文件分類在醫學資訊片段檢索之應用初探
A Preliminary Study on Applying Text Categorization to Medical Passage Retrieval

林宗興 Tsung-Hsing Lin
慈濟大學醫學資訊學系(所)
96325114@stmail.tcu.edu.tw

劉瑞瓏 Rey-Long Liu
慈濟大學醫學資訊學系(所)
rlliutcu@mail.tcu.edu.tw

摘要

以網際網路來尋找醫學資訊之相關資訊與知識，已漸成為大眾化之途徑，使用者常需於大量資訊中尋覓具有特定主題或概念之資訊片段，故醫學資訊片段檢索系統相當重要。本文實作並分析常見之三種資訊片段檢索方法，發現此三種以字詞頻率為主要考量的方法於實務上尚有待改良之處。所以我們揭示一個以語意為主要考量的方法，探究文件分類在醫學資訊片段檢索上之可能貢獻，並探討其與其他方法結合之方式，期能獲得兼備字詞頻率及語意考量之醫學資訊片段檢索系統，做為醫療決策支援與醫學資訊探索之用。

關鍵字：文件分類、資訊片段檢索、醫學資訊探索

Abstract

The Internet has been a popular platform on which users find medical information and knowledge. In the huge information space on the Internet, users usually need to seek passages that convey specific topics or concepts. Medical passage retrieval is thus essential. In this paper, we implement and analyze three popular passage retrieval methods, which rank passages on term frequency computation. We identify their weaknesses in supporting medical passage retrieval, and accordingly propose a framework that employs text categorization to rank passages. We also discuss several ways to combine the framework with other methods, so that the system may consider both term frequencies and semantics of passages. The contribution is significant to the development of medical passage retrieval systems to support healthcare decision making and medical information exploration.

Keywords: Text Categorization, Passage Retrieval, Medical Information Exploration.

1. 簡介

對醫護專業人員及一般民眾而言，網際網路已漸成為取得醫療資訊的主要管道[1][3]。醫護專業人員可經由網際網路獲得專業資訊，而一般民眾則除了可從醫護人員的口中得知醫療資訊之外，亦可經由網際網路獲取感興趣之醫療資訊來促進身心健康。

但現今環境中，使用者常無法針對特定醫療資訊進行更詳細之探索[1]，此乃因網際網路上有著大量之文章，而每一篇文章中常含有一個以上的主題或是概念，但使用者所感興趣之內容往往只有文章中的一個片段（passage）[2][5][7][8]。一個文章片段是一個傳遞特定觀念或想法的語意單元。於醫療領域中，使用者常從不同管道（如醫囑、醫藥新知等）獲得感興趣之文章片段（Passage Of Interest，POI），並以此片段為起始，進行相關醫學資訊之瞭解與探索，找出更多相關之文章片段。然文章片段通常相當簡短，造成其語意不易辨識，使得其檢索更形困難。

本文提出上述醫學資訊片段檢索在理論及實務上之意義（第二節），並在實際環境中測試分析現有相關技術在醫學資訊片段檢索上之不足（第三節），進而發現文件分類在此方面之可能貢獻（第四節）。本研究之發現可做為未來相關研究之重要基礎。

2. 醫學資訊片段檢索

醫學資訊片段檢索（Medical Passage Retrieval，MPR）是要支援一般民眾及醫護人員進行深入之醫學資訊探索，故可成為一個醫療決策支援之輔助工具。其運作模式是由使用者輸入一個POI當作探索之起點，進行高度相關片段之探索。因輸出為與POI相關之片段，故可更直接、快速地符合使用者需求，方便進行特定議題之深入探索。

