省索引空間之中文全文檢索系統

魏世杰
淡江大學
資訊管理學系
seke@mail.im.tku.edu.tw

于致文
崑山科技大學
資訊管理學系
leo.yu1987@msa.hinet.net

沈英謀
崑山科技大學
資訊管理學系
shenyin@maill.ksu.edu.tw

王建仁
崑山科技大學
資訊管理學系
cjw@mail.ksu.edu.tw

摘要

全文檢索一直是文件查詢者的最愛，因為能毫無遺漏的查到所有內文出現查詢字的文件。中文全文檢索系統受限於字碼及斷詞問題，一直要到 1995 年蓋世引擎的面世才有較快的進展，但是蓋世引擎並不開放原始碼，無從改善功能。目前提供開放原始碼之中文全文檢索系統以 DataparkSearch 表現最佳，但其需要外來資料庫存放索引檔作法較耗硬碟空間。MG 為一提供開放原始碼之英文檢索系統，採用特殊壓縮檔案結構儲存索引資料，不需外來資料庫輔助，能節省硬碟資源，唯其缺點為不能識別中文碼。本文利用 MG 系統原有之索引及查詢能力，為其加上中文識別模組，方便使用者對中文文件製作索引及進行查詢。結果部份將呈現中文化 MG 系統之網頁檢索介面，並比較相同中文文件集下，MG、DataparkSearch、Google Desktop 及 Windows Search 在索引檔大小及編製索引時間上差異。

關鍵詞：開放原始碼、索引壓縮、文件檢索系統

Abstract

Fulltext retrieval is useful as it promises to return all documents containing the query words. Due to the coding and word segmentation issues, Chinese fulltext retrieval system didn't make much progress until the GAIS search engine came out in 1995. But GAIS didn't provide open source. Currently DataparkSearch is the open source fulltext retrieval system which best supports Chinese. But DataparkSearch uses external database for storing index which is not efficient in disk space usage. MG is an open source fulltext retrieval system for English. It adopts a special compressed index structure which saves disk space and uses no external database system. However MG does not support Chinese. This paper aims to provide the Chinese processing functionality to the MG's indexing and query capabilities. As result a web interface for Chinese query will be demonstrated. Furthermore, given a Chinese document dataset, comparison will be made among MG, Google Desktop, Microsoft Windows Search, and DataparkSearch in terms of the index size and the indexing time.

Keywords: Open Source, Index Compression, Document Retrieval System

1. 前言

相對於目錄檢索，全文檢索一直是文件查詢者的終極最愛，因為能毫無遺漏的查到所有內文出現查詢字的文件。但是傳統的檢索系統受限於電腦容量及數位資料尚未普及，多只支援以作者，標題，關鍵詞，出版年份等欄位為對象的目錄檢索，例如常見的圖書館線上公眾存取目錄(online public access catalogue，簡稱 OPAC)系統。

近年來隨著電腦發展，數位資料愈形普及，逐漸有愈多愈多的檢索系統開始提供以摘要，甚至全文內容為對象的全文檢索，其應用領域涵蓋從期刊，專利，法律判例，電子書等專門資料庫，到查詢一般網頁或個人硬碟的搜索引擎等。

中文全文檢索系統受限於中文特有的字碼及斷詞問題，其發展一直比較慢。1995 年蓋世引擎(GAIS, Global Area Information Servers)[1]的出現才引領中文全文檢索進入一個真正實用的階段。蓋世引擎提供各平台執行檔的免費下載，方便資料供應者提供全文檢索服務，但不開放原始碼。

開放原始碼的中文全文檢索系統目前以 DataparkSearch [2]為代表，其採用外部資料庫儲存索引檔的作法，可加速系統的開發，但在節省硬碟索引檔空間上不見得有利。

近年來隨著個人資料快速增加，逐漸出現應用於個人硬碟的中文全文檢索系統，例如

Google Desktop [3]、Microsoft Windows Search [4]等。但是由於未開放原始碼，功能上無從自由改進。

為找尋一開放原始碼，不依賴外來資料庫之中文全文檢索系統，本文嚐試以開放原始碼及壓縮索引檔著稱之 MG 全文檢索系統[5]為基礎，自行增添中文識別模組，以達到節省硬碟索引檔空間之目的。

