斷詞系統對於 Queried keywords 的影響

陳宜惠[1] 呂瑞麟[2*] 黃政傑[2]
[1]亞洲大學資訊多媒體應用系
chenyh@asia.edu.tw
[2]國立中興大學資訊管理學系
[*]jllu@nchu.edu.tw, rogerhaug0729@gmail.com

摘要

部落格在近幾年來蓬勃發展，且讀者已經逐漸習慣從部落格擷取資訊，blogeosphere 也有些改變，從原本部落客只和親朋好友間互動，到現在越來越多的部落客分享資訊，因此根據關鍵字來查詢特定文章為其主要的搜尋方式之一，而取得關鍵字的方法之前得經過斷詞，中文斷詞系統主要使用的為兩種，一種為由中研院所開發的中文斷詞系統(CKIP)，根據中研院所維護的十萬目詞典，以整個句子為單位切成獨立的詞，一種為 MMSEG 斷詞系統，主要透過兩種演算法以及四種規則，並搭配可以自訂的詞庫，並根據斷詞結果帶入 full-text keywords retrieval process(FKRP)來取得文章內容。

我們所開發的一套 Blog Connect 平台，其為透過 blog readers 輸入關鍵字致搜尋引擎，取得相關的部落格資訊是為最普遍之方法，並希望集合眾人查詢某一特定文章所用關鍵字(queried keywords)，來代表該文章的主題。過去的研究成果雖然已經顯示可以利用關鍵字來代表文章的主題以及關鍵字(queried keyword)有斷詞後的成果比未斷詞的成果要來的好許多，但 queried keywords 組成方式有別於一般句子的組成方式，本研究中我們希望進一步利用 queried keywords 以及 full text keywords 並使用 MMSEG Complex Matching Algorithm，來測試 queried keywords 進行斷詞後的辨識率，其實驗結果 MMSEG Complex Matching Algorithm 優於 CKIP 與 MMSEG Simple Matching Algorithm 的斷詞系統。

關鍵詞：MMSEG、CKIP、部落格、queried keyword。

1. 動機與目的

近年來網路科技的普遍，Web 使用者於網路上張貼內容已變得簡單，Web 使用者可以輕易分享他們的經驗與表達他們的想法。而彈性與易維護的特色使得 Blog 變成主要的資訊分享平台[1][2]。根據 blogpulse[3]網站，目前存在於網路上的 Blog 已達 1.7 億多。根據 Technorati's 的報告有 40%的 Blog 讀者同意 Blog 的內容勝過主流媒體並且有 48%的 Bloggers 相信在未來五年內網路使用者會透過 blog 取得知識。而在 Technorati's 2011 Report[4] 內指出在 2010 這樣的現象已有不同，其指出 Blogger 漸漸不在只是與朋友的 Blog 互動往來的趨勢變化，而至 2011 該趨勢更提升至 68%的 Bloggers，其可說明 Blog 不再封閉，越來越多的 bloggers 著重於資訊內容的分享。儘管 Blog 的重要性提升且 Blog 不再封閉但 Blog 仍像是一個孤島[5]，彼此間並無關聯。

由於部落格文章的關鍵字集(keyword set)可以代表該文章的主題[7,8,9]，為了能夠為 Blogs 建立正確的關係或者分群，一般做法是計算 Blog 之間相似度的關鍵字集，其常見 full-text keywords retrieval process(FKRP) 而 FKRP 處理一篇部落格文章 A_i 基本上包括以下幾個步驟：依據網址直接下載部落格文章的原始碼，掃瞄整篇頁面原始碼去除網頁標籤而取得部落格網頁內容($A_{i\text{-}no\_tags}$)，接著移除如部落格邊欄及廣告等其它與部落格本文非相關資料，而取得部落格文章本文(A_{im})，並進行文章斷詞(tokenization or segmentation)，接著是計算每個詞(term)的權重透過權重的排序來選出文章關鍵字，再透過關鍵字相似度計算進而取得文章的相似度，接著利用不同的分類或分群的方法將其分類分群。