MPR 不同於以往的以文找文方法。其輸出入皆為 passage，長度遠比一篇完整的文章要短的多，使主要的概念較不易突顯及擷取[2]，此為其主要的挑戰之一。

另MPR也不同於一般直接鍵入關鍵字來查詢的方法。雖說POI字數已不到整篇完整文章數量，但往往還是超過某些搜尋引擎可接受字數的上限（例如：Google 的查詢上限為 32 個字詞）。此外，單以關鍵字來查詢往往僅能考量字詞頻率，無法兼顧完整語意。一般使用者常將所感興趣之片段直接輸入搜尋引擎，舉例來說，實際輸入以下POI至Google搜尋引擎[1]，希望可以獲知更多mRNA與tRNA於擺動學說中的關係：「在蛋白質合成時，mRNA上的密碼和tRNA上的反密碼是對應的。已知道 20 種氨基酸有 61 種對應的密碼子，按照擺動學說（wobble hypothesis），最少需要 32 種tRNA才能完全識別mRNA中的 61 個密碼子」。系統只回覆與POI相同出處的文章。原因可能是Goole是以字詞頻率為主要考量，而與POI有較多重疊字詞者只有此一篇文章。

類似情形也可於其他搜尋引擎中發現。例如，我們於奇摩搜尋引擎[2]輸入以下POI，希望可以瞭解茄紅素對抗乳癌、攝護腺癌、及心臟病的療效：「蕃茄中的茄紅素，近幾年來已被醫學證實具有防癌功效，特別是濃縮成蕃茄醬，蕃茄湯和蕃茄汁的蕃茄，可以幫助對抗乳癌。並且，曾有研究表示食用蕃茄可以減少男人罹患攝護腺癌和心臟病的風險」。系統回傳排名前十篇中有六篇恰為此POI出處之文章，其餘四篇中，一篇提到其他關於茄紅素的介紹，兩篇提到其他與POI無關之健康資訊，一篇為損毀網頁。

由以上兩例中得知，鍵入關鍵字於搜尋引擎方法似乎無法支援使用者針對 POI 做更進一步的探索。為方便使用者對醫學資訊做深入探索，且跳脫出以往的以文找文或關鍵字搜尋方法， MPR 即顯得相當重要。MPR 因其直觀的 POI 輸入，及簡短直接的片段輸出，可讓使用者方便做後續性的搜尋，在實務上提供更快速且準確地探索。所以，如欲深入瞭解上述兩例中之相關資訊，可於 MPR 輸入 POI，即可獲得具有高度相關之 passages，幫助使用者有進一步的瞭解，進而達到醫學資訊探索之目標。

1 於 2007 年 7 月 18 日搜尋。Google網址http://www.google.com.tw/。
2 於 2007 年 7 月 20 日搜尋。奇摩網址http://www.yahoo.com.tw/。

3. 現有方法於 MPR 上之效能分析

傳統之資訊片段檢索技術多運用於 question answering（QA）上，且植基於字詞頻率相似度的測量[9]。本節針對目前常見之資訊片段檢索方法進行實際效能分析，希望從中探得未來改良 MPR 之方向。

Figure 1 為本效能分析之整體流程。圖右為訓練階段，圖左為測試階段。我們實際設計幾個常見之方法，依照與POI間之相似度來排序並輸出資料庫中之片段（Database Passages, DPs）。

