，使文獻分數能相較於只出現一個配對詞之文獻高分。故一篇文獻相對於一組生醫物件 p 之分數為：

$$TEE(d,p) = NGC(d,p) + log_2(p\text{內物件出現在}d\text{之個數}) \qquad \text{公式(3)}$$

透過上述加分程序，若一篇文獻中有出現所有目標配對詞，則 TEE 能將其排序在只出現一個配對詞之文獻前面(因 NGC 分數介於 0 及 1 之間)。此設計之意義是能將配對詞均有出現的文章優先給於生醫學家觀看，並依照其配對詞出現於文章的位置計分並排序。

以本文 1.2 節中所討論之文獻為例，針對高血壓(hypertension)及化合物(magnesium)之配對，目標化合物與目標疾病均出現在圖 1 之文獻的標題，在摘要中，magnesium 出現在 12、122 位置，hypertension 出現在 136 的位置：

(1)Magnesium 分數 = 研究目標(Goal)分數 + 研究結論(Conclusion)分數 = 1 + (12+122)/136 = 1.985

(2)Hypertension 分數 = 研究目標(Goal) + 研究結論(Conclusion) = 1 + 136/136 = 2

(3)上述配對分數加總 = 1.985 + 2 = 3.985

(4)該文獻所有生醫物件分數加總：4.496

(5)正規化分數 = 3.985/4.496 = 0.886

(6)log2(物件出現在文章之個數) = 1

(7)文獻得分 = 正規化分數 + log2(配對出現次數)= 0.886+1= 1.886

相同的方法於圖 2，標題只出現了 magnesium；在摘要中 magnesium 出現在 28、64、88 位置，hypertension 出現在 25 的位置再分別除總字數 106：

(1)Magnesium 分數 = 研究目標(Goal)分數 + 研究結論(Conclusion)分數= 1 +

(28+64+88)/106= 2.698

(2)Hypertension 分數 = 研究目標(Goal) + 研究結論(Conclusion) = 0 + 25/106 = 0.235

(3)上述配對分數加總 = 2.698 + 0.235= 2.933

(4)該文獻所有生醫物件分數加總：8.2664

(5)正規化分數 = 2.933/8.2664= 0.354

(6)log2(物件出現在文章之個數) = 1

(7)文獻得分 = 正規化分數 + log2(配對出現次數) = 0.354+1 = 1.354

由於其分數較圖 1 文章之分數為低，故 TEE 可成功地將圖 1 之文獻排在前面。

然而 BM25 在面對這兩篇文獻時，圖 1 之 magnesium 分數為

$$\frac{3*(2+1)}{3+2\left(1-0.75+0.75\frac{|147|}{189.2}\right)}\log_2\frac{61796}{112}$$，

hypertension 分數為

$$\frac{2*(2+1)}{2+2\left(1-0.75+0.75\frac{|147|}{189.2}\right)}\log_2\frac{61796}{328}$$，

故總和為 28.94 分；

圖 2 之 magnesium 分數為

$$\frac{4*(2+1)}{4+2\left(1-0.75+0.75\frac{|112|}{189.2}\right)}\log_2\frac{61796}{112}$$，

hypertension 分數為$\frac{1*(2+1)}{1+2\left(1-0.75+0.75\frac{|112|}{189.2}\right)}\log_2\frac{61796}{328}$，

故總和為 29.92。因此，BM25 會誤將圖 2 之文獻排在前面。故當專注在文章長度與次詞出現的頻率時，而未參考字詞出現的位置，容易錯失與查詢配對相關之文獻。TEE 除了考慮配對詞出現的位置來評估關聯性外，還額外的考慮文獻中是否更著重於討論其他生醫物件，當文獻更專注地討論其他生醫物件時，查詢配對分數就會降低，因此 TEE 可將在談論查詢配對關聯之文獻排序在前面。