MG 原名取自 Managing Gigabyte [6]書名之縮寫，號稱適用於大小為 Gb 等級之文件集檢索，其特色如下。

- 支援英文全文檢索，檢索單位可以是段落，自訂分隔標記單元，目錄內文件等。
- 支援原文壓縮及索引壓縮，兩者相加容量往往小於未壓縮原文容量。
- 支援命令列查詢指令，方便輸出入導向及外部呼叫，適合包裝成網頁查詢介面。
- 支援四種查詢模式：文件編號(docnums)，布林(boolean)，向量(ranked)，近似向量(approx-ranked)。
- 支援五種顯示模式：文件編號(docums)，標題(heads)，文件數(count)，靜音(silent)，全文(text)。

本文除了在 MG 上加入中文識別模組，提供中文文件之索引及查詢能力之外，另外將提供一網頁查詢介面，方便使用者輸入中文查詢字串及查看符合文件。最後，就給定之新聞文件資料集，將比較使用 MG、DataparkSearch、GoogleDesktop、Windows Search 在索引檔大小及編製索引時間上差異，以呈現 MG 前處理省硬碟空間及時間程度。

2. 系統架構圖

原始的 MG 檢索系統未支援中文識別能力，不認得中文雙位元字碼中，Ascii 碼為 128 以上字碼。因此，本文將提供轉換代碼模組，將所有中文雙位元組字碼轉換為 127 以下的 16 進位 Ascii 代碼，以利沿用 MG 原有之編製索引及檢索文件功能。

整個完成的中文化 MG 檢索系統架構圖如圖 1 所示，主要呈現編製索引及檢索文件兩部份流程所需程式模組及資料。編製索引時，由管理者將事先準備好的中文 Big5 碼文件集，利用轉換 Ascii 代碼程式，轉換為 Ascii 代碼文件集。再由管理者呼叫 MG 原有之編製索引程式（執行 mgbuild 呼叫 mg_get 及 mg_passes 等程式），產生所有 Ascii 代碼文件索引檔及原文壓縮檔等。

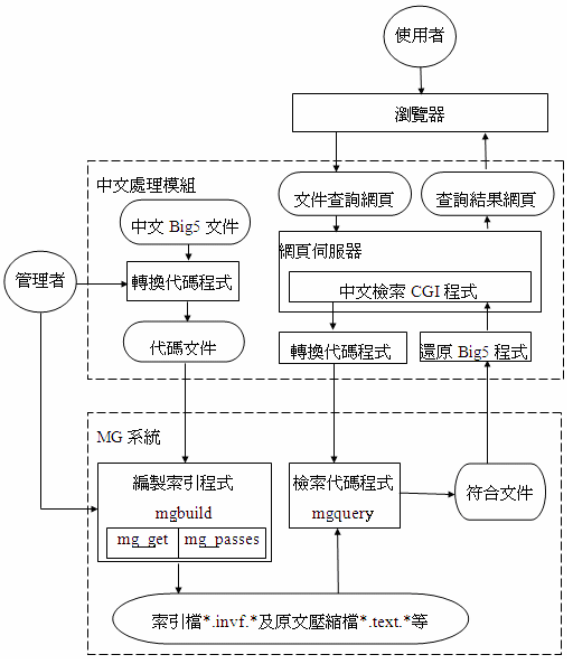

圖 1 中文化 MG 檢索系統架構圖

查詢文件時，則提供文件查詢網頁，方便使用者由瀏覽器輸入查詢句。由網頁伺服器之中文檢索程式，利用轉換 Ascii 代碼程式，將查詢句中文轉成 Ascii 代碼。接著，呼叫 MG 原有之檢索程式(mgquery)，自索引檔中找出符合之 Ascii 代碼文件。再利用還原 Big5 碼程式，將符合文件以原始中文呈現給使用者。