另有些研究為利用 Bloggers 常用來分類管理文章的標籤(tags)/分類(categories)，為 bloggers 建立具有相同興趣的部落格社群，但標籤的選定(tagging)是以 bloggers' 角度為出發點，其存在著固有的模糊性(inherent ambiguity)，相同的內容會因不同的 blogger 而有不同的 tags/categories 選擇[11, 12]，因此可能會造成同義詞的問題；例如:中油、中國石油都是相同的意思。近年來因為 Web 2.0 的影響，讓在網路上的需求者也可以成為資訊的提供者，因此興起了一種 Social Tagging 的概念[6]，該做法為集合眾人的力量為網路賞的資源進行 Tagging(標記)，目前常見的網站有:Flicker.com 讓使用者可以替照片貼標籤，del.icio.us 替任何一個網路連結貼標籤，last.fm 替任何一首歌或是歌手貼標籤。

因此結合眾人查詢某一特定文章所用的關鍵字，應該可以代表該文章的主題，而在過去的研究利用 Kendall tau coeficient 與 ma-ratio 來測量 FKRP 與 queried keywords 相似程度，其研究成果可以顯示利用關鍵字代表文章的主題[13]，並且根據[14]研究顯示，斷詞後的 queried keywords 比尚未斷詞後的 queried keywords 其辨識率要來的高出許多。

由於利用 queried keywords 來幫助部落格文章分群或者計算 TF-IDF 來更精確的選擇 queried

keywords 來改善分群的之前，必須都得先經過斷詞[14]，但是由於 queried keywords 的組成方式與一般中文斷詞系統主要以一整個句子為主的斷詞方式，並非完整的相同，在本研究中先利用兩種斷詞系統 CKIP、MMSEG[15,16]來驗證 queried keywords 與 FKRP 之間的辨識率，在本文當中我們根據不同的斷詞系統分別利用了 Jaccard [17,18] 兩種方式進行相似度比較；利用 Blog Connect 平台所收集的三年資料，並比較其成果，MMSEG-Complex Algorithm 的 Jaccard 值高達 0.8628，高於與之比較之斷詞系統(CKIP、MMSEG-Simple Algorithm)。

本文結構如下，第二章為文獻探討，關於不同種的斷詞系統以及斷詞演算法，以及相似度方法的描述及定義，第三章則描述我們蒐集資料、處理的方式，利用改進的 term frequency(tf)來計算每一篇部落格文章的 queried keywords 以及進行的架構，在第四章詳細說明相關實驗以及實驗結果；最後第五章為本篇總結與未來發展與可發展方向。

2. 文獻探討及相關研究

2.1 中文斷詞系統

在自然語言處理上，最基本的處理單位通常是詞，這裡的詞指的是語言學家所定義的「能夠獨立運用，具有完整語意的最小語言成分」。很多自然語言應用的研究，例如:文件檢索、中文輸入、語意辨識等等，都需要先將本文切割，以詞為單位後才能進行後續處理。在英文中，每個單字(word)就可成詞，且都以空白當作字與字之間的符號，因此無須進行斷詞的前置作業。反之中文字詞和詞之間並無空白或特定符號區隔，因此將正確的詞切分出來，就成為自然語言處理的最基礎工作。同樣地，文章的摘要或者文章本身也都是得經過斷詞處理，才進行後續的處理。

鑒於中文斷詞系統為一專精之研究領域，不僅需先經過長期及系統性的蒐集文件才能累積足夠的詞庫或語料庫以進行文件的分析及比對，對於歧義性及未知詞的比對更需利用不同的演算法:如長詞優先、法則式、統計方法等等，才能提高斷詞的正確率，因此在斷詞功能方面採用中央研究院資訊科學小組所開發的中文斷詞系統(ckip paper)以及採用 MMSEG4J 進行斷詞(MMSEG PAPER)，以下分別介紹：

2.1.1 中研院所開發之斷詞系統(CKIP)

中研院資訊所、語言所於民國七十五年成立一個跨所合作的中文計算語言研究小組共同合作建構中文自然語言處理的資源 與研究環境，為國內中文自然語言處理及相關研究提供基本的研究資料與知識架構。中文斷詞系統為其研究之一，其特色為:

2.1.1.1

包含一個約十萬詞的詞彙庫及附加詞類、詞頻、詞類頻率、雙連詞類頻率等資料。分詞依據為此一詞彙庫及重疊詞等夠詞規率及線上辨識的新詞，並解決分歧義問題。

2.1.1.2

採用的「中央研究院平衡語料庫」，是世界上第一個有完整詞類標記的漢語平衡與料庫。1997 年開放的研究院語料庫 3.0 版已達到五百萬目詞的預計規模，目前正朝向一千萬詞的目標邁進。

2.1.2 MMSEG 斷詞系統

MMSEG4j 乃為基於詞典之斷詞方式。MMmseg[15](Tsai,1996) 演算法主要區分為簡單(Simple)與複雜(Complex)兩種方式進行解析，此兩種方式都是使用最大匹配演算法 (maximum matching algorithm)(Chen &Liu,1992)進行處理，其簡單的方式準確率達 95%，而複雜方式準確率達 98%，由於 mmseg4j 之詞庫可以自行擴充，因此可以根據新的詞(未知詞)自行添加，並依此來進行斷詞，分詞規則如下:

2.1.2.1 規則一

最大匹配(Maximum matching) 簡單最大匹配(Simple maximum matching)：以找出資料庫中最長的詞彙為原則。

複雜最大匹配 (Complex maximum matching)：當分割詞彙時，若有歧義的分詞，再往前分析兩個詞彙，以分析三個詞彙為最長的長度為原則，並且取第一個詞彙為最終選擇。

2.1.2.2 規則二

最大平均單詞長度(Largest average word length) 以最大平均單詞長度從 chunk 中取得第一個單詞。

2.1.2.3 規則三

單詞長度的最小方差(Smallest variance of word lengths) 取 chunk 中擁有單詞長度最小方差的作為單詞。

2.1.2.4 規則四

單字單詞的語素自由度的最大和(Largest sum of degree of morphemic freedom of one-character words) 選取 chunk 中擁有最大頻率的第一個詞。

並且 Simple Maximum Matching Algorithm 只會用到最大匹配(Maximum matching) ，Complex Maximum Matching Algorithm 則會運用到四種規則，因此我們將實驗會將 Complex 與 Simple 分別進行實驗。

經過進行斷詞後，為了要確認 queried keywords 的辨識率是否有提升，我們根據[19]中所利用到的 Jaccard 是在經過 TF-IDF 所算出來的 queried keywords 與 FKRP 所算出來的 keywords 來進行驗證，是計算 BCi(K)裡與 FKRPi(K)相符合的個數的比率，即|CKi ∩FKRP(Ai)|/ |CKi|，如果計算結果越高其表示兩者之間的相似程度越高，並且根據相似程度的高低便可判定斷詞系統何者為優。

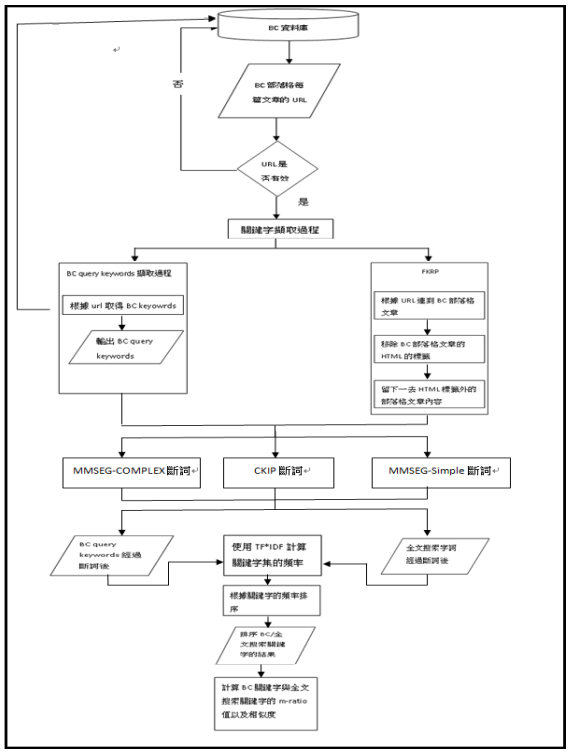

圖 1　實驗流程

3. 驗證方法

為了比較出適合 queried keywords 的斷詞方式，我們透過不同種斷詞方式進行，並比較利用不同斷詞方式所產生的 queried keywords 與 FKRP 之間的相似度，3.1 與 3.2 分別介紹 queried keywords 及 FKRP 的收集方式，圖 1 為整個研究流程。