# 第四章 實驗評估

為驗證 TEE 於推薦生醫關聯文獻之成效，本研究於 CTD 收集疾病、基因、化合物的交互配對。這些配對是由生醫學家自生醫文獻中仔細審確認後之配對。CTD 並為每一個配對標註相關文獻之 PubMed ID，故我們亦從 PubMed 中收集個別配對之文獻。在此實驗資料基礎上，我們比較了 TEE 和其他較佳之文獻檢索方法的效能。

## 4.1 實驗資料來源

本實驗的目的是要評估 TEE 是否可更有效地將真正在談論查詢特定配對的文獻排序在前面，故在基本實驗資料部分需要生醫配對、經生醫專家人工編選過的目標文獻、及用於排序的候選文獻。故我們從 CTD 收集專家確認過之生醫物件配對，並從 PubMed 收集目標及候選文獻[8]。我們從 PubMed 收集文獻之標題及摘要。實務上標題及摘要為公開資訊(全文則否)，故此實驗設計可完整分析所有文獻，以驗證 TEE 之實務效能。CTD 所收錄之關聯配對包括(1)疾病-基因、(2)化合物-基因、(3)化合物-疾病配對三種。例如，一個基因可能同時擁有多個關聯疾病，故將一個基因對應一個關聯疾病視為一個配對。而該基因也可能與不同化合物有關聯，故每一個關聯亦可視為一個配對。

在文獻中為辨識生醫物件(疾病、基因、化合物)，我們採用需要 CTD 之字詞庫[9]。該字詞庫之組成是分別從 MeSH[10]的化合物字詞資料庫(共 4,196,860 筆)、OMIM 的疾病字詞資料庫(共 77,829 筆)及 NCBI[11]的基因字詞資料庫(共 2,013,675 筆)彙整而來。為考慮一致性，將所有字詞資料和文獻皆依第 3.1 節所述之前處理方式進行處理。

[8] 資料收集日期：2015 年 11 月
[9] http://ctdbase.org/downloads/
[10] MeSH：https://www.ncbi.nlm.nih.gov/mesh
[11] NCBI：https://www.ncbi.nlm.nih.gov/

疾病、基因與化合物除了全名、縮寫和符號，尚有其他別名(平均疾病別名數有 6.5 個，基因平均別名個數為 2 個，化合物別均別名個數為 2.5 個)，亦一併予以前處理(參見 3.1 節)。

表一 實驗設計

| 項目 | 內容 |
|---|---|
| 資料來源 | (1)<基因, 疾病>、<化合物, 基因>、<化合物, 疾病配對>：取自 The Comparative Toxicogenomics Database(CTD)。<br>(2)目標文獻：針對每一個配對，取自 CTD 中生醫學家人工編選所列之參考文獻，並從 PubMed 取其標題及摘要。<br>(3)生醫物件字詞：取 CTD 之基因、疾病與化合物之字詞資料庫(含同義詞)。 |
| 評估準則 | Mean Average Precision(MAP) |
| 對照組 | (1)以字詞權重為主之排名技術<br>(1A)BM25<br>(1B)Lucene<br>(2)以字詞位置為主之排名技術<br>(2A)標題_前一句_末一句<br>(2B)標題_前兩句_末兩句<br>(2C)標題_前三句_末三句<br>(2D)標題_前五句_末五句<br>(2E)標題_前八句_末八句 |

針對每個配對，我們以 CTD 生醫學家編選過的生醫文獻作為其目標文獻。我們從 CTD 之資料庫中隨機取 300 個生醫物件，並將這些物件之所有配對共 96,450 組及生醫學家編選之文獻共 61,796 篇擷取出來。而生醫學家編選過的生醫文獻並非全部都在其摘要中提及目標物件，故實務上即便文獻只出現一個配對詞，仍然需要依照其重要性排序給生醫學家觀看，以促進生醫領域之進展。因此實驗中僅將完全沒出現配對詞的文章移除，過濾之後配對共 77,659 組配對，文獻共 17,343 篇。各配對之生醫學家編選文獻列為目標文獻，其餘有出現配對詞之文獻為候選文獻。

