中文健康資訊搜尋結果自動分群

Clustering Search Results of Chinese Health Information

陳品守, 林紋正*
Pin-Shou Chan, Wen-Cheng Lin*
慈濟大學醫學資訊研究所
**通訊作者: 林紋正*, denislin@mail.tcu.edu.tw

摘要

在本研究中我們提出了一個搜尋結果分群方法，針對中文的醫療健康資訊搜尋結果來加以自動分群，讓使用者能方便地瀏覽。我們採用詞彙的共現資訊來作為分群時的依據，依此建立基礎群，然後再將基礎群合併。實驗結果顯示我們的系統能產生較大的群組，涵蓋較多的網頁，而且並沒有降低準確率，可以提供較完整的群組給使用者查閱。

關鍵字：健康資訊檢索、搜尋結果分群、詞彙共現資訊

Abstract

In this study, we propose an online clustering method to cluster search results of Chinese health information. We use word co-occurrence information to build base clusters, and then the base clusters are merged into final clusters. Experimental results showed that our system constructs larger clusters and covers more web pages, while the precision is not decreased. The result clusters can help users conveniently browsing search results and finding the health information they need.

Keywords: Health Information Retrieval, Search Result Clustering, Word Co-occurrence Information

1、 前言

網際網路的發達改變了人們找尋資訊的習慣，資訊的來源從以前的電視、書籍、報紙擴展到網路。對於各種不同資訊的需求，可經由網際網路的查詢與瀏覽來尋找所需的資訊。在多樣化的訊息中，醫療與健康的資訊是令人重視的，原因是這些資訊與人們的生命健康最息息相關，而透過網際網路的便利性及豐富的資料，使用者可以方便地找尋所要的健康資訊。例如現今社會上有許多的健康與醫療問題，如黑心食品、流感、癌症等等，這些問題都是與使用者切身相關的議題。拜網路發達所賜，關於這些問題的資訊，現在都可以經由網路搜尋引擎來查詢相關資訊。

雖然搜尋引擎的功能不斷地發展，但是目前主流的搜尋引擎的檢索結果，依舊是採用了條列式的呈現方式。而這些資料排序的依據主要是根據搜尋引擎所計算之網頁與使用者輸入之查詢(query)的相關程度，相關度越高的網頁排名就越前面。如果使用者的查詢不是很精確，那些排名高的網頁不一定都是真正與查詢很相關，也很可能包含不同主題的資訊。這些不同的議題可能不是全部都是使用者所關心的，而條列式的呈現方式往往導致不同議題的資訊交互攙雜在一起呈現給使用者，因此使用者必須一筆一筆的瀏覽確認是否是自己所要的資料。條列式的呈現方式增添閱讀搜尋結果的困難度，無形中浪費使用者許多的時間，增加找尋資料的麻煩。

上述的問題若是能透過將搜尋結果進行分群的動作，就可以將不同議題的網頁加以區隔，再呈現給使用者瀏覽。使用者就能夠避開他所不關心的議題，縮小瀏覽的範圍，節省不必要的瀏覽工作。

目前英文語系已經有實際上線的搜尋結果分群系統如 Vivisimo[1] 、iBoogie[2] 等，但這些系統卻不能夠滿足中文使用者的需求，原因是中文與英文的處理方法不盡相同，所以縱然系統能夠涵蓋中文的搜尋結果，

---

[1] http://www.vivisimo.com/

[2] http://www.iboogie.com/

但所呈現的分群結果就十分不理想。因此本研究希望針對中文的醫療與健康資訊搜尋結果來加以自動分群，讓使用者能更快地找到所需的資訊。

2、 相關研究：

將資料分群的基本方法主要可分成兩大類：階層式分群與非階層式分群。階層式分群是在資料集中建立像樹狀結構般的從屬架構，大的群組裡還包含著小的群組。階層式分群分成由下而上的凝聚(agglomerative)方式的分群，與由上而下的分裂(divisive）方式的分群。在非階層式分群裡，群與群之間沒有關係，可以看成是只有一層、平面的架構。通常非階層式的分群方法會重覆進行很多回合，不斷地將資料重新分群，直到分群結果穩定為止。例如 K-means [5]就是一種常用的反覆進行之非階層式分群方法。

