資訊檢索斷詞處理方法

馬德葳 洪朝貴 陳毓璋
樹德科技大學 樹德科技大學 樹德科技大學
資訊工程系 資訊工程系 資訊工程系
研究生 副教授 助理教授
s95739103@mail.student.stu.edu.tw ckhung@mail.stu.edu.tw sclass@mail.stu.edu.tw

摘要

本文主要在探討資訊檢索中所使用的中文斷詞方法，其目的為提供資訊檢索系統一個適用的中文斷詞依據。本方法是一種基於 N-gram 與詞彙法所結合的演算法。其中 N-gram 的部份透過配合詞類組合處理高頻率的虛詞；詞彙法則採用了改良式的正向長詞優先。

這個方法相關的研究及數據資料均於本文中有詳細的說明。同時這個演算法已實作於一個開放原始碼的自由軟體專案 Ozearch 之中，所有程式的原始碼以及字典檔都可以輕易的在 Internet 上取得。

期望本文能對中文資訊檢索的相關研究有所助益，並藉由開放原始碼的架構而使得此方法在未來能夠持續的改良。

關鍵詞：斷詞、資訊檢索、N-gram、詞彙法、Ozearch

Abstract

This paper presents an algorithm for segmenting Chinese phrases. It utilizes both the N-gram algorithm and the word-based algorithm.

Highly frequent characters (stop words) not carrying useful meanings are processed by identifying phrase categories in the N-gram algorithm we use. The word-based algorithm adopts the rule that favors the longest meaningful phrases.

This algorithm is implemented as an open source project "Ozearch" for easy peer review and verification. By releasing its code to the internet under a free license, we expect it to be useful to the research community as well as to the entire society.

Keywords: segmentation, Information retrieval, N-gram, word based, Ozearch

1. 前言

隨著資料量不斷的增加，需要資訊檢索系統的地方越來越多，卻苦於資訊檢索系統通常具有較高的設計門檻，使得國內所自行發展的檢索系統應用於實務上仍寥寥無幾。也因此，中文方面的索引技術仍具有許多可以的改善空間。

本文主要是基於資訊檢索系統中對於中文斷詞方法的探討與改進，並透過這樣的方法開發一套可供使用的資訊檢索斷詞系統。

傳統的斷詞方法一般分為兩大類別，分別是 N-gram 與詞彙法（Word based）。近年來資訊檢索的研究主要以詞彙法為主導，但 N-gram 仍為實務所重視。並且由於本方法是一種結合 N-gram 與詞彙法的斷詞演算法，故在說明本方法前先對這兩大類別做一些介紹。

1.1 N-gram

此方法用於自然語言分析(Natural language processing)時，假設每一個字元皆與前一個字元相關，是一個 N-1 的馬可夫模型(Markov Model)。[2]

當 N 等於 1 時稱為 unigram；等於 2 時則稱為 bigram；3 則為 trigram，大於 3 之後可直接稱為 N-gram，同時由於 N 越大，將使得索引鍵所能形成的組合越多。所以一般認為中文索引使用 unigram 或 bigram 即可提供足夠的檢索成效。[7]

若採用 bigram 對於未知詞的切割會將成語「歲不我與」(出自: 論語陽貨篇) 這樣的字串可分別切割為「歲不」、「不我」、「我與」等索引鍵。

從上例應可看出 N-gram 的切割具有位置的標注、免於維護詞庫等優點，但一個基於 N-gram 所切割的索引鍵也會帶來很多問題。由於 N-gram 基本的假設是每一個字皆與前一個字元相關，但一個有意義的詞彙之前通常會帶

有一個中文的虛詞或英文裡會出現的 stop word。舉例而言，中文的虛詞包括了副詞、介詞、連接詞、助詞、嘆詞與擬聲詞…等。而英文的 stop word 常見的則有 a、an、am、and、are、as…等。

為了解決這個問題，通常 N-gram 會配合詞類組合去除高頻率的虛詞。[3]但去除高頻率虛詞也會帶來一定的影響。如同上例，在 2007 年使用「歲不我與」在 Yahoo 的搜尋結果則表現的相當差。第一頁除了第一筆資料正確，其他全與搜尋標的無關，而需要特別加入雙引號標注才能搜尋到相關的資訊。會發生這種情況應該是虛詞判斷時的條件所形成的問題。