2.1 Big5 碼和 16 進位 Ascii 代碼轉換模組

上述架構圖中多處用到 Big5 碼和 Ascii 代碼兩者間轉換程式。其中，Big5 碼轉換成 Ascii 代碼程式用於管理者編製索引前，將原始中文文件集轉為 Ascii 代碼文件集；及使用者輸入查詢句後，將中文部份轉為 Ascii 代碼。Ascii 代碼還原成 Big5 碼程式則用於符合文件回傳使用者前，將 Ascii 代碼還原成中文供顯示用。以下將針對此轉換原理作介紹。

由於一般繁體中文文件採用 Big5 碼為其內碼，因此能代表其原始中文碼的最方便 Ascii 代碼不外是 Big5 碼本身的 16 進位 Ascii 字串。例如，『由』字的 Big5 碼為 a5d1，其對應的 16 進位 Ascii 代碼即『a5d1』字串。如此，『由』和『a5d1』之間存在一對一的對應關係，可作任一方向之轉換。

要將 Big5 碼轉換成代碼時，只要依照 Big5 標準字碼範圍識別出 Big5 碼，再將其對應的

16 進位 Ascii 代碼輸出成一檔案，即可輕易將 Big5 文件轉換成代碼文件。一個 Big5 內碼由前後高低位元組 BH 及 BL 組成，其有效字碼範圍以 C 語言條件式表示如下：

BH 範圍：BH>=0xA1 && BH<=0xF9
BL 範圍：(BL>=0x40 && BL<=0x7E) ‖
(BL>=0xA1 && BL<=0xFE)

其中，0xhh 表示一個位元組所能表現的 16 進位常數 hh，&&表示布林且(AND)運算子，‖表示布林或(OR)運算子。BH 及 BL 兩者需同時滿足範圍限制，才是有效 Big5 字碼。例如下列這段繁體中文字：

『建管處指出，圓山飯店十二樓的總統套房已恢復原來模樣，原來宴請國賓的大宴會廳四周改成賞景迴廊。』

經過轉換代碼程式後，其內容將變為如下一串 Big5 的 16 進位 Ascii 代碼字串。

『abd8 bade b342 abfc a558 a141 b6ea a473 b6ba a9b1 a451 a447 bcd3 aaba c160 b2ce ae4d a9d0 a477 abec b45f adec a8d3 bcd2 bccb a141 adec a8d3 ae62 bdd0 b0ea bbab aaba a46a ae62 b77c c655 a57c a950 a7ef a6a8 bde0 b4ba b06a b459 a143 』

其中，abd8 為『建』的代碼，bade 為『管』的代碼，b342 為『處』的代碼。以此類推，直到 b459 為『廊』的代碼，a143 為『。』的代碼。特別的是上述轉換過程不需查表，直接擷取中文字內碼，將其 16 位元整數以 16 進位格式（利用 C 語言 printf 函數的%x 格式指令）列印即可。另外，轉換後每一代碼後面都加一個空格，以利代碼之區隔及還原。

要由代碼還原成 Big5 碼時，只要識別出 4 個 16 進位 Ascii 字元加 1 個空格，將其還原（利用 C 語言 scanf 函數的%x 格式指令）成 16 位元整數，再測試其高位元組 BH 及低位元組 BL 是否符合 Big5 有效字碼範圍。符合的話再分別輸出其高低位元組 BH 及 BL，即可還原成 Big5 碼，顯示原來繁體中文字。

2.2 編製索引及檢索文件模組

經過上述 Big5 轉換 Ascii 代碼動作，所有文件字碼都在標準 Ascii 字碼 127 以下。MG 依照原來英文的標點符號及空格斷詞法，即可斷出每一個中文字，進行單字詞編製索引工作。因為目前尚未引入字典檔，故中文索引仍限定以單字詞為對象。