當分群的技術應用到分析整理網際網路搜尋結果時，一般文件分群的方法就不見得適用，原因在於網際網路搜尋結果的分群對於時間的要求十分嚴格，往往要在數秒內將所得到的搜尋結果分群完畢。而影響系統運算時間的最大因素往往就在網路的傳輸效能。基於時效性的考量，網際網路檢索結果的分群多直接採用網頁的片段摘要(snippet)，而非把網頁全文擷取回來做分群。

為了加快分群的速度，有許多學者提出網際網路搜尋結果分群方法。Zamir 和 Etzioni[7]使用後綴樹(Suffix Tree)來進行分群，對每一篇文章(即網頁片段摘要)，利用文章內的詞和詞組來建立 Suffix Tree 的資料結構，再透過合併 Suffix Tree 中子樹的方式來對文章做分群。這個方法的缺點是符合的子樹過少所以形成的群可能都不大，甚至不少單獨一篇文章自成一群的情形，這造成不易於進一步合併。因此 Crabtree 等人[4]在 2005 年提出用預先探索的方式來選擇子樹。

Chang 等人[2]在 2005 年也提出用 Topic Keyword Clusters 來做網際網路搜索結果分群的方法，首先找出文件集中具有代表性的關鍵字，然後將所有關鍵字分群形成關鍵字群組，完成後再將文章依分類的方式分至各關鍵字群組中。

Carrot^2[1]為一開放原始碼的分群系統，此系統採用 Lingo[6]演算法進行搜尋結果分群。Lingo 演算法使用 singular value decomposition (SVD)找出文件集中的潛藏概念(latent semantic)，然後找出與潛藏概念最相似的片語或字彙做為群組的標籤。最後計算文章與每個標籤的相似度，如果相似度夠高，則把文章加入此標籤形成的群組中。實驗結果顯示 Carrot^2 效能優於 Suffix Tree Clustering。

3、 搜尋結果分群系統

本研究的目的是將使用者經由搜尋引擎所搜尋而得的醫療健康相關的資料加以分群，呈現給使用者瀏覽。考慮到網路傳輸時間，甚至是原網頁資料已經移除等因素，若採用原網頁全文資料來分群，系統運算時間勢必很長，非一般人所能容忍。因此要將網路搜尋引擎搜尋結果分群時，使用搜尋結果頁裡每個網頁的摘要資訊來分群會比較節省時間。通常在分群時所使用的是文字資訊，也就是網頁的標題及片段摘要(snippet)。但這部份的文字僅有數十字，資訊太少，可能會影響分群的效能，如果能使用額外的資訊來協助分群，或許可以提升分群的效能。

一般而言若只根據片段摘要內的單一詞彙來進行分群，往往會有錯誤發生。原因是詞彙本身可能擁有不同用途、意義，若是在不同場合與不同的詞彙相鄰，同一詞彙所表達的意義可能就會不一樣。為避免上述問題，在本研究裡將採用詞的共現資訊(co-occurrence information)來作為分群時的依據。