今日許多實務上運作的商業搜尋引擎，幾乎都有採用 N-gram 做為索引鍵的切割方法之一。

1.2 詞彙法

以詞彙為主的未知詞偵測與斷詞方法為國內相關研究主流。其中又有陳稼興等人歸納出三種不同的方式[9]，分別為統計式(statistical methods)、法則式(heuristic rule-based methods)與結合法則式跟統計式兩種方法的混合式(hybrid methods)。

對於未知詞偵測的演算法多仰賴詞庫的正確性與完整性。而近年來已有相當多的研究宣稱有 95%以上的正確率，但仍難免有遺珠之憾。且詞彙法對於未知詞的偵測具有針對性，對於不同的訓練文章會產生截然不同的正確率。舉例而言，極少有人提到未知詞偵測用於切割「古文詩歌」或「方言漢字」時的正確率。

同時近來新興的「火星文」與「注音文」也會帶給未知詞偵測一些新的挑戰。其實就算不是未知詞，詞彙法對於已知詞的切割有時也會有一些不應該會發生的錯誤，舉剛剛用於 N-gram 切割的例子。「歲不我與」在中研院的 CKIP 系統也會依詞類組合而誤斷。

詞彙為主的方法索引鍵相較於 N-gram 而言少的很多，因此查詢時的效能較好。但缺點則是檢索遺漏的可能性高、建立索引所耗費的成本高、詞庫的依賴性高。

1.3 各種方法間的比較

本切割方法主要是基於 N-gram 與詞彙法所提出的綜合方法，主要的目的為使用於資訊檢索系統之上。表 1 為不同切分方法於資訊檢索時所形成之簡易比較表。

表 1 斷詞方法比較表

	N-gram	詞彙法	本方法
索引鍵	多	少	中
檢索遺漏	無	高	少
檢索正確	中	高	高
建立索引	快	慢	中
詞庫	無	重視	不重視

未知詞的判斷一直以來都具有正確率的問題，正確率伴隨而來的不確定性使得資訊檢索系統的索引精準度降低。

這也就是為什麼國內對於未知詞研究多年，而大部分的搜尋引擎卻仍大量使用 N-gram 索引法進行詞彙的切割。但在中文分割中，N-gram 雖然能解決部份的問題，但仍會有索引錯誤的情況存在。不光是因為 N-gram 存在高頻率的虛詞與名詞錯誤組合所帶來的誤判，而是字典字可能包含名詞的情形也會出現在傳統的 N-gram 切割上面。舉例而言，不論是在 Google 或 Gais 查詢「飛狗」，都會得到關於「雞飛狗跳」的查詢結果。而查詢此詞彙的使用者通常希望得到的其實是國內大眾運輸業者的相關資訊。

本文主要是基於資訊檢索系統中對於中文斷詞方法的探討與改進，並透過這樣的方法開發一套可供使用的資訊檢索斷詞系統。

第二節在除了描述方法外，還提出了一些特殊的情況的處理方法。而第三節則是分析本方法切割後的數據，包括實驗設計時所使用的虛詞、字典檔與樣本等。接著是精準度測量及測量過程中所發現的錯誤探討。最後第四節則是結論與未來的研究方向。

2. 斷詞方法

本斷詞法實作採用 N-gram 與詞彙法的特殊結合方法，藉此處理資訊檢索時所面臨的各種狀況。本斷詞法建立於以下的基本假設。分別是：

(1)N-gram 的基本假設「每一個字元皆與前一個字元相關」。

(2)以及從 N-gram 所延伸的假設「詞彙的

第一個字元與詞彙前一個非虛詞且非字典字的字元相關」。

本方法於切割時所進行的程序，包含結合方法的描述；傳統法則式中所使用的方法與改良；虛詞的處理辦法...等。

2.1 先比對詞庫並取得詞彙位置

一連串文字與符號連續的輸入，由符號做為中斷點，並採用正向長詞優先依循序比對辭典。比對詞庫的方法與正向長詞優先的實作方法可以參考蔡志浩先生於 MMSEG[10]的實作，故本文不於此詳述。