MG 於編製索引時提供以文件內之段落，自訂標記分隔之單元，或目錄內所有文件為索引對象。本文目前採用第三者，即以文件為索引對象，對給定目錄內所有文件編製索引。

另外，MG 於檢索時提供布林及向量排名兩類檢索模式。布林檢索模式只考慮文件內有無查詢字來決定相關度，其相關度非 0 則 1，故同為相關度 1 之文件容易很多，無法細分排名。向量排名檢索模式則考慮文件內查詢字出現之多寡來決定相關度，其相關度為實數值，故同值相關度之文件不易出現，足以細分文件排名。

MG 的布林檢索模式支援由布林且(&)，或(|)，非(!)三運算子及小括號等組成之查詢字串。例如，查詢字串『陽&明』表示想找出有『陽』和『明』皆出現的文件。又查詢字串『(陽|明)&山』表示想找出有『陽』或『明』皆可，但一定要有『山』之文件。因為資源有限，本文後面將只就布林檢索模式作測試。

3. 結果

3.1 測試資料集

本文採用 CIRB030 中文新聞資料集[7]作為測試對象，其組成如表 1。

表 1 中文新聞測試資料集之組成

編號	資料集	名稱	文件數	大小 (Mb)
1	中時晚報	cte1998-1999	5747	10
2	中央日報	cdn1998-1999	27770	41
3	工商時報	ctc1998-1999	25811	45
4	中華日報	chd1998-1999	34728	48
5	中國時報	cts1998-1999	38116	65
6	聯合報系	udn1998	101865	133
7	聯合報系	udn1999	147338	185

3.2 檢索介面

以查詢『陽』字有出現之文件為例，其檢索介面，及回傳結果畫面分別如圖 2 及圖 3 所示。依據 MG 原來功能，目前檢索介面計提供文件編號(docnums)、布林(boolean)、向量排名(ranked)、近似向量排名(approx-ranked)等四種檢索模式。另提供文件編號(docums)、前頭某

給定字數內容(heads)、不擷取全文只顯示文件數(count)、擷取全文只顯示文件數(silent)、全文(text)等五種顯示模式。silent 模式一般只供測試檢索速度用，實用上可由 count 模式取代。

回傳結果時，目前設計為先顯示查詢字串，其統一碼(Unicode)及大五碼(Big5)供確認，然後若為全文顯示模式，再顯示其符合文件之路徑檔案名，及內文。內文以原始文件之行為單位，有出現查詢字之行才作顯示，否則以『.』作省略。行顯示時遇查詢字會特別以醒目之紅色大字體標記，以方便核對。

圖 2 中文化 MG 檢索系統輸入查詢句畫面

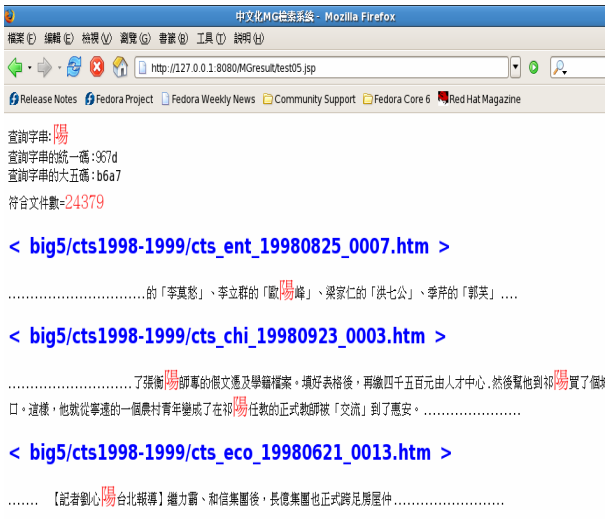

圖 3 中文化 MG 檢索系統回傳結果畫面

3.3 布林檢索

為檢驗 MG 檢索系統中文化之後是否仍保持布林查詢正確性，以下將就二項及三項布林查詢作測試。

3.3.1 二項布林驗證

以『陽』、『明』兩字，測試其交集(&)、聯集(|)四種布林組合之查詢句，得到結果如表 2。其中，**no(Q)**表示查詢句『Q』回傳的文件數。由表 2 可知，兩字之聯集文件數，等於兩個別字文件數相加總和，扣掉兩字之交集文件數。 驗證式子如下。