系統之流程圖如 Figure 1 所示，主要分成兩部份：一是建立醫學與健康領域共現詞資料庫，二是將搜尋回來的網頁進行分群並輸出結果。

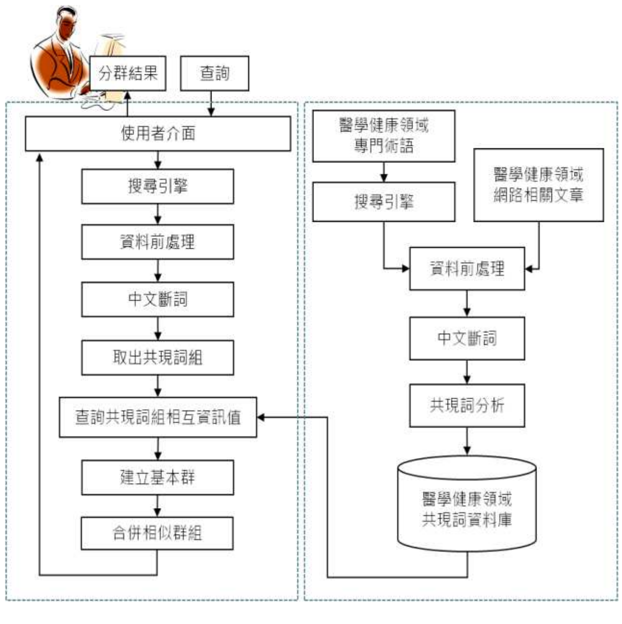

Figure 1 系統流程圖

3.1 醫學與健康領域共現詞資料庫

針對醫學與健康的特定領域，本研究建構一個醫學健康領域詞彙之共現詞資料庫，目的是透過詞與詞之間的相關性來做文件分群。我們從網路上蒐集了約一萬篇的醫療健康相關資訊之報導與文章做為訓練資料。這些網頁文章經過前處理後過濾掉不必要的資料，如html 標籤，網頁廣告等，最後留下網頁文章的本文內容，以 XML 的格式儲存下來。接著我們使用中研院詞庫小組的中文斷詞系統[3]對每篇訓練文章進行斷詞處理，將每個詞區隔開來。

訓練資料在經過整理與斷詞後，接著就要進行共現詞的分析。本研究中對於兩個詞是否具有共現的關係主要是用相互資訊(Mutual Information, MI)[3]來判斷。在計算 MI 時，需要限制資料的共現範圍。通常關聯性比較強的共現詞不會分散在二句話當中，所以在本研究中用一個句子作為計算 MI 時資料的共現範圍。在過濾掉不必要的標點符號、停用詞(stop words)之後，便使用下列公式計算各個詞彙組合的相互資訊值。

$$\mathrm{MI}(x, y) = \log \frac{p(x, y)}{p(x)\,p(y)} \qquad (1)$$

其中 x 與 y 為相異的詞彙，p(x) 表示 x 出現的機率，p(y) 表示 y 出現的機率，p(x, y) 表示 x 和 y 同時出現在一個範圍裏(本研究中設定為一個句子)的機率。當 x 和 y 的 MI 值大於 0 時，表示兩者之間具有正向的相關關係；若 MI 值小於 0 時，表示兩者間具有負向的相關關係；若 MI 值接近 0，表示兩者間沒有關係。所以我們可以設定一個門檻值，當 MI 值超過門檻表示此二字之間共現度夠高。最後把計算所得之詞組間的相互資訊存入資料庫中供分群系統查詢使用。

3.2 網頁分群

分群的步驟分成四個階段來處理：(1)資料前處理 (2)建立基礎群 (3)合併群組 (4)過濾不適當的群。各階段之方法分述於下。

3.2.1 資料前處理

系統先將使用者的查詢送到搜尋引擎進行檢索，接著將搜尋結果頁裡的資料進行前處理，以利後續的分群動作。我們將搜尋結果頁裡的網頁片段摘要加以整理後，送入中研院斷詞系統進行斷詞處理。在過濾掉 html 標籤、標點符號、停用詞之後，以一個網頁片段摘要為單位，將字彙與字彙組合成詞組，查詢共現詞資料庫以獲得該詞組的相互資訊值。如果在共現詞資料庫中查詢不到資料，便給予該詞組一個趨近於 0 的相互資訊值，避免後續計算時出錯。

3.2.2 基本群的建立

於前一步驟所找出之共現詞組可以當成網頁分群的依據，包含有同一個共現詞組的網頁便可以被視為屬於同一個群組。不過由於共現詞組太多，而且相關程度不一，所以要做篩選，保留少數重要的共現詞組。

我們依相互資訊值由高至低將共現詞組進行排序，然後訂定一個相互資訊值的門檻，將高於門檻值的共現詞組留下來。如果這些留下來的共現詞組所出現的文章集合沒有涵蓋全部的檢索結果，則依相互資訊值高低依序往下取，直到全部檢索結果的文章均能被涵蓋為止。這些留下來的共現詞組便各自形成一個小群組，包含有同一個共現詞組的網頁即組成一個基礎群，而共現詞組即是基礎群的標籤。要注意的是此種建立基礎群的方法是允許同一篇文章被歸屬到一個以上的群組。