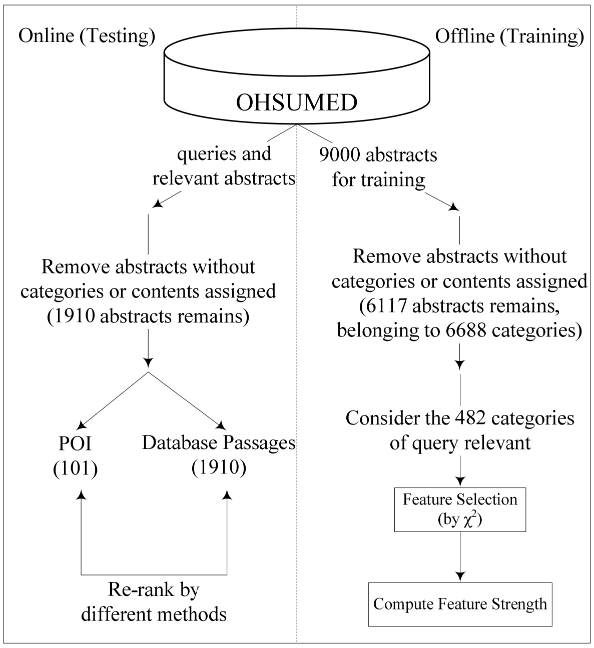

Figure 1 現有方法於 MPR 效能分析之流程

3.1. 實驗資料

我們使用 OHSUMED 資料庫作為實驗資料[10]，此資料庫為一臨床導向之 MEDLINE 文件之子集合，由 348,566 篇摘要（abstracts）組成，涵蓋 1987-1991 年份。首先，從 OHSUMED 資料庫 1987 年份隨機取出 9000 篇作為訓練文章。接著濾除無類別或是無摘要之文章，產生 6117 篇訓練文章。

在訓練階段，於訓練文章所屬類別中，選出與 Query 相關文章中的 482 個類別，先以 chi-square （χ^2） technique 擷取出特徵詞，並依 Term Frequency-Inverse Document Frequency（TFIDF）計算這些特徵詞的強度。

在測試階段，採用 OHSUMED中作為測試用之 101 個

queries[3]，將各自之patient information及information request合併為 101 個POI。接著將經過專家判定為與此 101 個queries相關的文章，濾除無類別或是無摘要之文章後，共有 1910 個DPs。之後依不同方法計算分數，決定輸出DPs之順序。

3.2. 現有常見之方法

常見之資訊片段檢索方法是以字詞重疊為主要考量。本研究實作其中三種常見方法：

第一個方法為 Density Based Passage Retrieval（DBPR）[6]，計算 DP 出現在 POI 中字的密度（Density）作為排序的依據。Density 定義如下：

$$Density\ (s) = \frac{number\ of\ terms\ in\ both\ s\ and\ POI}{Size\ of\ s}$$

其中 s 為一個 DP ，Size of s 為 s 中總字數。

此法希望能從字詞頻率的考量為 DPs 做排序。當交集之密度越高，代表此 DP 越與 POI 相似。

第二個方法為 Weighted Density Based Passage Retrieval（WDBPR）[4]，是改良上述 DBPR 方法，以 feature strength 作為權重，來計算 Density。Weighted Density 定義如下：

$$Weighted\ Density\ (m) = \frac{\sum_{t \in m; t \in POI} Feature\ strength\ (t)}{\sum_{f \in m} Feature\ strength\ (f)}$$

其中 m 為一個 DP；

Feature strength (x)：字詞 x 之 TFIDF 特徵詞強度值；

此法是以 feature strength 為權重，希望依不同重要性之特徵字給予不同權重，其目的是提升相似度計算之準確度。

第三個方法為 Cosine Similarity Passage Retrieval（CSPR），運用 chi-square （χ^2）technique 擷取出特徵詞，以 feature set 作為向量維度，計算 POI 與每一個 DP 之向量的 cosine 相似度，並依此為排序依據。此方法亦常被用來計算文章間之相似度。cosine 相似度的定義如下：

$$Similarity\ (DP) = \frac{Vector_{POI} \cdot Vector_{DP}}{|Vector_{POI}| \cdot |Vector_{DP}|}$$

其中 $Vector_{POI}$：POI 以 feature set 為維度之向量；
$Vector_{DP}$：DP 以 feature set 為維度之向量；

3.3. 評估準則

為衡量各種方法之排序成效，我們採用三個常被引用之評估準則：MAP、MRR 及 P@5。說明如下：

（1）MAP(Mean Average Precision)＝ $\frac{\sum_{i=1}^{101} P(i)}{101}$, $P(i) = \frac{\sum_{j=1}^{k} \frac{j}{R(j)}}{k}$