$$172841=no(陽|明)=no(陽)+no(明)-no(陽\&明)$$
$$=24379+161655-13193$$
$$=172841$$

表 2 『陽』、『明』兩字各種布林組合查詢結果

陽	cts1998-1999/cts_ent_19980825_0007.htm cts1998-1999/cts_chi_19980923_0003.htm cts1998-1999/cts_eco_19980621_0013.htm ...等，共 24,379 份文件
明	cts1998-1999/cts_for_19981221_0006.htm cts1998-1999/cts_soc_19980714_0009.htm cts1998-1999/cts_pol_19981222_0004.htm ...等，共 161,655 份文件
陽 & 明	cts1998-1999/cts_spo_19990421_0008.htm cts1998-1999/cts_soc_19980908_0015.htm cts1998-1999/cts_soc_19990317_0014.htm ...等，共 13,193 份文件
陽 | 明	cts1998-1999/cts_for_19981221_0006.htm cts1998-1999/cts_soc_19980714_0009.htm cts1998-1999/cts_pol_19981222_0004.htm ...等，共 172,841 份文件

3.3.2 三項布林驗證

驗證 1：表 3 為三個字『陽』、『明』、『山』各種組合查詢結果。由表 3 可計算驗證式如下。

$$204031=no(陽|明|山)$$
$$=no(陽)+no(明)+no(山)$$
$$-no(陽\&明)-no(明\&山)-no(山\&陽)$$
$$+no(陽\&明\&山)$$

= 24379 + 161655 + 63232
－13193－29235－7795 + 4988

= 204031

驗證 2：圖 4 為三項集合文氏圖。由圖 4 可計算驗證式如下。

204031 = no(陽|明|山)

= no(陽!明!山) + no(!陽明!山) + no(!陽!明山)
+ no(陽明!山)+ no(!陽明山) + no(陽!明山)
+ no(陽明山)

= 8379 + 124215 + 31190 + 8205 + 24247
+ 2807 + 4988

= 204031

驗證 3：由圖 4 可計算驗證式如下。

32042= no((陽|明)&山)

= no(陽!明山) + no(!陽明山) + no(陽明山)

=2807 + 24247 + 4988

=32042

驗證 4：由圖 4 可計算驗證式如下。

140799= no((陽|明) & !山)

= no(陽!明!山) + no(!陽明!山) + no(陽明!山)

= 8379 + 124215 + 8205

=140799

以上驗證式中，查詢字串『陽明』和『陽&明』意義相同，因為布林查詢中『&』可省略；查詢字串『!陽明!山』和『(!陽) &明&(!山)』意義相同，因為『!』順位高於『&』順位。由以上結果可知，中文化 MG 檢索系統仍具有布林查詢正確性。

表 3 『陽』、『明』、『山』三字各種布林組合查詢結果

查詢字串	符合文件數
陽	24,379
明	161,655
山	63,232
陽&明	13,193
明&山	29,235
山&陽	7,795
陽&明&山	4,988
陽|明|山	204,031

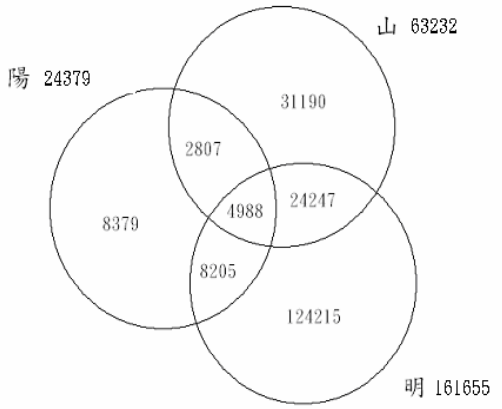

圖 4 『陽』、『明』、『山』出現文件集合文氏圖，其中數字代表各組合下之文件數。

3.4 索引量及索引時間之比較

為呈現 MG 檢索系統使用專屬資料庫的優點，本文將其和另外一個使用外來資料庫的 DataparkSearch (DP)檢索系統、及另兩個同樣使用專屬資料庫的 Google Desktop (GD) 、Windows Search (WS) 作比較，其中 DataparkSearch 的外來資料庫使用常見的 MySQL 資料庫。比較將分為索引量及索引時間兩部份。由於 MG 系統編製索引時，除了對索引檔(*.invf.*)壓縮，還對本文檔(*.text.*)壓縮，本文計算索引量時將只針對索引檔大小作統計。另外，為了解隨著原始文件集變大之下，各檢索系統之索引量及索引時間差異，本