3.2.3 群組合併

建立基礎群之後要進行群組合併的動作將相似的群組合併在一起，並刪減群組數量。首先找出成員完全相同的基礎群，刪除重複的群組，只保留一份基礎群。接著將留下的基礎群做合併以產生最終群組。合併方法是找出全部群組中文章重疊比率最高的兩個群，若這兩群重疊的部分超過一門檻，則將這兩個群組合併。接著再重複上述步驟，持續合併文章重疊比率最高的兩個群組，直到重疊比率低於門檻值為止。

合併後的群組的標籤則是取用原始群組中權重較高的標籤。若是兩個群組標籤的權重一樣，則將這兩群的標籤合併組成新的標籤。群組標籤權重的計算方式是將共現詞組的相互資訊值乘上群組內文章數量加 1 的對數。

---

[3] http://ckipsvr.iis.sinica.edu.tw/

急性心肌梗塞

- 急性心肌梗塞
  - 冠狀、動脈[39]
  - 壓迫感、下巴、左肩、胸痛[35]
  - 冷汗、冒[20]
  - 行政院、衛生署[15]
  - 緊急、心導管[13]
  - 所謂、指[12]
  - 脫離、險境[11]
  - 人員、醫護[8]
  - 心臟、衰竭[8]
  - 冠狀、心臟病[8]
  - 悶、胸[8]
  - 中廣、新聞網[8]
  - 擴張術、氣球[8]
  - 心絞痛、發作[7]
  - 不整、心律[7]
  - 冠狀、心導管[6]
  - 溶解劑、血栓[6]
    - 溶解劑、血栓[6]
    - 溶解劑、腦出血[3]
    - 溶解劑、心導管[2]
  - 格、部落[6]
  - 動脈、硬化[5]
  - 氧氣、供應[5]
  - 西化、日漸[5]
  - 誘因、冰水、降溫、誘使、冷飲[4]
  - 硝化甘油、舌下、胸痛[4]
  - 打通、心導管[4]
  - 報紙、名人、細心、友人[4]
  - 輕者、李孟儒、重、放射科[4]
  - 內壁、時效、爭取、粥狀[4]
  - 嘔吐、噁心[4]
  - 死、初中、孔明、考試、發作、成績[4]
  - myocardial、infarction[3]
  - 人聲、鼎沸[2]
  - 理財、投資[2]
  - Other[55]

溶解劑、血栓

[17] er_jason 網路經歷: November 2006
振興醫院心臟內科主治醫師任勗龍表示,在急診室中,有7成的
要馬上進行急救,否則會有生命危險。 ... 一旦出現急性心肌
解劑的注射。 ...
http://jyhshengluwebsurf.blogspot.com/2006_11_01_archive

[87] 醫師,為什麼我的心臟血管老是會再度狹窄呢? --談冠狀
的,應屬血栓溶解劑,在急性心肌梗塞發生後12小時內,如果沒
栓溶解劑, ... 隊,對急性心肌梗塞病患施行直接經皮冠狀動脈
今的心導管技術. 產品支架。 ...
http://www.ncku.edu.tw/~clpharm/YHKYang/CV.pdf

[117] www.ncku.edu.tw/~clpharm/YHKYang/劉秉彥醫師上課
藥物治療方面對急性心肌缺氧及心絞痛極有幫助的,應屬血
剝離,腦內腫瘤或腦出血病史者,可以考 慮施予血栓 ... 這樣的
早期數小時內的病患而言,近年來也已證實有相當大的幫助,
http://www.ncku.edu.tw/~clpharm/YHKYang/%BCB%AA%C
1-7.doc

[123] 衛教園地--急性心肌梗塞常見治療用藥
急性心肌梗塞患者的胸痛在前六小時是心肌受損最危險的時
溶解劑. ・urokinase、tpa ... 藥物的治療可以有效的治療急
血脂,並且戒菸、減重及做適當的運動等,是防止心血管疾病
http://w3.tyh.com.tw/htm/epaper/0078/medicalnews2.htm

[128] 大頤事業有限公司
納豆是日本人的傳統食物,有人說,日本人為什麼長壽,「綠茶
族,納豆是將煮熟的黃豆加入一種特殊的發 酵納豆的 ... 已經
便,一般僅止於急性心肌梗塞患者使用,同時投與後的效果也
http://www.dain888.com.tw/health.php