3.1 Query Keywords

針對取得 queried keywords 的方式，主要是透過使用者在搜尋引擎(google、yahoo)等輸入關鍵字來查詢資料，搜尋引擎會將搜尋結果呈現在網頁上，以此為例當使用者點選部落格文章有掛載 Blog Connect (BC) widget[21]，BC widget 會收集使用者的 queried keywords，並且把資料儲存到 BC 平台的資料庫，圖 2 為利用 BC widget 收集關鍵字流程。

利用蒐集來的 queried keywords 對其進行斷詞的動作，進而觀察在不同的斷詞方式下 queried keywords 與 FKRP 之間的相似程度，下面為處理 queried keywords 利用不同的斷詞方式以及計算相似度的過程。

3.1.1

我們分別利用 mmseg4j[15]分為兩種演算法 Simple Matching Algorithm 與 Complex Matching Algorithm 以及中研院所開發之斷詞系統 CKIP[16]將 BC 資料庫的欄位 keyword 內的 queried keywords 進行中、英文斷詞、移除 stop words。以 Table2 的 queried keywords “中文亂碼”為例，經過不同的斷詞方式分別斷為”中文|亂碼”、”中文|亂|碼”。如 Table1 所示。

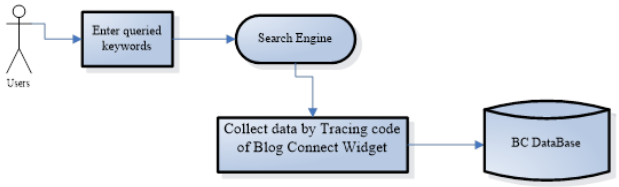

圖 2　利用 BC widget 收集關鍵字流程。

斷詞系統	例子	斷詞結果
CKIP	中文亂碼	中文|亂碼
MMSEG4j	中文亂碼	中文|亂|碼

表格 1　不同斷詞系統斷詞後的 queried keywords

因此可以發現現透過不同的斷詞方式會產生不一樣的結果，經過斷詞處理後會將移除其標點符號。

經過前置處理後，queried keywords 可能會出現在雜訊，例如：”php5.4”經由斷詞後成為”php5”和”4”，為了避免影響相似度的計算時產生雜訊影響，我們透過計算 queried keywords 的 TF-IDF (term frequency-inverse term frequency)[20]給予重要的 queried keywords 較高的權重。TF-IDF 是一種統計方法，用於評估一個 queried keywords 在語料庫(corpus)裡的一份部落格文章所佔的重要程度。傳統 TF、IDF 的定義如下：

$$TF = tf_{i,da} = \frac{n_{i,da}}{\sum_{i=1}^{k} n_{i,da}} \quad (1)$$

TF 的分子 $n_{i,da}$ 為一個 queried keywords i 出現在部落格文章 Ai 的次數，分母為所有 queried keywords 出現在部落格文章 da 的次數總和。

$$IDF - idf_i - \log \frac{|D|}{|\{t \in da : da \in D\}|} \quad (2)$$

IDF 的分子 $|D|$ 表示為在同一個 domain 底下的所有部落格文章總數，分母表示為包含 queried keywords i 的部落格文章總數。

而為了計算是否在同一個 domain 我們採用[13]所述我們將透過專家人工分類的方式，將 BC 所蒐集的部落格文章手動分類，為了配合 queried keywords，將傳統的 TF 定義進行修改，其中 TF 的分子改為在同一篇文章內一個 queried keywords 經過斷詞處理後的出現次數，分母改為在同一篇文章中所有 queried keywords 出現的次數加總，以 Table 2 說明，Ai 的 queried keywords 在斷詞處理前有兩個—分別是”php5.4”和” php5”，且利用”php5.4”查詢到該文章的次數（frequency）為 2，而利用” php5”查詢到該文章的次數為 1。經過斷詞處理過後”php5.4”將會拆成” php5”和”4”，因此” php5”