## 4.2 評估準則

Mean Average Precision(MAP)為針對全部候選文獻依據目標文獻排名順序來計分。MAP 計算方式如下所示。當排序方法將各配對中的目標文獻排在越前面，AP 值越高，代表此排序方法有機會排序好相關文獻。反之當排序方法將配對中之目標文獻排名越後面，則 AP 值會越低，表示此排序方法無法將相關文獻排在前面。

$$\text{MAP}=\frac{\sum_{i=1}^{|Q|}P(i)}{|Q|},$$

$$AP(i)=\frac{\sum_{j=1}^{k}\frac{j}{Doc_i(j)}}{k}$$

變數定義：

$i$ ： 第$i$個已關聯配對；

$Doc_i(j)$ ： 第 i 個關聯配對之第 j 個目標文獻之排名；

k ： 第$i$個配對之目標文獻之個數；

Q ： 總測試之配對個數(= 61796)。

順序

| | 1 | 2 | 3 | 4 | 5 | AP |
|---|---|---|---|---|---|---|
| $Pair_1$ | V<br>$p=\frac{1}{1}$ | V<br>$p=\frac{2}{2}$ | | | | $\frac{(\frac{1}{1}+\frac{2}{2})}{2}=1$ |
| $Pair_2$ | | V<br>$p=\frac{1}{2}$ | | V<br>$p=\frac{2}{4}$ | | $\frac{(\frac{1}{2}+\frac{2}{4})}{2}=0.5$ |

V：目標文獻

圖 6 AP 之計算範例

以圖 6 舉例說明，假設有一個排序器面對兩組不同配對(Pair$_1$ 及 Pair$_2$)。假設有五篇候選文獻，其中包含兩篇目標文獻。在 Pair$_1$ 中目標文獻排在第一名和第二名，故 Pair$_1$ 的 AP 會等於 1。在 Pair$_2$ 中目標文獻被排在第二名及第四名，故其 AP 會等於 0.5。由於 Pair$_1$ 的 AP 大於 Pair$_2$，故 Pair$_1$ 排序結果優於 Pair$_2$。將各配對之 AP 平均後可以得到 MAP。MAP 值越大的排序器之效能越好。

## 4.3 實驗之對照組

我們目前實作了(1)BM25、(2)標題_前兩句_末兩句、(3)標題_前一句_末一句、(4)標題_前三句_末三句(5)標題_前五句_末五句(6)標題_前八句_末八句對照組。BM25 之作法以第 2.3 節所述。BM25 在資訊檢索有其重要地位，常被廣泛運用於資訊檢索且在生醫領域有不錯的表現(Boyack et al. 2011)，故我們將 BM25 作為對照組。用於調整字詞頻率(TF)權重之參數 $k_1$ 和 b 分別設為 2 和 0.75。

如第 2.5 節所述，以往已有方法考慮了基因之出現位置來檢索生醫文獻(Domedel-Puig and Wernisch 2005; Tudor et al. 2010)。故我們實作「標題_前一句_末一句」、「標題_前兩句_末兩句」、「標題_前三句_末三句」、「標題_前五句_末五句」、「標題_前八句_末八句」五個對照組，分別作法是考量第一句或最後一句、第二句或最後二句、第三句或最後三句、第五句或最後五句、第八句或最後八句，列為對照組進行實驗。

Lucene 為開放軟體可依不同需求進行修改與設定，容易架設客製化之搜尋引擎，提供全文檢索與搜尋(Hirsch et al. 2007; Michael and Otis 2010)。由於 Lucene 提供開放程式碼，故能更方便地運用在許多不同的領域，我們使用 Lucene6.5 版來實作 Lucene[12]。

[12] Lucene 6.5：http://lucene.apache.org/core/6_5_1/

## 4.4 實驗結果

由於所考慮的是 TEE 與對照組效能之間有無顯著變化，故針對實驗結果進行雙尾 T 檢定(95%信心水準)，檢驗實驗組在效能上是否有顯著差異。圖 7 列出系統之 MAP 效能，TEE 之 MAP 為 0.1387，且其與其它對照組之 MAP 進行 t-test 檢定後，p-value 皆小於 0.01，故 TEE 能將相關文獻排序較前面且顯著優於對照組。

