誌謝

感謝指導教授曾元顯老師細心指導，給予許多建議，並讓我有機會見識國際會議(NTCIR)的學術環境，更加充實研究所的知識。能夠成爲碩士生；在研究所學習到充實的知識，須要特別感謝曾老師的指導栽培。

感謝林頌堅老師不辭勞苦趕到輔大爲學生評審口試論文，並點出許多學生未發現的問題，讓學生能更加了解研究核心；感謝張淳淳老師，對於學生詞不達意的文句能耐心訂正，並適時給予學生許多建議，再次感謝兩位評審委員。

再來要感謝隆基助教，能提供學生研究的環境，才能在完成多次馬拉松式的實驗，並適時協助解決問題；靜宜助教能在許多行政程序上有效的協助，節省學生不少時間，小童助教也幫忙不少。三位助教讓系辦成爲我們聊天散心的好地方。

感謝熱心的炳魁、莉萱、小頭、陳威、宏偉、維哲、宣宇等…同學們大力協助，最重要的要感謝學長育欽，不厭其煩，教導許多研究細部流程，才能使我更快進入情況，再次感謝。

最後，感謝父母給予機會讓我修習完成學位，及女朋友芸嬋的陪伴。

謝謝

摘要

從過去 NTCIR3、4、5 屆的會議論文中，分析出能提升的檢索效能的機制，假設在，「最佳效能的參賽團隊提供的機制能有效的提升檢索成效」。日本 NTCIR 會議學術交流環境之下，參賽團隊會將檢索機制說明於會議文件之中，即便說明不清楚，也可能於相關技術文章之中討論。本研究目的，在於有效的分析 NTCIR 跨語言檢索任務的最佳技術文件，從中取得該團隊的技術，供本研究實驗驗證。

分析實驗系統、各優秀團隊提出的機制之後，本研究最終篩選出「I2R 文件二次排序」、「HKPU 文件標題排序」、「Label Propagation 歸類二次排序」，並且和「PRF 查詢擴展」比較。實驗結果發現文件二次排序的方式都無法提升成效，而 PRF 查詢擴展能有效穩定的提升成效。

關鍵字：Document Reranking, NTCIR, CLIR, PFR, Label Propagation, 資訊檢索, 文件排序

Abstract

From the experience gained from participating in the past NTCIR workshops, we learn that the major factors that affect retrieval effectiveness are: indexing schemes, retrieval models, query expansion techniques, and document re-ranking methods.

We compared I2R document re-ranking, HKPU title re-ranking, label propagation, K-nearest neighboring, and pseudo relevance feedback for document re-ranking and found that pseudo relevance feedback is a more robust technique for performance improvement, while label propagation and K-nearest neighboring are sensitive to the choice and the number of relevant documents for successful document re-ranking. I2R document re-ranking and HKPU title re-ranking cannot improve performance.