其中

k：與 query 相關之 DPs 之總數；

R（j）：與 query 相關之第 j 個 DP 在排序後之名次。

可以從此準則看出，所有與 query 相關之 DPs 若能被排於越前面（名次越小），則 MAP 值越大，代表整體表現度越好。

（2）MRR(Mean Reciprocal Rank)＝ $\frac{\sum_{i=1}^{101} \frac{1}{M(i)}}{101}$

其中

M（i）：對第 i 個 query 而言，第一個相關 DP 之排序名次。

MRR 值越大，代表與每個 query 第一個相關之 DP 於排序中被成功地排在前面。

（3）P@5(Precision at five)＝ $\frac{\sum_{i=1}^{101} \frac{N(i)}{5}}{101}$

其中

N(i)：對第 i 個 query 而言，相關 DP 被排於前 5 名之個數。

P@5 值越大，表示越多排序前 5 名之 DPs 是符合需求的。

3.4. 結果與討論

測試結果如 Table 1。整體而言，WDBPR 及 CSPR 在 feature set越大時，表現最好。在feature set大小為30000 時（共有 30360 個不同字詞於訓練文章中），三個測試方法優劣依序為 CSPR>WDBPR>DBPR。可以看出加了feature strength做為權重的WDBPR較只考慮字詞頻率的 DBPR 表現度佳，而考慮 feature set 向量相似度的 CSPR，其排名準確度又比 WDBPR 高。

另外，從結果中發現 WDBPR 及 CSPR 皆有 feature 數越大，表現越好的趨勢，會有此現象推測原因是，訓練文章中之字詞恐不足以包含測試文章之字詞。此應為實務上常見之情況，也是未來改進方向之一。

[3] 原 OHSUMED 中有 106 個 queries，在濾除無相關文章的 queries 後，共餘 101 個 queries。

Table 1 DBPR、WDBPR、CSPR 表現結果

方法 / Feature 數	DBPR	WDBPR	CSPR
10000		MAP:0.13798 MRR:0.29859 P@5:0.2	MAP:0.290422 MRR:0.46811 P@5:0.34455
20000	MAP:0.20757 MRR:0.44678 P@5:0.26930	MAP:0.24611 MRR:0.46062 P@5:0.27326	MAP:0.46333 MRR:0.66345 P@5:0.52871
30000		MAP:0.39042 MRR:0.60496 P@5:0.4099	MAP:0.48408 MRR:0.68873 P@5:0.54653

以上三種方法皆是從字詞頻率方面做為輸出 DPs 排序的考量，但就實務上而論，另應考慮語意上的相似程度，方可達到更精確之檢索結果。

4. 文件分類於 MPR 之意義

為進一步增進 MPR 之效能，我們深入分析上述實驗中表現最佳之做法（即 CSPR）之缺失，並進而探討可能之改進方法。

我們以第 21 個 Query 為例，Table 2 列出以 CSPR 搜尋 Query21 前十名之文章，表格中文章以 OHSUMED 編號及其歸屬之 Query 編號表示，並依類別出現次數做計數。我們取出 Query21 之所有相關文章中之 MeSH Terms，並將其與經 CSPR 排序後的前十名文章中之 MeSH Terms 相比對。結果發現，原排在第十名的相關文章之 MeSH Terms，與 Query21 相關文章之 MeSH Terms 之重疊數最高，而原排在第一名的非相關文章則低許多。原排在第一名文章討論的內容為"colonoscopy"，但 Query21 是談"hypertension"，CSPR 使其排在第一名是因為"strategy"出現字詞頻率很高，但語意上卻完全無關。另外，第八名非相關文章與第九名相關文章之字詞重疊數一樣，其原因為第八名文章內容為"煙、酒成癮"，第九名文章內容為"煙、酒易引起 hypertension"，此例可說明某些文章在字詞頻率上與 Query 相符合，但卻不是相關的文章。