我們再以前述之例子(圖 1 及圖 2)來進一步分析此實驗結果。BM25 在為此兩篇文章計分的過程如前面所述（參見 3.2 節)，它無法將圖 1 之高相關文獻排在前面。而在 Lucene 方面，圖 1 得分為 1.15 分，圖 2 得分為 1.9 分，因此 Lucene 亦無法將高相關文獻排在前面。

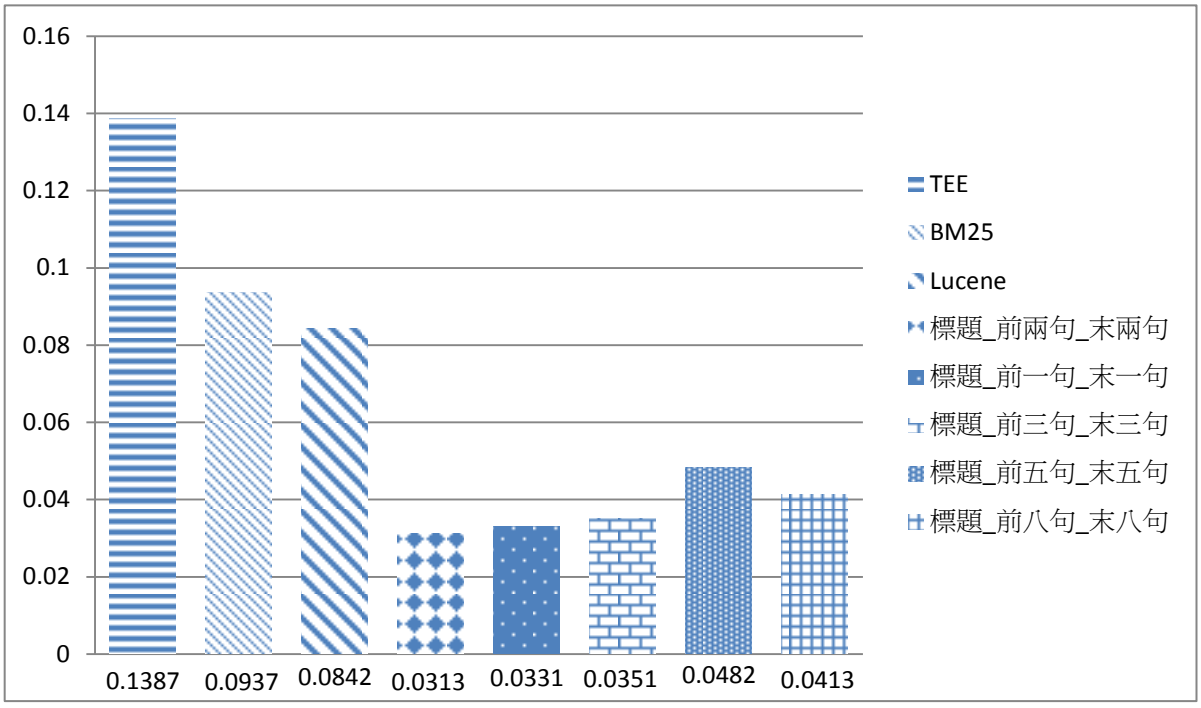


圖 7 TEE 與對照組之 MAP

而在「標題_前五句_末五句」方面，查詢之生醫物件均出現在圖 1 文章的標題各一次，故得 2 分；且亦出現在摘要的前五句及後五句，故得 3 分，總分為 5 分。查詢之生醫物件在圖 2 文章中只出現在標題一次，故得 1 分；且亦出現在前

五句及後五句，故得 4 分，總分為 5 分。由於兩篇文獻均為 5 分，故此無法有效的推薦高相關文獻。而 TEE 可以針對多個查詢物件進行計分，並且可依生醫物件在文獻中出現位置來進行更妥善的計分，進而將高相關的文獻排在前面。