iii

目次

誌謝 ..... **i**

摘要 ..... **ii**

目次 ..... **iv**

第一章 緒論 **1**

第一節 研究背景與動機 1

第二節 研究目的 2

第三節 研究貢獻 3

第四節 名詞解釋 4

第五節 研究限制 8

第二章 文獻探討 **9**

第一節 日本 NTCIR 檢索會議 9

第二節 FJUIR 於 NTCIR 使用的檢索機制 20

第三節 文件排序研究 31

第三章 研究方法與設計 **44**

第一節 研究方法 44

第二節 研究流程與架構 46

第三節 研究設計 50

第四章 實驗結果與分析 **59**

第一節 以查詢詞特徵調整相似度 60

第二節 以初次查詢排序調整相似度 66

第三節 綜合評估 79

第五章 結論與建議 **85**

第一節 結論 85

第二節 建議 86

參考文獻 **87**

附錄 A　NTCIR 5 CLIR 中文查詢問題.....91

表 目 錄

表 2-1： NTCIR 6 重要日期.....9
表 2-2： NTCIR 各屆測試文件集.....11
表 2-3： NTCIR5 所使用文件集.....13
表 2-4： 答案集相關判斷層級.....15
表 2-5： 答案集的範例.....16
表 2-6： 精確率、回收率和雜訊比之２乘２表格.....16
表 2-7： NTCIR4 全域擴展及局域擴展的實驗成績.....28
表 2-8： NTCIR3-5 屆 CLIR 任務使用 Re-ranking 技術的團隊.....33
表 2-9： I2R 團隊於 NTCIR5 之成績.....37
表 3-1： CIRB040r 文件集收錄資料來源.....50
表 4-1： 基礎檢索系統平均精確度.....59
表 4-2： I2R 文件二次排序實驗結果.....60
表 4-3： HKPU 標題二次排序實驗結果.....63
表 4-4： Label Propagation 歸類二次排序實驗結果.....66
表 4-5： KNN 歸類文件二次排序實驗結果.....68
表 4-6： KNN 及 Label Propagationt 查詢題型比較(Rigid).....70
表 4-7： 歸類一起文件非同爲相關文件.....70
表 4-8： 和判斷文件相似度高但被判斷爲非相關文件之內容.....71
表 4-9： PRF 查詢擴展實驗結果.....72
表 4-10： 前 6 篇 x 個最佳詞與 MAP 分佈情況.....74
表 4-11： 最佳 15 個詞於前 X 篇文件 MAP 成效.....77
表 4-12： NTCIR6 文件集 PRF 穩定成長情況.....78
表 4-13： 本實驗研究數據表.....79
表 4-14： 各機制送入判斷正確文件數比較表.....82

v

圖 目 錄

圖 2-1：測試文件集資料標示.....12
圖 2-2：NTCIR5 題目範例.....14
圖 2-3：精確率及回收率關係.....17
圖 2-4：TREC_EVAL 執行結果範例畫面.....18
圖 2-5：FJUIR 於 NTCIR5(CLIR)檢索流程圖.....20
圖 2-6：關鍵詞擷取演算法.....21
圖 2-7：關鍵詞演算法執行流程.....22
圖 2-8：反向索引檔(inverted file)範例.....22
圖 2-9：向量模式示意圖.....23
圖 2-10：FJUIR 歷年檢索技術方向.....29
圖 2-11：I2R 團隊檢索系統流程圖.....34
圖 2-12：文件二次排序演算法.....36
圖 2-13：NTCIR4 查詢問題第二題.....38
圖 2-14：NTCIR4 查詢問題第 13 題.....38
圖 2-15：Label Propagation 歸類調整文件排序.....41
圖 2-16：Label Propagation 計算情況示意圖.....42
圖 2-17：KNN 及 Label Propagation 歸類方式示意圖.....43
圖 3-1：研究流程圖.....46
圖 3-2：控制組檢索系統.....51
圖 3-3：I2R 文件二次排序架構圖.....53
圖 3-4：標題文件二次排序架構圖.....54
圖 3-5：Label Propagation 歸類演算法.....55
圖 3-6：Label Propagation 歸類模組進行文件二次排序.....55
圖 3-7：KNN 歸類演算法.....56
圖 3-8：KNN 歸類模組進行文件二次排序.....56