文自表 1 依量小到量大，單獨抽取 5 種（編號 1、2、5、6、7）文件集做比較。

本文測試所使用硬體設備為 CPU Pentium 4@3.0GHz、RAM 512Mb 之個人電腦。表 4 為使用軟體的版本。測試結果方面，表 5 及圖 5 為中文化 MG、DataparkSearch、Google Desktop、Windows Search 編製索引後佔用硬碟空間之索引量比較。由結果可知，表現由好到差依序為 MG、DataparkSearch、Windows Search、Google Desktop。其中，DataparkSearch、Windows Search、Google Desktop 編製索引後佔用硬碟空間的索引量約為本 MG 系統之 2.4~49 倍以上。( ) 內數字為索引量與原始資料量之百分比，MG 索引量則約為原始資料量之 13~19%。

表 6、圖 6 為中文化 MG、DataparkSearch、Google Desktop、Windows Search 編製索引所花費時間之比較。由結果可知，表現由好到差依序為 MG、Windows Search、DataparkSearch、Google Desktop。其中，DataparkSearch、Windows Search、Google Desktop 系統之編製索引所花費時間約為本 MG 系統之 1.8~91 倍以上。

由以上結果可確認，本中文化 MG 系統在索引量及索引時間上都比 DataparkSearch、Google Desktop、Windows Search 優異。

表 4 本文測試所使用各軟體版本

軟體	版本
Linux OS	Fedora Core 6
Google Desktop	5.8.0809.08522
MG	1.2.1
DataparkSearch	4.51
MySQL	5.0.51a
Windows OS	Windows XP
Windows Search	4.0

表 5 編製索引後佔用硬碟空間的索引量(Mb)比較

資料集	cte1998-1999	cdn1998--1999	cts1998-1999	udn1998	udn1999
原大小	10Mb	41Mb	65Mb	133Mb	185Mb
MG	1.9 (19%)	5.8 (14%)	8.8 (13%)	21.2 (16%)	28.9 (16%)
DP	5.6 (57%)	26.1 (63%)	28.2 (43%)	51 (38%)	136.9 (74%)
WS	13.0 (133%)	46.3 (112%)	71.2 (109%)	156.0 (117%)	280.0 (151%)
GD	58 (596%)	256 (619%)	435 (666%)	710 (532%)	1100 (594%)
DP/MG	2.9	4.5	4.1	2.4	4.7
WS/MG	6.8	8.0	8.1	7.4	9.7
GD/MG	30.7	44.1	49.5	33.5	38.1