所以，我們嘗試以文件分類方法來進行 DPs 重新排序，以下為進行步驟(如 Figure 2)：首先，將 OHSUMED 中專家所判定與 query 相關文章之 MeSH terms 記錄起來，接著個別與 DPs 中的 MeSH terms 做比對，累計出該 DP 與 query 間重疊之類別數，並依此分數進行排序，表現結果如 Table 3，與上述實驗最佳作法之 CSPR（feature set 大小為 30000）相比，此法確可明顯提升表現效能。

Table 2 以 CSPR 搜尋 Query21 前十名文章

DocID/QueryID	出現一次	出現兩次	出現六次	符合之類別數
53134/46	1	1	1	3
53134/99	1	1	1	3
283240/70	4	1	1	6
250599/69	2	1	1	4
16246/50	3	2	1	6
195496/54	1	0	1	2
108878/50	2	3	1	6
43600/35	5	2	1	8
250601/21	5	2	1	8
142210/21	6	2	1	9

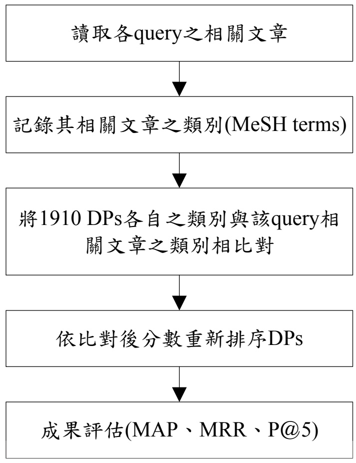

Figure 2 文件分類實驗流程圖

Table 3 文件分類於 MPR 表現結果

方法 / Feature 數	文件分類法
MAP	0.64923
MRR	0.98349
P@5	0.73663

由以上分析可得知，文件分類方法很可能大幅改善只考慮字詞頻率的方法，使其得到更佳之結果。醫學資訊片段檢索實與文件分類有著不可分離的關係。所以，我們中揭示一個架構將 MPR 植基於文件分類（如 Figure 3 所示），期能達到更精確之搜尋結果。圖右為訓練階段，先用訓練資料建立分類器（步驟 1），並使用此分類器將資料庫中之 passages 建成向量（步驟 2），其中每一個維度代表此 passage 可能被歸於對應類別之程度。圖左為測試階段，使用分類器將 POI 建成以類別相似度為維度之向量（步驟 3），與資料庫中

passage vectors 做相似度比較（步驟 4），接著輸出資料庫中重新排序過之 passages（步驟 5），提供給使用者做醫療決策支援或醫學資訊探索。

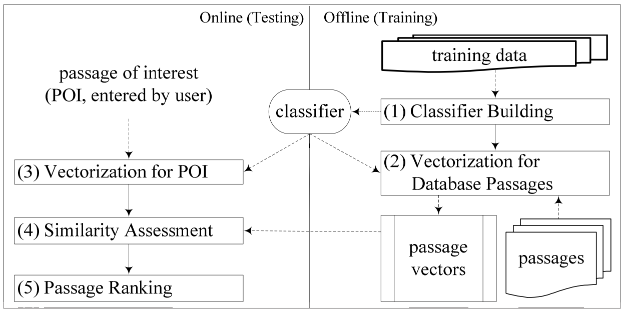

Figure 3 植基於文件分類的醫學資訊片段檢索系統

會以類別為基礎做為向量之維度，是因實務上我們常將各事物歸入一至數個類別。對於 POI 而言，也有其所歸屬的類別，所以依 DPs 中與 POI 對應類別之相似度來進行排序。

雖然類別上的相近也隱含了語意上是有相關的，但是被歸入同個類別的片段，其語意上卻不見得與 POI 有極大程度的相關。例如，兩個分別談論肺癌及肝癌之文章，雖同歸於”癌症”類別，但其語意內容則相當不同。故輔以字詞考量應可增進搜尋之準確度。以下我們列舉出四種結合文件分類與字詞相似度的方式：