圖 3-9：PRF 查詢擴展實驗架構圖 57
圖 3-10：檢索系統查詢結果.....58
圖 4-1：文件二次排序各題查詢問題成效提升(imp)圖(Description-run rigid) 61
圖 4-2：文件二次排序各題查詢問題成效提升(imp)圖(Title-run rigid) 61
圖 4-3：文件排序 f(i)對權重的影響(假設該詞文件數爲 1000) 62
圖 4-4：DF(t,C)對權重的影響(假設排序於第 500 篇文件).....62
圖 4-5：標題二次排序各題查詢問題成效提升(imp)圖(Description-run/rigid) 64
圖 4-6：標題二次排序各題查詢問題成效提升(imp)圖(Title-run/rigid) 64
圖 4-7：Label Propagation 歸類二次排序各題查詢問題成效提升(imp)圖 (Description-run/rigid) 67
圖 4-8： Label Propagation 歸類二次排序各題查詢問題成效提升(imp)圖 (Title-run/rigid) 67
圖 4-9：KNN 歸類文件二次排序各題查詢問題成效提升(imp)圖 (Description-run/rigid) 69
圖 4-10：KNN 歸類文件二次排序各題查詢問題成效提升(imp)圖 (Title-run/rigid) 69
圖 4-11：PRF 查詢擴展各題查詢問題成效提升(imp)圖(Description-run/rigid) 73
圖 4-12：PRF 查詢擴展各題查詢問題成效提升(imp)圖(Title-run/rigid) ..73
圖 4-13：前 6 篇 x 個最佳詞與 MAP (Relax) 75
圖 4-14：成長度(imp)與回饋詞數 76
圖 4-15：最佳 15 個詞於前 X 篇文件 MAP(Relax)分佈圖 77
圖 4-16：Title run 的 imp 長條圖 80
圖 4-17：Description run 的 imp 長條圖 81
圖 4-18：送入正確文件數提升成效(Title-run/rigid).....83
圖 4-19：送入正確文件數提升成效(Description-run/rigid) 83

vii

第一章 緒論

第一節 研究背景與動機

資訊系統爲提供使用者解決問題所需資訊的系統，要設計好的資訊系統，則必須充分掌握使用者的問題和解答問題所需的資訊(黃慕萱，1996)，資訊使用者所下達的查詢詞彙，不一定出現於所需文件之中；此外，使用者也會因爲外部環境、地位、知識等…不同，其需要表達的查詢詞彙也會因此不同(蔡育欽，2005)。

爲了幫助使用者滿足資訊需求，資訊檢索技術因而發展。在美國 TREC、日本 NTCIR 等…皆提供了一個完善的實驗環境供學者評估其開發的資訊檢索系統（陳光華，2004），該項國際性的評比成爲了解自己發展的檢索系統優缺點，各個國家團隊在此環境上交流研究心得，以精進自身發展的系統，更推動資訊檢索技術上的發展。

FJUIR(Fu Jen Catholic University, Information Retrieval Laboratory)團隊，參加了 NTCIR3、4、5 屆會議跨語言檢索任務的單語檢索的部分，由於檢索技術發展快速，因此本研究希望能整理過去三屆成效最佳參賽團隊技術，了解目前高成效之檢索系統運作機制，有效實作於 FJUIR 團隊的檢索系統之中，實驗是否能因此提高該系統效能，以獲得最佳的檢索結果。

第二節 研究目的

爲了分析出能有效提升成效的機制，將從 NTCIR3、4、5 屆的優秀團隊會議論文爲主要的研究文獻，分析成效較佳的參賽團隊技術，也試著藉此取得相關經驗及研究，實驗這些技術，希望能有效提升檢索效能。研究目的大致分成二個階段。

第一階段，從過去 NTCIR3、4、5 屆的會議論文中，分析出能提升的檢索效能的機制，假設在，「最佳效能的參賽團隊提供的機制能有效的提升檢索成效」。NTCIR 會議學術交流環境之下，參賽團隊會將檢索機制說明於會議文件之中，即便說明不清楚，也可能於相關技術文章之中討論。第一階段的研究目的，在於有效的分析 NTCIR 跨語言檢索任務的最佳技術文件，從中取得該團隊的技術，供研究參考。

第二階段，在完成第一階段目的之後，本研究將盡可能的將第一階段所分析出的技術運用於「FJUIR 團隊」的檢索系統之中，並且將實作後的系統，再模擬於 NTCIR5 的跨語言檢索任務，了解成效提升情況，並且以該項最佳分數爲目標。