## 4.5 各計分項目之影響

TEE 對每一篇文獻都會有兩項分數(參見公式 3)，這些數值加總才能得到該文獻的相關性分數，進而依此分數將文獻依相關性分數進行排序。因此我們針對 TEE 的兩項計分項目個別計算其效能(結果如圖 8)。我們發現，兩項分數單獨使用時之 MAP 均低於 TEE(第一項 p=0.0011；第二項 p=0.0043)。BM25 加上 TEE 之第二項(Log 分數)後其效能仍顯著遜於 TEE(p-value=0.0076)，故 TEE 之兩項分數融合有助於提升效能。

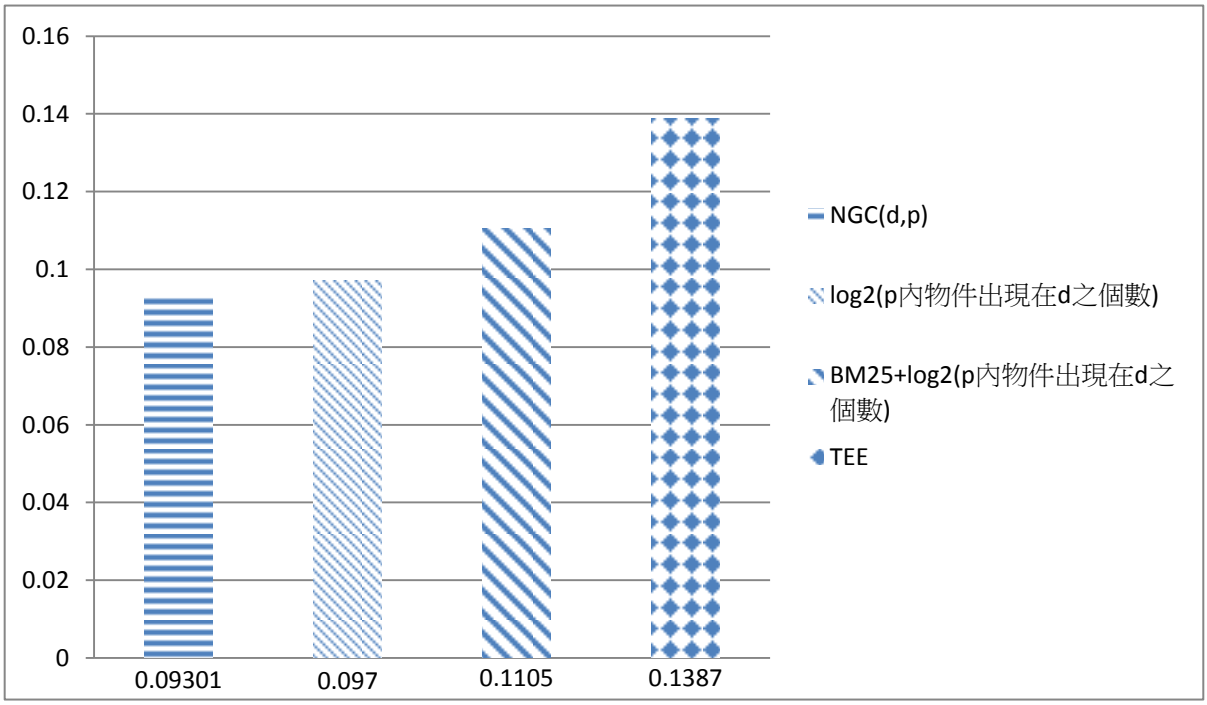


圖 8 各計分項目之 MAP

## 4.6 討論與未來展望

為了探討未來可以更進一步改進TEE的方法，我們分析圖9與圖10之例子。圖 9 是一篇針對葡萄糖(Glucose)與發炎(Inflammation)的文章且由 CTD 所選。其內容有談論查詢配對之關聯，但 TEE 卻將這篇文獻排在第十一名。我們觀察後發現造成如此結果之原因可能為，此篇文章所談論的核心內容是關於 trans-resveratrol(一種化合物)抑制高血糖誘導(hyperglycemia-induced)發炎(Inflammation)，故也提到 Glucose 與若干個基因之反應，並著重於其之反應機制，且於末端中也出現了其他生醫物件。因此，當文獻中同時著重敘說其他生醫物件時，其正規化分數(即 NGC 分數)將下降使得 CTD 所選的生醫文獻排序在後面。而在圖 10 文獻中，其內容談論 Glucose 對 CD14(一個基因)之影響及其機轉，故非討論目標配對之關聯。但此文之最後一句附帶提及高血糖及炎症和組織破壞，因此 TEE 將該文獻排於第一名。從此例可以瞭解到 TEE 之不足。於圖 10 最末句提及高血糖症及發炎和組織破壞，但從語意分析來說，此文章主要是討論高血糖症影響某些基因，並非著重高血糖症與發炎之關聯。因此未來可以考慮加入更經濟有效的語意分析，增加語法及語意之訊息，期望獲得更優秀的排序結果。

此外，文章中作者自行引用之縮寫字詞之處理也很重要。例如圖 12 之文章中提及生醫物件 acute coronary syndrome(一種疾病)，並在其後以 ACS 為縮寫，又因 ACS 並非生醫詞庫之字詞，故沒有被辨識出來。反之，若縮寫之字詞為詞庫中的生醫字詞，就會誤標成另一個生醫物件。由於出現縮寫的字詞通常是作者會一再重覆提到的物件，未來期望為每篇文獻加入縮寫字詞庫，增加字詞辨識的正確率。為達此目標，需能正確辨識縮寫字詞所代表之完整字詞(可能為單字詞或多字詞，亦不一定由第一字母組成縮寫)。

Title: trans-resveratrol inhibits hyperglycemia-induced **inflammation** and connexin downregulation in retinal pigment epithelial cells.

Abstract: The purpose of this study was to determine the inhibitory activity of trans-resveratrol against hyperglycemia-induced **inflammation** and degradation of gap junction inte……. Retinal (ARPE-19) cells were incubated with 5.5 mM **glucose**, 5.5 mM **glucose** and 10 microM resveratrol, 33 mM **glucose**, or 33 mM **glucose** and 0-10 microM trans-resveratrol at 37 degrees C and 5% CO(2) for 9 days. Cell viability was determined by the crystal violet assay. The levels of low-grade **inflammation** biomarkers …… endothelial growth factor (VEGF) were determined by the enzyme-linked …...The expression levels of protein kinase Cbeta (PKCbeta), connexin 43 (Cx43), transforming growth factor-beta1 (TGF-beta1), and cyclooxygenase-2 (COX-2) were determined by Western blot. Incubation of retinal cells with 10 microM trans-resveratrol in the presence of 5.5 mM **glucose** did not affect any of the biomarkers investigated. Incubation of ARPE-19 cells with 33 mM **glucose** for …… accumulation of VEGF, IL-6, IL-8, TGF-beta, and COX-2, activation of PKCbeta, and reduction of Cx43 and GJIC. Incubation of ARPE-19 cells with 33 mM **glucose** in the presence of 0-10 microM trans-resveratrol dose-dependently inhibited VEGF, TGF-beta1, COX-2, IL-6, and IL-8 accumulation, PKCbeta activation, and Cx43 degradation and enhanced GJIC. These data suggest that trans-resveratrol can protect the retinal pigment epithelial cells against hyperglycemia-induced low-grade **inflammation** and GJIC degradation.

（節錄來源：https://www.ncbi.nlm.nih.gov/pubmed/?term= 20578705）

□：目標疾病； ▒▒：目標化合物； ＿＿：其他生醫物件

圖 9 CTD 所選之參考文獻：<glucose, inflammation>

Title: High **glucose** enhances lipopolysaccharide-stimulated CD14 expression in U937 mononuclear cells by increasing nuclear factor kappaB and AP-1 activities.