第一種結合方式：先用 CSPR 取前特定數目之文章（如前十篇），再用文件分類方式重新排序。

第二種結合方式：先用文件分類方式取出前特定數目之文章（如前十篇），接著再用 CSPR 重新排序。

第三種結合方式：將文件分類及 CSPR 各自與 POI 相似度的分數融合成為一個分數，以進行重新排序。由於各自分數差異很大，融合前需做正規化之調整，此外融合之權重係數亦需有恰當之方法來設定。

第四種結合方式：將文件分類以類別做維度的向量，與 CSPR 以特徵詞做維度之向量結合，形成一個涵蓋語意及字詞維度之向量，並依此向量計算 POI 及 DPs 之相似度。此方法同時考慮到語意及字詞層面，唯需克服的困難為特徵詞數量遠大於類別數量，應盡量平衡此兩層面之維度與權重。

5. 結論

網際網路漸成為一般民眾及醫護人員獲取醫學資訊相關訊息不可或缺的來源。而使用者於過濾及篩選符合興趣之資訊的過程又需花費許多時間與精力，故醫學資訊片段檢索相當重要，此檢索方式可提供更直接、快速的搜尋結果，可讓使用者做更深入且廣泛之醫學資訊探索。一般民眾可從醫囑或健康資訊文件中選定一個感興趣之片段，進行更深入之探索；而醫療專業人員亦可在有特定資訊需求片段下，直接將此片段輸入系統，進行專業知識之研究。

本研究實作目前常見的三種資訊片段檢索方法，經分析後，發現僅以字詞頻率為主的評估方式於實務考量上尚有待改良空間。我們更進一步嘗試將 MPR 植基於以語意為主的文件分類，在實際測試過後，其表現度確有明顯提升。據此，我們進一步提出文件分類植基於醫學資訊片段檢索的系統架構，並討論幾種結合的方法，期能獲得兼備字詞頻率與語意考量之醫學資訊片段檢索系統，提供做為醫療決策支援或醫學資訊探索之用。此研究成果可為後續相關研究提供相當有意義之參考。

6. 誌謝

本研究承國科會研究計劃補助 (計劃編號：NSC 96-2221-E-320-001-MY3)，謹此誌謝。

參考文獻

[1] K. S. Shuyler and K. M. Knight, “What Are Patients Seeking When They Turn to the Internet? Qualitative Content Analysis of Questions Asked by Visitors to an Orthopaedics Web Site”, Journal of Medical Internet Research, Vol. 5, No. 4, 2003

[2] Gerard Salton, J. Allan, C. Burckley, “Approaches to Passage Retrieval in Full Text Information Systems”, ACM-SIGIR’93-6/93/Pittsburgh, PA, USA

[3] Heather A. Liszka, Terrence E. Steyer, William J. Hueston, “Virtual Medical Care: How Are Our Patients Using Online Health Information?”, Journal of Community Health, 2006

[4] Ittycheriah A., Franz M., Zhu W.-J., and Ratnaparkhi A., “IBM’s statistical question answering system”, In Proceedings of the 9th

Text Retrieval Conference, 2000.

[5] James P. Callan, "Passage-Level Evidence in Document Retrieval", Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1994

[6] Light M., Mann G. S., Rilloff E., and Breck E., "Analysis for elucidating curruent question answering technology", Natural Language Engineering 7 (4): 325－342. , 2001.

[7] Marcin Kaszkiel, Justin Zobel, "Passage Retrieval Revisited", SIGIR' 97, Philadelphia PA, USA

[8] Renxu Sun, "Mining Dependency Relations for Query Expansion in Passage Retrieval", SIGIR' 06, August 6 – 11, 2006, Seattle, Washington, USA.

[9] Tellex S., Katz B., Lin J., Fernandes A., and Marton G.. "Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering", SIGIR'03, July 28–August 1, 2003, Toronto, Canada.

[10] William Hersh, "OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research", 1994_SIGIR_pp 192-201.