本研究的目的如下：

1. 分析 NTCIR 中 CLIR 任務的技術文件，研究提升主題檢索成效之機制
2. 實作「分析所得的檢索機制」以提升主題檢索成效

第三節 研究貢獻

NTCIR 檢索會議，是一個讓參賽者交換檢索技術及心得的環境(TREC, 2004)，因此會有許多成效佳的參賽團隊，將自身的經驗及技術提供出來，供其他團隊參考，有許多的機制，值得去研究並實作，更進一步的深入探討。本研究藉由分析 NTCIR3、4、5 屆的會議論文，分析出最佳檢索系統的機制，希望能提供技術資訊供相關研究者參考、技術交流或研究；或運用本研究中所提及的技術，解決其他研究領域的問題，例如：「圖書館知識組織」、「資料探勘」、「企業資訊管理」等…，提升自動化「資訊處理」能力，進而協助使用者組織廣大的資訊。

過去參加 NTCIR3、4、5 屆的檢索任務中，FJUIR 成績未盡理想，因此參考其他參賽團隊的研究技術及經驗以提升檢索成效，更進一步的驗證機制，確實將該機制實作，實驗於 NTCIR5 的文件集中，了解該實驗系統是否能提升成效。假設研究成功，成效確實提升，在此爲基礎之下，運用更佳的檢索系統，有助於發展更佳的檢索機制，希望能在 NTCIR 之後幾年的參賽中，取得更好的成績。

第四節 名詞解釋

一、TREC 檢索會議

爲了促進資訊檢索的研究與發展，美國國家標準暨技術局(National Institute of Standards and Technology, 簡稱 NIST)與美國國防部高等研究計劃局 (Defense Advanced Research Projects Agency, 簡稱 DARPA)共同舉辦了文件檢索會議(Text REtrieval Conference, 簡稱 TREC)，爲 TIPSTER 計劃一部分，其目的在於支援資訊檢索的研究，提供一個可評估文件檢索的環境及方式；制定各種測試項目、測試程序及測量準則、組合成一評估檢索系統的機制(江玉婷、陳光華，1998)，以模擬大型真實的資訊環境。TREC 的主要目標有下(TREC, 2004)：

(1) 以大型測試集爲基礎，鼓勵文件檢索的研究。

(2) 經由開放式的論壇，使與會者能交換研究的成果與心得，以增進學術界、產業界與政府的交流互通。

(3) 經由對真實檢索環境的模擬與實質的論證，加速將實驗室研究技術轉至商業化產品。

(4) 發展適當且具應用性的評估技術，或使評估技術更適合當今系統，供產業及學術界採用。

TREC 在 1992 年舉辦了第一屆，至 2007 年底已進行了十六屆。在 2003 年共有來自 22 個國家 93 個團隊參加。除了與會者依據大會提供的測試集送回各測試項目的資料進行評估之外，尚有一個爲期三天的研討會，與會者可以在會中發表系統的架構，評估結果，並相互討論切磋。現今參與 TREC 的檢索系統，相較於前六年，成效大約提升二倍左右(TREC, 2004)，由此可知，TREC 會議提供的環境，對於研究者發展評估自身的檢索系統是有正向的幫助。

以下爲 TREC 2006 年的任務簡述：(TERC, 2007)

(1) Blog Track：TREC 2006 年新增項目，2007 年繼續舉辦。其目的在於從部落格環境中發覺使用者「資訊尋求(information seeking)行爲」。

(2) Enterprise Track：其目的在於研究企業搜尋，滿足使用者搜尋一個組織的資訊以完成某些任務。

(3) Genomics Track：提供一個評估生物基因(genomics)領域檢索系統的討論會議。

(4) Legal Track：TREC 2006 年新增項目，2007 年繼續舉辦。其目的在於去發展一種搜尋方式，讓法律人士能在「數位文件集」(digital document collections)中有效率的探索文件。

(5) Question Answering Track：其設計於提供一個讓「文件檢索」更接近於「資訊檢索」。

(6) Spam Track：對現今的垃圾郵件過濾方式提供一個評估標準，因此是讓許多普遍的信件過濾器和相關檢索任務有一項評估基礎。

(7) Million Query Track：2007 年新增項目，其目的於測試一個假設性問題；從許多不完全判斷(incompletely judged)的 topic 所建立的測試文件集，會比使用傳統 pooling 所組成的文件集要好。