[184] 冠狀動脈疾病
在急性心肌梗塞發生後12小時內,如果沒有內出. 血或動脈剝
動脈內血塊,挽救已梗. 塞,或將塞而尚未塞住的缺血性心肌組
言,近年來也已證實有相當大的幫助,不僅可以縮. 短住院天數
http://nr.stpi.org.tw/ejournal/NSCM/9104/9104-08.pdf

Figure 2 系統分群結果圖

3.2.4 過濾不適當的群

群組在合併完之後，有可能還是會留下很多的群組，或者有些群組的文章很少。為了避免分群結果群組數目過多或是過多零碎的群組，我們將過小且重要性不高的群組濾除，放入一個特定的”其他”群組中。若一個群組內的文章數低於一門檻，則表示這個群很小，當此群組的標籤的相互資訊值也很低時表示此群組就顯得不重要，會被放入”其他”群組中。反之一個群組雖然不大，但是它標籤的相互資訊值夠高，表示這個群仍具有一定的參考價值，可以保留呈現給使用者。

3.3 **分群結果呈現**

搜尋回的網頁分群的結果是用二層樹狀結構顯示給使用者瀏覽。最上面顯示使用者所下的查詢，接下來第一層是分群後的最終群組，第二層則是每一個群組所包含的基礎群。分群結果呈現的介面如 Figure 2，左半邊為給使用者點選的樹狀架構，右半邊則為所點選群組所包含的文章片段摘要，文章是依照搜尋引擎原始排名順序由高至低排列。

4、 **實驗**

4.1 **實驗資料**

為了評估系統的效能，我們收集線上搜尋引擎的搜尋結果加以分群，然後與人工建立之黃金標準做比較，以了解分群的結果是否與使用者的觀點相似。我們參考中華民國行政院衛生署國民健康局網站的熱門查詢排行[4]，從中篩選出 10 個作為實驗之測試查詢。為了測試分群系統在不同搜尋情況下的效能，測試查詢詞含括了不同型態的查詢，例如有查詢範圍較為精確的查詢，如『骨質疏鬆症狀』，或是較為模糊的查詢，如『牙齒』。

為求評估資料的一致性，不會因為搜尋引擎在時間上的差異，導致檢索結果出現變化，本實驗採用封閉型態的測試資料。測試資料的蒐集是將上述的測試查詢，透過 Yahoo![5]搜尋引擎進行搜尋而得，每個查詢均取搜尋結果的前 200 篇網頁作為測試資料。除了搜尋引擎的搜尋結果頁外，我們也將每一筆搜尋結果的原始網頁抓回，以利評估人員在建立黃金標準時參考。

在分群系統的評估標準上，因為每個人對於事物的看法不一定相同，所以每個人認定的分群依據並不一致，導致有多種可能的分群結果，而且並沒有一定哪一種結果是最佳的，可能有好幾種分群方式都是可行的。所以有些分群系統的評估方式是請評估人員查閱分群結果，依主觀意見判定分群的優劣。此種評估方式的缺點是缺少詳細量化的評估，而且對每一種分群結果都要評估人員重新審閱，非常耗費人力。所以本實驗不請評估人員直接審閱分群結果，而是針對每一篇網頁標記其內容之主題，再依據主題來分群。

前述的測試資料收集好後便請評估人員以人工審閱標記的方式來建立黃金標準。評估人員會查看每篇網頁全文，根據網頁內容所講述的重要主題，給予此篇網頁適當的標籤(Tag)，標籤並不限只有一個。每個查詢的搜尋結果均有兩位評估人員進行標記，並且要做討論以取得一致性的最後標記結果。這些標記完成的資料便用來當作分群系統效能評估的黃金標準，標注有同一個標籤的網頁即可視為屬於同一個群組。

4.2 **評估準則**

有了人工建立的黃金標準後，我們便可以使用多種量

---

[4] http://www.bhp.doh.gov.tw/BHPnet/Portal/Hot.aspx

[5] http://www.yahoo.com/

化評估準則來評估分群效能。本研究所採用之評估準則如下所述。

(1) 準確率(Precision)：表示一個系統建立的群組中，有多少比率的文章是真正應該被分在同一群。對於一個系統自動分群的結果群，找出此群中具有最多文章所共有的人工標籤，以作為該群的標準標籤。沒有標注此標準標籤的文章，則被認為不應該被分到這一個群中。若在一個群組中每個人工標籤都相異，則此群不具有標準標籤，且其準確率設為 0。

$$\mathrm{precision}_{C_i} = \frac{\underset{t_j}{\mathrm{Max}}\left|\{d_k \mid t_j \in \mathrm{Tag}_{d_k}, d_k \in C_i\}\right|}{|C_i|} \qquad (2)$$