和"4"的次數都是 2；再加上原來利用"php5"查詢到該文章的次數為 1，最後"php5"的 frequency 為 2+1=3。

Before Segmentation		After Segmentation	
Queried Keywords	Frequency	Queried Keywords	Frequency
php5.4	2	php5	2+1=3
php5	1	4	2

表格 2　計算斷詞後的 queried keywords 的 frequency

最後，若以 Table 2 的" php5"來計算 TF，分子為"php5"的 frequency，也就是 3，而分母為 da 的所有 queried keywords 的 frequency 總和，也就是 2+3=5。

$$TF = tf_{1,da} = \frac{n_{1,da}}{\sum_{t=1}^{k} n_{t,da}} = \frac{3}{5}$$

3.2 FKRP

FKRP 的計算方式為透過[13]裡面所介紹的方式執行，其中修改，透過不同的斷詞方式，來計算出每篇文章的關鍵字集。

3.3 Jaccard

為了驗證 queried keywords 根據不同的斷詞方式(CKIP、MMSEG-Complex、MMSEG-Simple)所斷出來的方是哪種為優，我們採用 jaccard 根據 queried keywords 與 FKRP 之間的相似度，根據其相似度高，則代表該斷詞方式為好，Jaccard 的計算公式如下:

$$Jaccard = \frac{N_i}{T_i} \quad (3)$$

i 代表第 i 篇部落格文章，T_i 為第 i 篇部落格文章的 queried keywords 總數，N_i 為 queried keywords 與 FKRP 所交集的總數。每篇部落格文章都會有 Jaccard 因此為了計算平均，$\overline{Jaccard}$ 的公式如下:

$$\overline{Jaccard} = \frac{\sum_{i=1}^{TN} Jaccard_i}{TN} \quad (4)$$

TN 為在 Blog Connect 中所有的部落格文章總數。

4. 實驗

我們從 BC 資料庫裡的 click_time 欄位取出 2011/07/01~2013/07/01 的資料，總共 913 篇文章，並且根據不同的斷詞方式 MMSEG-Complex、MMSEG-Simple、CKIP 根據[15]所述 Complex Matching Algorithm 會用到四種規則，而 Simple Mathcing Algorithm 則只會用到最大匹配規則，CKIP[16]為中研院開發的中文斷詞系統，分別計算根據同一篇文章而言，queried keywords 與 FKRP 之間的 Jaccard 值，並且設定門檻值(t)設定為 A*i* 的長度的比率，每次累加 10%到 100%，做驗證比較:

根據表格 3 顯示，我們利用 MMSEG4J-Complex 方式所計算出來的 jaccard 值大於其他兩種，因此我們可以判斷為，透過該斷詞方式可以針對辨識 queried keywords 有較好的結果。

segmentation type / (*t* %)	Complex	Simple	CKIP
10	0.5321	0.5343	0.4933
20	0.6220	0.6190	0.5784
30	0.6701	0.6696	0.6319
40	0.6949	0.6954	0.6593
50	0.7229	0.7228	0.6841
60	0.7502	0.7502	0.7125
70	0.7777	0.7778	0.7340
80	0.8005	0.8016	0.7591
90	0.8284	0.8279	0.7861
100	0.8628	0.8626	0.8185
Average	0.7262	0.7261	0.6857

表格 3　實驗結果

我們的實驗結果主要是看斷詞對 queried keywords 的影響，因此我們採取的方法為看其相似值的部分，若相似程度越高，則代表其斷詞效果較好。

5. 總結與未來發展

我們為了找出最適合 queried keywords 的斷詞方式，因此我們使用了不同種的斷詞系統，來比較何種斷詞系統針對 queried keywords 有較好的結果，在我們的實驗當中，MMSEG-Complex 所產生的效果最佳，因此往後的研究再針對斷詞部份，我們將會採用 MMSEG-Complex 的斷詞方式進行，對於未來所要做的部分，我們欲結合 Association Rule，來增進 queried keywords 的辨識率。

參考文獻

[1] J. Gao and W. Lai, "Formal Concept Analysis Based Clustering for Blog Network Visualization," in: Proceedings of International Conference on Advanced Data Mining and Applications, Berlin: Heidelberg, 2010, pp. 394-404.