二、NTCIR 檢索會議

NTCIR（NII -NACSIS Test Collection for Information Retrieval）會議，由 MEXT 資訊學領域科學研究補助金及日本國立情報學資源研究所（Research Center for Information Resources at National Institute of Informatics，簡稱 RCIR/NII）共同資助主辦的，提供一個完善的實驗環境，供各家學者評估其開發的資訊檢索系統（陳光華，2004）。

5

相較於歐美語系的 TREC(Text REtrieval Conference)檢索競賽， NTCIR 主要針對亞洲語系所設計出的一套檢索環境，其中可供實驗的語言有：(1)日語 (2)繁體中文 (3)簡體中文 (4)韓文 (5)英文。其中主要的目標包含(NTCIR Workshop, 2006)：

(1) 推動資訊處理技術(information access technologies)研究，提供大型的測試集爲實驗及共同的評量基礎，並且允許跨系統間的比較。

(2) 對研究團隊來說，其爲「跨系統間的評估」及「交換研究心得」的討論場所。

(3) 爲「資訊處理技術」研究評估的方法，並爲實驗建構一個大型可取得的資料集。

NTCIR 和 TREC 最大的差異處在於：

(1) TREC 近幾年著重於將檢索系統應用於各種環境之中，例如：部落格(blog)、垃圾郵件過濾、法律等…；而 NTCIR 偏重系統對查詢問題的資訊處理能力。

(2) TREC 以英文文件檢索爲主；而 NTCIR 以亞洲語系(日、中、韓文)爲主。因爲英文語系字與字間大都有明確的空格或符號標示，必須做「詞幹(word stemming)還原」的處理，例如：動詞時態變化等…及「字頭語判斷」例如：WHO、NBA、IR 等…；而亞洲語系大多必須斷字/詞，單一字的時態變化情況比較少，但字與字的組合變化較多。

三、文件二次排序(document re-ranking)

資訊檢索競賽評估中，參賽隊伍送出去的結果爲「系統查詢與主題相關的前 1000 篇文件」，除了依照相關文件左右分數外，相關文件的排序也是評估主因，也是爲了能在真實環境之中，符合使用者需求的文件能排序於前。因此，有些參賽團隊除了將查詢出的結果，依相關性作排序外，還依其他因素來調整文件權重，讓更相關文件能排序在前。

許多參賽團隊調整權重因素考慮不一，例如，成效最佳的 I2R 團隊，是將前

1000 篇文件依「關聯詞」、「詞彙長度(term length)」、「文件排序位置」計算賦予權重，再根據新的權重調整排序(Yang & Ji, 2005)；HKPU 是將「查詢標題」和「文件標題」作相似性運算，重新給予權重。由過去參賽者提供實驗可知，文件二次排序技術的確能有效的提升檢索成效(Xiao, Luk, Wong & Kwok, 2005)。

7

第五節 研究限制

爲了研究出有效提升成效的機制，將從 NTCIR3、4、5 屆的會議論文爲主要的研究文獻，分析成效較佳的參賽團隊技術，也試著藉此取得相關知識及研究，適當實驗這些技術，希望能有效提升檢索效能。但由於研究方向及領域不同，未能有效的涵蓋各方面研究主題，本研究以資訊檢索領域爲主，將研究範圍限制如下：

(1) 本研究針對 NTCIR 檢索會議 CLIR 單語檢索所提供的模擬環境進行探討

(2) 本研究以文件二次排序爲主要研究方向

(3) 以繁體中文爲主要研究語系

(4) 以 FJUIR 使用的檢索系統爲本研究實驗系統

第二章 文獻探討

第一節 日本 NTCIR 檢索會議

一、會議及任務項目

如同美國 TREC 一樣，日本 NTCIR 會議每年一次，第一屆會議在 1998 年十一月開始，爲期約約一年。以下爲 NTCIR 第 6 屆的重要日期供作參考，也可以從中了解 NTCIR 會議的運作模式。