2.2 未於詞庫內的字串採用 bigram 切割

有了詞彙的位置之後本方法會與 bigram 進行連結的動作。分別是非詞庫字的連結，預設讓非詞庫字元與前一個字元形成 bigram。若字元前後皆出現符號中斷點或是可判別的虛詞則列為 unigram。而詞庫字的連結，則由詞彙前一個非虛詞的字元與該詞彙的第一個字元連結。其中虛詞的判斷又包含各種虛詞類型。

2.3 關鍵字與詞類組合檢查，去除虛詞所形成的 bigram。

本研究將去除虛詞分為三種不同的情形，根據各種虛詞連結的條件做為斷詞方法所需要參考的標準。

一般情況

根據香港中文大學對現代漢語常用字頻率統計[6]，以 90 年代台灣地區所取樣的文章而言，99.7%的文章裡都會使用「的」這個字，而其中頻率之高，幾乎是每一百個中文字中，「的」這個字就會重複出現四到五次。(頻率為 4.534%)接著是一、是、了、不、在...等，與英文中的 stop word 類似。這些虛詞可以根據一些簡單的規則去除掉。

由虛詞跟名詞搶字的情況

由於發現到，一般情況並不能解決所有的問題。有很多時候是由虛詞組成名詞形成的錯誤。舉例而言{ 和 }多用於連接詞，但也有部份名詞由{ 和 }組成，如: 和服、和室、大和...等。

在碰到名詞當中有虛詞時有一定的機會發生錯誤。如:{ 禮貌和服務 } 在本方法中會被切割成 { 禮貌_和服_服務 }而正確的切法為{ 禮貌_和_服務 }，此時本方法會特別判斷那些包含有連接詞性之虛詞詞彙後者是否為另外一個詞彙。

當然，若後詞彙為未知詞仍有判斷錯誤的可能。就上例而言，此時使用者搜尋{ 和服 }將會有與期望中不符的回應但未知詞於本切割方法中將不受影響而交由 N-gram 處理。

未知詞遇到詞彙包含虛詞的情況

多時候未知詞會本身不含虛詞，而結合後誤切字典字反而會包含虛詞。舉例而言，如:{ 白目的小孩 } 會切割成 { 白目_目的_小孩 } 白目為河洛語漢字且非字典中收錄的名詞，而字典中收錄的是{ 目的 }，但以搜尋字串的效果而言,不論是 { 白目 } 或是 { 白目的 }都可以正確的找到目標，缺點是若搜尋資料為 { 目的 } 也會出現該筆資料。以{ 目的小孩 }做為索引字串可以輕易比較出目前常見的商用搜尋引擎(i.e.: Google，Yahoo，Baidu)對於這個問題的斷詞方法其實也是一致的。
而中研院 CKIP 中文斷詞系統則斷成{ 白(?)(VH) 目的()(Na) 小孩()(Na) } 。

少數能夠正確處理此問題的搜尋引擎 Gais，則直接把詞彙中的「的」濾掉。這個作法雖然可以解決此範例，但其實有點飲鴆止渴的味道。舉例而言，若搜尋「演講的目的」反而得到的是整頁不相關的回應。

2.4 正向長詞優先法(Forward Maximum Matching Algorithm)

本方法在比對詞庫時的做法採用的是長詞優先法。舉例而言 { 作業系統 } 一詞若採用短詞優先而分為 { 作業 系統 } 其意義不如完整的長詞正確。所以一般都採用長詞優先作為斷詞方法。而長詞優先又可分為正向或是反向，依據其比對詞庫的方向而定。

改良式的正向長詞優先

改良式的正向長詞優先指的是對於正向長詞所造成的一些問題而進行規則的套用。舉例而言{ 法務部門 }在意義上為兩個詞彙，分別為{ 法務 }及{ 部門 }，但正向長詞優先容易形成搶字的問題而變成{ 法務部_門 }[6] 。在此方法中，則透過一些簡單的規則來解決搶字的問題。根據前述規則，{ 法務部門 }會被拆成{ 法務部_部門 }其中{ 部門 }為 bigram 結