C_i 是一個群組；d_k 是群組中的一篇文章；Tag_{dk} 是 d_k 的人工標籤集合；t_j 是一個標籤。

(2) 查全率(Recall)：在一個人工標記標籤所形成的群組中，有多少比率的文章是真正被分到同一群組。因為可能有數個系統所分的群組擁有相同的標準標籤，我們對這些群組分別計算查全率後，取其平均值以做為此一標準標籤之查全率。

$$\mathrm{recall}_{ST_i} = \frac{1}{\left|\{C_i \mid st_{C_i} = ST_j\}\right|} * \sum_{C_i, st_{C_i} = ST_j} \frac{\left|\{d_k \mid ST_j \in \mathrm{Tag}_{d_k}, d_k \in C_i,\}\right|}{\left|SC_{ST_j}\right|} \qquad (3)$$

ST_j 是一個標準標籤；SC_{STj} 是由標注有 ST_j 標籤的文章所組成的群組；st_{Ci} 是群組 C_i 的標準標籤。

(3) F_1-Measure：為了不讓準確率與查全率失衡，避免系統為了凸顯準確率而忽略了查全率，導致一個群組內的文章過少，或是凸顯了查全率而忽略了準確率，使得一個群組內文章過多，因此採用了 F_1-Measure 作為評估標準。

$$F_1 = \frac{2 * \mathrm{precision} * \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \qquad (4)$$

(4) 群組分裂數(Split)：平均有多少個系統所分的群組擁有相同的標準標籤，即對應到同一個人工標定的群組。

(5) 群組數量(Group)：一個查詢的搜尋結果平均被分成多少個群組(不包含”其他”群)。

(6) 群組文章數(GroupDoc)：平均一個群組包含多少篇文章(不包含”其他”群)。

(7) 文章遺失數量(Loss)：對某一群組的標準標籤，平均有多少篇文章擁有此標籤，卻沒有被分到一般的群組，而被分到”其他”群組中。

Table 1 分群系統評估結果

分群系統 評估項目	HISC	Carrot^2
Precision	0.331	0.314
Recall	0.538	0.544
F_1-Measure	0.394	0.381
Split	2.797	1.575
Group	35.4	30.8
GroupDoc	7.072	3.614
Loss	1.379	2.553

4.3 實驗結果與分析

實驗評估結果如 Table 1 所示，表中之數值為十個查詢之平均值。HISC 為本研究的系統，作為比較的分群系統為 Carrot^2。Carrot^2 為開放原始碼的分群系統，因 Carrot^2 系統具有良好的效能，並且支援中文文章的分群，恰能作比較的對象。

HISC 的平均準確率及 F1-Measure 都略高於 Carrot^2，不過兩個系統的準確率都不高。觀察群組的分裂數，HISC 會將一個人工標定的群組內的網頁劃分到 2~3 個群組，較 Carrot^2 多出 1 個群左右，這表示系統分的群還不夠集中，而且讓群組數量也偏多。就群組的大小來看，HISC 的群較大，會含括較多的網頁，只有少部分的網頁會被分到”其他”群組中。相較而言，Carrot^2 的群組過小，很多群組只包含 2 篇左右的網頁，而且有不少網頁被放到”other”群中，這對使用者來說可能並沒有多大的幫助，因為很可能要到”other”群中慢慢找尋所需的資訊。

分析系統分群的結果，發現兩個主要的問題會造成準確率不高。第一是僅利用網頁片段摘要來進行分群所能使用的資訊過少，可能無法得知原來網頁內文所討論的主題。當系統要即時將搜尋結果進行分群時，為了避免耗時過久，分群系統是採用搜尋引擎傳回的網頁片段摘要作為分群的依據。這些片段摘要僅顯示出查詢詞附近的字，字數不多，資訊不是很足夠，而且也不一定能代表原來網頁的內容。因此只使用片