Abstract: We have demonstrated recently that high **glucose** augments lipopolysaccharide …… monocyte-derived macrophages. Since CD14 is a …… that high **glucose** enhances CD14 expression. In the present study, we determined the effect of high **glucose** on CD14 expression …… to normal or high **glucose** for 2 weeks…… high **glucose** by itself did not increase CD14 expression significantly, it augmented LPS-stimulated CD14 expression …… membrane-associated and soluble CD14 protein levels by high **glucose**. Further investigations using transcription factor activity assays and gel shift assays revealed that high **glucose** augmented LPS-stimulated CD14 expression by enhancing transcription factor nuclear factor kappaB (NFkappaB) and activator protein-1 (AP-1) activities. Finally, studies using …… showed that CD14 expression is essential …… by high **glucose**. Taken together, this study has demonstrated a robust augmentation by high **glucose** of LPS-stimulated CD14 expression through AP-1 and NFkappaB transcriptional activity enhancement, elucidating a new mechanism by which hyperglycemia boosts LPS-elicited gene expression involved in **inflammation** and tissue destruction.

（節錄來源：https://www.ncbi.nlm.nih.gov/pubmed/?term= 18180316）

□：目標疾病； ▒▒：目標化合物； ＿＿：其他生醫物件

圖 10 非 CTD 所選之參考文獻：<glucose, inflammation>

Title: Investigation of the inhibitory effect of N(G)-nitro-L-arginine methyl ester on the **antihypertensive** effect of the angiotensin AT1 receptor antagonist, GR138950.

Abstract: 1. The effect of systemic administration of the nitric oxide synthase inhibitor, N(G)-nitro-L-arginine methyl ester (L-NAME) on the **antihypertensive** effects of …… inhibitor, **zaprinast** on the **antihypertensive** effect …… that of **zaprinast** in conscious…… the **antihypertensive** effects …… rats. 4. **Zaprinast** pretreatment did not affect the **antihypertensive** effect …… nitroprusside. **Zaprinast** alone caused……The **antihypertensive** effect ……whereas **zaprinast** caused a small…… inhibit the **antihypertensive** action…… of **zaprinast** to potentiate the **antihypertensive** effects …… is unknown.

(節錄來源：https://www.ncbi.nlm.nih.gov/pubmed/?term= 9421286)

□：疾病； ▒：目標化合物； ＿＿：其他生醫物件

圖 11 CTD 所選之參考文獻：<zaprinast, hypertension>

Title: Presence and extent of coronary artery disease by cardiac computed tomography and risk for **acute coronary syndrome** in cocaine users among patients with chest pain.

Abstract: Cocaine users represent an emergency department (ED) population that has been shown to be at increased risk for **acute coronary syndrome** (**ACS**); however, there is controversy about whether this higher risk is mediated through advanced atherosclerosis. Thus, we aimed to determine whether history of cocaine use is associated with **ACS** and coronary artery disease. ……, 44 had a history of cocaine use (9%) and were matched to 132 controls (mean age 46 +/- 6 years, 86% men). History of cocaine use was associated with a sixfold higher risk for **ACS** (odds ratio 5.79, 95% confidence interval 1.24 to 27.02, p = 0.02), …… In conclusion, in patients presenting to the emergency department with acute chest pain, history of cocaine use is associated with an increase in risk for **ACS**; however, this was not attributable to a higher presence or extent of coronary atherosclerotic plaque.

(節錄來源：https://www.ncbi.nlm.nih.gov/pubmed/?term= 19231323)

□：疾病； ▒：疾病縮寫

圖 12 CTD 之參考文獻

# 第五章 結論

生醫專家致力於了解生醫物件之間的關聯，嘗試找出其之間的關聯，以發現關聯為基礎，進而發展疾病診斷、基因治療、預防和新藥來幫助患者。生醫學家將研究成果發表生醫文獻，而生醫科技的日新月異使得文獻數量快速成長，導致疾病、基因與化合物關聯之文獻檢索不易。本研究提出 TEE 之技術，可針對使用者輸入之若干生醫物件，將生醫文獻依其與該生醫物件之關聯分數排序。TEE 的目的是推薦真正在談論查詢物件之文獻給生醫領域研究學者、關聯編撰人員與文獻探勘系統等使用者。經實驗證實 TEE 顯著優於其他文件排序技術。