合的結果，此時若偵測到這個 bigram 為字典字，系統則會向後進行{ 法務 }是否為字典字的判斷。若{ 法務 }加上{ 部門 }兩個字典字的總長度為四，大於{ 法務部 }的長度為三則取用較多詞彙的斷詞方式{ 法務_部門 }。同樣透過改良式的正向長詞優先可處理的例子還有{ 網球拍賣 }容易被斷成{ 網球拍_賣 }、{ 研究生命 }斷成{ 研究生_命 }等。

其實更好的解決方法應該是正向長詞優先做完後再進行反向長詞優先(Backward Maximum Matching)或者是短詞優先，再比對其異同與選擇較優之斷詞方法。但目前尚無需要應用此方法的範例，在未來的研究中應該是可以透過大幅度的比對來尋找需要用到此方法的範例。但目前若納入這樣的作法對於索引鍵建立的成本將大幅提昇，故仍不在考慮範圍之內。

3. 實驗設計與結果

3.1 N-gram 的收斂

一般未知詞偵測會使用正確率(Accuracy)做為其演算方法是否優良的其中一個判斷標準，但其實正確率大多跟訓練語料有著極大的關係。然而使用 N-gram 的切分詞的方法則不存在正確率問題，而是文字間各種不同組合的問題。由於 N-gram 的基本假設是建立於文字是個有限的組合，是一個馬可夫鏈機率問題。因此對於一般未知詞偵測所強調的正確率，於本斷詞方法索引鍵的收斂才是更為重要的。

由於本方法會切割出 unigram 與 bigram 的索引鍵，所以這兩種索引鍵都被視為本研究所需要觀察的對象。

於實驗過程配合 N-gram 所使用的詞庫來源為「新酷音輸入法」當中所包含的 13 萬詞所轉出的詞庫。雖然比起中研院語料庫，其所收錄詞彙量相對少的許多。但其授權為更加開放的 GNU Library General Public License。跟本系統所採用的 GPL 授權一致。相較於中研院語料庫嚴格的授權，使用此詞庫更能將此研究結果利用於實務中不同的領域。

實驗樣本資料為民國 93 至 95 年間，共 57 萬筆的中英文新聞資料。主要集中在以下新聞媒體。分別為「中時報系，法新社，軍聞社，美通社，中央社，中廣新聞網，聯合新聞網…等」。本實驗中採用的虛詞判斷只包括了「的、是、在 跟、和、與、之、得、地、了、嗎、吧」等字，同時依詞性採用了不同的判斷條件。本表採每千筆資料進行一次比對，並於表三中列出前五筆分析資料。

一開始及有相當顯著的重複，數據顯示在第一筆資料新增 50847 個索引鍵後第二筆即可得到 15136 個重複出現的索引。

表 2 本方法前五千筆的分析資料

新聞資料	字典字	N-gram 索引鍵	新增索引鍵	總索引鍵
0~1k	22253	0	50847	73100
1k~2k	23269	15136	39626	78031
2k~3k	22494	20947	31795	75236
3k~4k	22943	24672	28410	76025
4k~5k	22469	27561	24741	74771

到第五筆就已經有 50%以上為重複索引鍵。從圖二可以看出完整的走勢。每千筆新聞的新增索引鍵會逐漸減緩而最後會接近百位數。

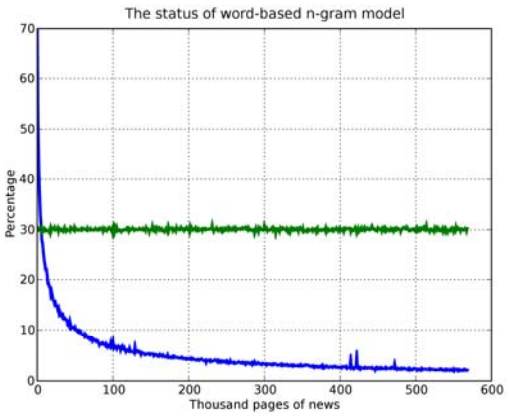

圖 1 索引鍵比例圖

從表 2 可以觀查出每千筆新聞的總量，雖然數字相近但畢竟內容不同。也因此總鍵數也不同. 所以圖 1 採用百分比繪置。藍線表示的是新增的索引鍵所佔總索引鍵的百分比。綠線表示的是於字典擋中的索引鍵所佔總索引鍵的百分比。因為有錯別字，未判別之虛詞和新增的未知詞(於本實驗中，通常是人名)，新增索引鍵要完全達到零的情況並不是本實驗所期待的。而 N-Gram 在龐大的資料量下顯示其收斂的情形相當快速。