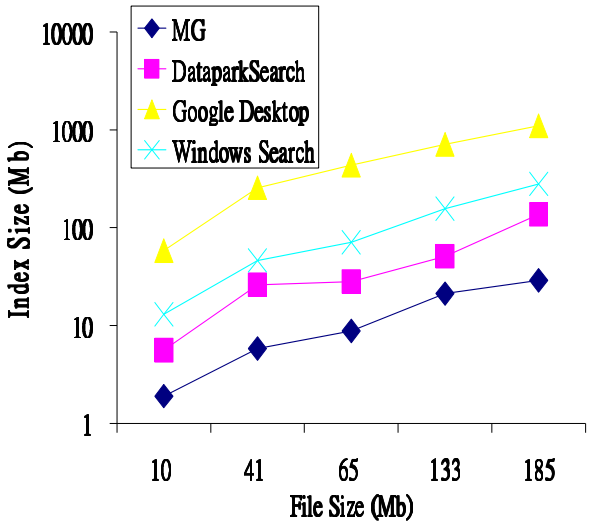

圖 5 編製索引後佔用硬碟空間索引量比較圖

表 6　編製索引花費時間(Min)比較

資料集	cte1998-1999	cdn1998-1999	cts1998-1999	udn1998	udn1999
原大小	10Mb	41Mb	65Mb	133Mb	185Mb
MG	0.7	3.9	11.2	34.6	55.5
DP	43	291	347	748	1112
WS	4.5	15.5	20.7	63.3	106
GD	64	382	600	2400	4200
DP/MG	62	75	31	22	20
WS/MG	6.4	4.0	1.8	1.8	1.9
GD/MG	91	84	54	69	76

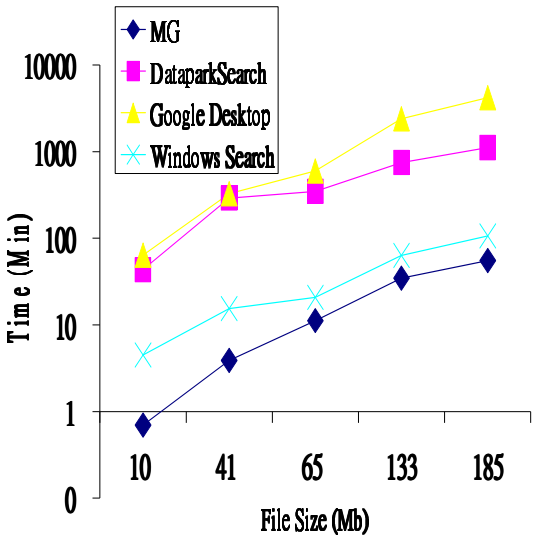

圖 6 編製索引花費時間的比較圖

4. 結語

MG 為一英文全文檢索系統，其特色為開放原始碼及支援索引壓縮。為方便開發及省索引空間，本文遂以 MG 為基礎，為其添加中文處理能力，俾提供一實用之中文全文檢索系統。

由實驗測試結果可知，本文之中文化作法仍維持原來 MG 檢索系統之布林查詢正確性。另外，就索引量而言，表現由好到差依序為 MG、DataparkSearch、Windows Search、Google Desktop。就索引時間而言，表現由好到差依序為 MG、Windows Search、DataparkSearch、Google Desktop。可見中文化 MG 系統採用檢索專用之反轉索引表，並針對索引作壓縮之作法，和 DataparkSearch 系統採用一般資料庫儲存反轉索引表比較，的確在索引量及索引時間上皆表現優異。即使和採用專用索引資料庫的 Google Desktop 及 Windows Search 相比，MG 系統在索引量及索引時間上也都毫不遜色。

由於 MG 為一開放原始碼之 GPL 軟體，在為其加上中文處理能力之後，將為開放原始碼之中文檢索軟體世界，提供一索引量小（約原始文件集容量之 14~19%），索引時間快之優質檢索元件，未來極適合套用在各種文件檢索應用上。

本系統之下一階段目標為測試查詢回應時間，處理中英文夾雜文件之索引及查詢，導入字典提供片語查詢(phrase query)，及查詢結果作向量排名測試等。

參考文獻

[1] GAIS2.0, ftp://ftp.ccu.edu.tw/pub2/packages/gais/gais20,1995.

[2] DataparkSearch, http://www.dataparksearch.org/, 2008.

[3] Google Desktop, http://desktop.google.com/zh/TW/, 2008.

[4] Microsoft Windows Search, http://www.microsoft.com/windows/products/winfamily/desktopsearch/default.mspx, 2008.

[5] The MG system, http://www.nzdl.org/html/mg.html, 1999.

[6] I.H.Witten, A.Moffat, and T.C.Bell, Managing Gigabytes - compressing and indexing documents and images, Morgan Kaufmann, 1999.

[7] 中文資訊檢索標竿測試集第三版(CIRB030)-文件集，中華民國計算語言學學會，http://www.aclclp.org.tw/use_cir_c.php，2004.