給定大量文獻，找出文獻中之主要討論物件，進而進行探索分析，藉由 TEE 之輔助，使生醫學家更能有效及時地收集這些已於文獻中發表之關聯，更可以找到高度相關之文獻，促進生醫研究之進展。

# 參考文獻


[1] Ahmed ST, Chidambaram D, Davulcu H, Baral C: IntEx: A Syntactic Role Driven Protein-Protein Interaction Extractor for Bio-Medical Text. Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics; Detroit, Michigan. 1641492: Association for Computational Linguistics 2005: 54-61.

[2] Baral C, Gonzalez G, Gitter A, Teegarden C, Zeigler A: CBioC: Beyond a Prototype for Collaborative Annotation of Molecular Interactions from the Literature. Computational systems bioinformatics / Life Sciences Society Computational systems bioinformatics 2007,6:381-384.

[3] Boyack KW, Newman D, Duhon RJ, Klavans R, Patek M, Biberstine JR, et al. Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. PLoS ONE 6(3): e18029, 2011.

[4] Domedel-Puig N, Wernisch L: Applying GIFT, a Gene Interactions Finder in Text, to fly literature. Bioinformatics 2005: 3582-3583.

[5] Frijters R, van Vugt M, Smeets R, van Schaik R, de Vlieg J, Alkema W: Literature mining for the discovery of hidden connections between drugs, genes and diseases. PLoS Computational Biology 2010, 6(9).

[6] Hirsch L, Hirsch R, Saeedi M: Evolving Lucene search queries for text classification. In: Proceedings of the 9th annual conference on Genetic and Evolutionary Computation ; London, England. 1277279: ACM 2007: 1604-1611.

[7] Jensen LJ, Saric J, Bork P: Literature mining for the biologist: from information retrieval to biological discovery. Nature Reviews Genetics 2006, 7:119-129.

[8] Kim J, So S, Lee HJ, Park JC, Kim JJ, Lee H: DigSee: disease gene search engine with evidence sentences (version cancer). Nucleic Acids Research 2013,41:W510-517.

[9] Liu RL, Huang YC: Ranker enhancement for proximity-based ranking of biomedical texts. Journal of the American Society for Information Science and Technology 2011, 62(12):2479-2495.

[10] Liu RL, Shih CC: Identification of highly related references about gene-disease association. BMC Bioinformatics 2014, 15:286.

[11] Michael McCandless EH, and Otis Gospodnetić: Lucene in Action, Second Edition; 2010.

[12] Ö zgür A, Vu T, Erkan G, Radev DR Identifying gene-disease associations using centrality on a literature mined gene. Bioinformatics 2008,24(13) : 277-285.

[13] Patnala R, Clements J, Batra J: Candidate gene association studies: a comprehensive guide to useful in silico tools. BMC Genetics 2013 14:39.

[14] Robertson SE, Walker S, Beaulieu M: Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive. proceedings of the 7th Text Retrieval Conference 1998: 253-264.

[15] Singhal A: Modern information retrieval: A brief overview. IEEE Data Eng Bull 2001, 24(4):35-43.

[16] Thomas C Wiegers, Allan Peter Davis1, K Bretonnel Cohen, Lynette Hirschman and Carolyn J Mattingly: Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD). BMC bioinformatics 2009,10:326.

[17] Tsuruoka Y, Tsujii J, Ananiadou S: FACTA: a text search engine for finding associated biomedical concepts. Bioinformatics 2008, 24(21): 2559–2560.

[18] Tudor CO, Schmidt CJ, Vijay-Shanker K: eGIFT: mining gene information from the literature. BMC bioinformatics 2010,11:418.

[19] Winnenburg R, Wächter T, Plake C, Doms A, Schroeder M: Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? Briefings in Bioinformatics 2008,9(6):466-478.

[20] Zhao J, Yun Y: A proximity language model for information retrieval. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval: 2009.ACM: 291-298.