[2] N. Johnson, 2008, "Google on User Intent in Search Queries, Search Engine Watch,"[Online]Available:http://searchenginewatch.com/article/2053806/Google-On-User-Intent-in-Search-Queries.

[3] J. Sobel, (2010, Nov. 3). "State of the Blogosphere 2010 Introduction.

Technorat,"Available:http://technorati.com/blogging/article/state-of-the-blogosphere-2010-introduction/.

[4]State of Blogosphere, “State of the Blogosphere 2011 : Introduction and Methodology,” Referencing: http://technorati.com/social-media/article/state-of-the-blogosphere-2011-introduction/ (2011, Nov. 4).

[5] Uldis Bojars, John G. Breslin, Vassilios Peristeras, Giovanni Tummarello, and Stefan Decker. Interlinking the Social Web with Semantics. *Journal of IEEE Intelligent Systems* 2008; 23 (3): 29-40.

[6] Lai, V. Rajashekar and C.Rand, W., 2011, "Comparing Social Tags to Microblogs," *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)* , vol., no., pp.1380-1383, 9-11 Oct.

[7] B. Larsen and C. Aone, “Fast and Effective Text Mining Using Linear-time Document Clustering” in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge discovery and Data Mining(KDD ’99), Aug. 1999, pp. 16-22.

[8] K. Ohtsuki, T. Matsuoka, S. Matsunaga, and S. Furui, “Topic extraction with multiple topic-words in broadcast-news speech” in : Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), May. 1998, vol. 1, pp. 329-332.

[9] X. Hu, and B. Wu, “Automatic Keyword Extraction Using Linguistic Features” in: Proceedings of the Sixth IEEE International Conference on Data Mining-Workshop(ICDMW), Dec. 2006, pp. 19-23.

[10] J. L. Elsas, J. Arguello, J. Callan and J. G. Carbonell, “Retrieval and Feedback Models for Blog Feed Search,” in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, July 2008, pp. 347-354.

[11] G. Hope, T. Wang, and S. Barkataki, “Convergence of Web 2.0 and Semantic Web: A Semantic Tagging and Searching System for Creating and Searching Blogs,” in: Proceedings of *IEEE International Conference on Semantic Computing (ICSC)*, Irvine: California, 2007, pp. 201-208.

[12] G. Srinivas, N. Tandon, and V. Varma. “A weighted tag similarity measure based on a collaborative weight model” in: Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents(SMUC ’10), Oct. 2010, pp. 79-86.

[13]M. F. Tsai, ” A Systematic Study of Queried Keywords vs. Full-text Extracted Keywords in Blog Mining” , National Chung Hsing university, Taichung, 2012.

[14] Y. H. Chen, Eric J.- L. Lu , and T.Y. Wu,” A Blog Clustering Approach Based on Queried Keywords”,

[15] Tsai, C.-H.”MMSEG: A Word Identification System for Mandarin Chinese Text” Based on Two Variants of the Maximum Matching
Algorithm ,”http://technology.chtsai.org/mmseg/,2000.03.12.

[16] Chen, K.J. & Wei-Yun Ma,” Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff”, In Proceedings of SIGHAN, pages 168-171

[17] R. Stokes, *Ultimate Guide to Pay-Per-Click Advertising*, Irvine, CA: Entrepreneur Press, 2010.

[18] State of Blogosphere, “State of the Blogosphere 2011: Introduction and Methodology,” Referencing: http://technorati.com/social-media/article/state-of-the-blogosphere-2011-introduction/ (2011, Nov. 4).

[19] Y. H. Chen, Eric J.- L. Lu , and M. F. Tsai,” Using Queried Keywords or Full-text Extracted Keywords in Blog Mining?”,IMECS,2012.

[20] G. Salton and M. J. McGill, Introduction to modern information retrieval, NY, USA: McGraw-Hill, Inc. 1986.

[21] Blog Connect. Referencing, http://bridge.nchu.edu.tw/BC/.