為了比較字典檔的效果，本實驗使用 unigram 與 bigram 去除了字典擋後並使用相同虛詞判斷產生出表 3 與圖 2 的資料。

表 3 去除字典檔後前五千筆的分析資料

新聞資料	N-gram 索引鍵	新增索引鍵	總索引鍵
0~1k	0	106384	106384
1k~2k	41985	69817	111802
2k~3k	56035	53715	109750
3k~4k	64356	45868	110224
4k~5k	68932	39004	107936

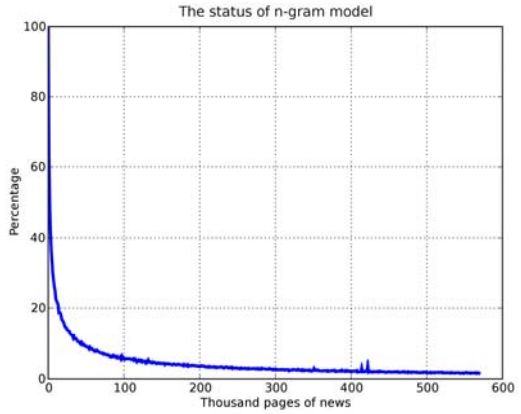

圖 2 N-Gram 索引鍵比例圖

從表 3 及圖 2 可以看出來雖然 N-gram 收斂的效果依然很明顯，但傳統 N-gram 的總索引鍵以及新增索引鍵的量增加太多。這兩種方法對於檢索的資料匯入與取出的效能影響是顯而易見的。

3.2 精準度測量

由於主要是針對斷詞方法對文件檢索的效能做出評估，在一開始我們想採用 F-Measure 做為無排序的搜尋結果(Non-Ranked Search Results)評估方法。F-Measure 是一個由精準度(Precision)與召回率(Recall)所構成的綜合指標。

$$\text{Recall}=\frac{\text{回應資料中有相關的文件}}{\text{所有相關的文件}}$$

$$\text{Precision}=\frac{\text{回應資料中有相關的文件}}{\text{總回應}}$$

F =F-measure
R=Recall
P=Precision

$$F=\frac{2PR}{P+R}$$

其中實驗所遇到的問題點就是"相關的文件"(Relevant documents) ，所謂相關的文件指的包括同義或與搜尋字串本身有相關的所有文件。而本方法是以字串為基礎，尚不具備詞彙關連的查詢能力。同時客觀的召回率驗證程序是相當困難的[8]。重新定義召回率雖然可以將這個問題簡單化，但會失去數字參考的意義。故於本實驗中僅透過人工查詢的方式確任關連後提出表 4 所列出的精準度。

實驗採用了 2007 年 10 至 11 月間的兩萬篇新聞來進行實驗比對。分別包含詞典組合字、未知詞、完整句，並大量使用了部份測試於 CKIP[4]中所斷錯的詞彙。

表 4 Precision

	完整句	詞典組合	未知詞	Total
平均值	0.8931	0.9810	0.9581	0.9574

另外我們也試著透過簡單的 Regular Expressions 進行 Pattern Matching 取得包含查尋字串的文件總數，藉以取得一個近似召回率的值。實驗數據依然有相當好的表現，同時針對以上這些實驗所產生的錯誤進行研究之後，大至上將這些索引所發生的錯誤型態分為三種，分別為：

字典錯誤

例句:{ 行政院所提法務部組織法 }被切成{ 行政院_院所_所提_提法_法務_務部_部組_組織法 }字典字{ 提法 }干擾了{ 法務部 }的斷詞，而形成{提法_法務_務部 }在刪除了{ 提法 }之後，重新切為{ 行政院_院所_所提_提法_法務部_組織法 }同樣的錯誤還有一個句子斷錯兩個字典字的。如：{ 應該是_你是 }會將{ 應該是你是我 } 拆成 { 應該是_你是_是我 }。

詞類組合形成的錯誤.