段摘要來做分群，很容易造成誤判，使得準確率偏低。若是要分析網頁全文來做分群，則時間上消耗甚巨，這也正是搜尋結果分群系統所遇到的共同難題。

<title>Steering Monthly, Issue 168, Article on Page 21</title>
<snippet>A Chinese christian publication from Overseas Evangelical Mission ... 第一類:心血管疾病,如冠心病、高血壓、心臟病、慢性肺心病、心肌炎、心肌病、風濕性心臟病和先天性心臟病等。 第二類:軀體疾病,如扁桃腺炎、胃十二指腸潰瘍、 道疾病、泌尿 ...</snippet>

Figure 3 文章”心臟早搏當心”的片段摘要

<title>十大死因 唯獨自殺率上竄-生命的華麗與冒險-新浪部落</title>
<snippet>去年國人十大死因依序為惡性腫瘤、腦血管疾病、心臟疾病、糖尿病、事故傷害、肺炎、慢性肝病及肝硬化、腎臟病變、自殺、高血壓性疾病。 ... 連續七年進入十大死因的自殺,去年奪走四千四百零六條人命,平均兩小時左右就有一人自殺身亡。 其中七成是男性,男性自殺身亡人數是女性的二點三四倍。 ...</snippet>

Figure 4 ”十大死因”搜尋結果之一 (片段摘要)

以 Figure 3 為例，該篇實際上是講述「心臟早搏」的文章，但是在片段摘要當中卻沒有”心臟早搏”這個詞彙。所以只利用片段摘要來分群時，HISC 將此篇文章分到”心臟病、高血壓”群組，而 Carrot^2 則是分到”心臟病”群組，造成錯誤之分群。

第二個問題是，即使網頁片段摘要包含了文章的主題，分群系統所找出的群組標籤可能與文章的主題不符，導致分群出現錯誤。以 Figure 4 的例子來說，文章內容是在說十大死因中的「自殺」率上升，但是 HISC 與 Carrot^2 都沒有找出文章主題，HISC 將此篇分至”男性、女性”群組中，而 Carrot^2 則將此篇分至”other”群組中，沒有分到與自殺有關的群組。雖然”男性、女性”是一個相互資訊值夠高的共現詞組，但實際上此共現詞組並不足以代表此篇文章的主題。

5、 結論

在本研究中我們提出了一個搜尋結果分群系統，針對中文的醫療健康資訊搜尋結果來加以自動分群，讓使用者能方便地瀏覽。實驗結果顯示系統的表現還不錯，略優於 Carrot^2。HISC 能產生較大的群組，涵蓋較多的網頁，而且並沒有降低準確率，可以提供較完整的群組給使用者查閱。

由於使用網頁片段摘要來進行分群，資訊過少無法代表網頁全文的內容，使得分群的準確率偏低。但是如果用網頁全文來做分群，又會耗費太多的時間，對線上即時系統來說並不適用。因此如何更有效地使用片段摘要內的資訊，以及在不影響系統效率的情形下整合額外的資訊，是改進自動分群系統效能需要考量的重點。未來我們將使用更多片段摘要內的資訊，例如詞頻、位置等，來協助評估共現詞組的重要性，以利篩選合適的詞組來進行分群。

參考文獻

[1] Carrot^2, http://search.carrot2.org/stable/search

[2] His-Cheng Chang and Chiun-Chieh Hsu, “Using Topic Keyword Clusters for Automatic Document Clustering,” Proceedings of the International conference on information technology and applications, pp. 419-424, 2005.

[3] Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle, “Parsing, Word Associations and Typical Predicate-Argument Relations,” Proceedings of International Workshop on Parsing Technologies, pp. 389-398, 1989.

[4] Daniel Crabtree, Xiaoying Gao, and Peter Andreae, “Improving Web Clustering by Cluster Selection,” Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 172-178, 2005.

[5] James B. MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations,” Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, pp. 281-297, 1967

[6] Stanislaw Osinski and Dawid Weiss, “A Concept – Driven Algorithm for Clustering Search Results”, IEEE Intelligent Systems, Vol. 20(3), pp. 48-54, May, 2005.

[7] Oren Zamir and Oren Etzioni, “Web Document Clustering: A Feasibility Demonstration,” Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval, pp. 46-54, 1998.