例句:{ 迷失於金錢遊戲的人耳裡 } 由於詞類組合而將使得本句型後面斷成 { 人耳_

耳裡 }，而非斷為 unigram 的{ 人 }。使得實際用於搜尋{ 迷失於金錢遊戲的人 }得依靠降低 Precision 才能找到正確的目標。

索引切分問題

例句:{ 花蓮縣政府 }其實指的是{ 花蓮_縣政府 } 而正向長詞優先會斷成 { 花蓮縣_政府 } 但使用者其實應該不管是搜尋{ 花蓮 }、{ 花蓮縣 }或{ 縣政府 }都應該要能找到正確的資料。

另外實驗也看到了一些特殊的結合句型，出現 N-gram 能有效處理的狀況，如：{ 民主進步黨籍立委 }。此筆資料其實包含了兩個常用的詞彙，分別可為{ 民主進步黨_黨籍立委 } ，而本方法可正確處理出任一詞彙的資料。

4. 結論

一連串的文字主要的目的在於傳遞訊息，但在應用上則被視為是一種藝術又稱為文學。在研究之初，其實是非常想要找到一個通用的方法來處理斷詞。但文字的規則是如此的複雜，並為 Large-Scale Search Engine 所需，很多極佳的方法與條件得一一取捨。而在這個研究當中提出了一些不一樣的方法，用來強化資訊檢索當中對於中文索引鍵的處理。

雖然國內對於這方面的研究已歷經多年，但是許多方法及中文詞庫都有不同程度上的授權限制。欲將這些寶貴的演算法應用到求技若渴的實務界，不僅因此而受到極大的法律障礙，也造成其他研究人員難以重複其實驗，亦無法對其結論做出獨立客觀的驗證。

本文所介紹的斷詞方法被應用在一個稱為 Ozearch 的開放原碼專案當中，該程式依 GPL 宣告其版權。所有的原始碼以及字典檔都可以輕易的在 Internet 上取得。這同時代表的是在未來這個研究的內容是可以被輕易的改良、發展與驗證的。

在未來的研究與發展方向，虛詞以及其規則判斷的增加應該可以更有效率的降低由虛詞組合而成的索引鍵所出現的頻率。由實驗中可以看出來詞庫可大量降低總索引鍵的數量，而可允許成本範圍內的未知詞偵測應該也是未來在拓展詞庫所必須考量的。

參考文獻

[1] Shafi, S. M., & Rather, R. A.，Precision and Recall of Five Search Engines for Retrieval of Scholarly Information in the Field of Biotechnology.，***Webology***，Vol. 2，2005

[2]Wikipedia，N-gram，http://en.wikipedia.org/wiki/N-gram

[3]王惠，基於組合特徵的漢語名詞詞義消歧，***Computational Linguistics and Chinese Language Processing*** Vol. 7，No. 2，pp. 77-88，August 2002

[4]中央研究院中文計算語言研究小組中文詞知識庫小組，http://godel.iis.sinica.edu.tw/CKIP/

[5]林千翔, 張嘉惠. 基於特製隱藏式馬可夫模型之中文斷詞研究. ***ROCLING XVIII: Conference on Computational Linguistics and Speech Processing***, 2006.

[6]林筱晴、陳信希，語料庫統計值與全球資訊網統計值之比較：以中文斷詞應用為例，***第十六屆自然語言與語音處理研討會***，pp.89-100，2004

[6]香港中文大學，現代漢語常用字頻率統計，http://humanum.arts.cuhk.edu.hk/Lexis/chifreq/

[7]曾元顯、林瑜一，模糊搜尋、相關詞提示與相關詞回饋在 OPAC 系統中的成效評估，***中國圖書館學會會報 61 期***，Vol. 6，pp.103-125，1998

[8]陳光華，資訊檢索系統的評估-NTCIR 會議，***國立台灣大學圖書資訊學系四十週年系慶學術研討會論文集***，pp.67-86，2001

[9]陳稼興、謝佳倫、許芳誠，以遺傳演算法為基礎的中文斷詞研究，***資訊管理研究第二卷第二期***，pp.27-44，2000

[10]蔡志浩，“A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm”，http://technology.chtsai.org/mmseg/