表 2-1：

NTCIR 6 重要日期

項目	日期
參加申請開始	2006-04-17
參加申請結束	2006-05-31
文件集發佈	2006-06-01
預行實驗(Dry Run):	從 2006-07 到 2006-09
正式實驗(Formal Run)	從 2006-10 到 2006-12
發佈部分「任務報告」	2007-02-01
送回評鑑結果	2007-02-01
提出成果報告會議論文	2007-03-01
成果報告會	從 2007-05-15 到 2007-05-18

資料來源：”Important Dates,” by The 6th NTCIR Workshop, 2006, Retrieved MAY 4, 2006, from http://trec.nist.gov/overview.html

NTCIR 提供了各種不同的資訊處理環境供參賽者實驗，其中的項目及子項目如下所示：(NTCIR Workshop, 2006)

(1) 跨語言資訊檢索(Cross-Lingual Information Retrieval，簡稱 CLIR)：三項子項目 a. 多語檢索 b. 雙語檢索 c. 單語檢索

(2) 跨語文問題解答(Cross-Language Question Answering，簡稱 CLQA)：著重在標示實體命名(Named Entities)，爲亞洲語系文本「跨語言資訊處理」的一項

9

問題。目前共有五項子任務，C->C, E->C, C->E, E->J, J->E(C：中文、E：英文、J：日語) 而 J->J 是屬於 QAC 任務之中。而韓語部分還待考慮。

(3) 專利文件檢索(Patent Retrieval，簡稱 PATENT) ：包含了以下幾個子任務：a. 文件檢索 b. 段落檢索(Passage Retrieval)c. 分類

(4) 問題解答(Question Answering，簡稱 QAC) ：現今只針對日本文件。

(5) 試驗性任務(Pilot Tasks)：在很快的時間解答問題，或是能對未來的任務有所研究。目前考慮的任務爲「論點摘錄」(Opinion Extraction)，「多語多文件摘要」及「網路搜尋引擎評估」等…。

(6) 試驗性會議(Multimodal Summarization for Trend Information，簡稱 MuST)：趨勢資訊的多模型摘要，現只針對日本文件。

二、測試集

早期對檢索系統評估最著名的研究是 Cleverdon 在 1950 年代末期開始進行的 Cranfield 計劃，它開創了以測試集(Test Collection)配合測量準則(Measures)來評估系統的模式。所謂測試集，是一在規範化環境中測試系統效能的機制，包括測試問題(Queries)、測試文件集(Document Set)、及相關判斷(RelevanceAssessment)等三個部分(陳光華、陳信希，2004)。其研究設計的概念是假設在給定查詢問句與文件集中，文件集的某些文件是與查詢問句相關的。系統的目的是檢索出相關的文件，並拒絕不相關的文件，因此採用回收率(Recall)及精確率(Precision)作爲測量準則。Cranfield 研究首開規範化系統評估之先趨，其評估模式亦成爲後世普遍採用的標竿。(江玉婷、陳光華，1998)

(一) 測試文件集(Document Set)

NTCIR 所使用的測試文件集，主要是由新聞、科學技術文件所組成，在 NTCIR4 屆 CLIR 任務所使用的 CIRB030 已經達到 38 萬多筆(CIRB030 爲 CIRB011 結合

CIRB020)，已經有相當的規模，如下表所示，爲各屆所使用的測試文件集：

表 2-2：

NTCIR 各屆測試文件集

測試集	任務	文件						任務資料	
		類型	檔案名稱	語言	年度	文件數	容量	標題/ 問題	
								語言	數量
NTCIR-1	IR	科學技術文獻：抄錄	ntc1-je	JE	1988-1997	339,483	577MB	J	83
			ntc1-j	J		332,918	312MB		
			ntc1-e	E		187,080	218MB		60
	詞彙擷取/角色分析		ntc1-tmrc	J		2,000	-?	?	-
CIRB010	IR	新聞	CIRB010	Ct	1998-1999	132220	132MB	CtE	50
NTCIR-2	IR	科學技術文獻：抄錄	ntc2-j	J	1986-1999	400,248	600MB	JE	49
			ntc2-e	E		134,978	200MB		
NTCIR-3 CLIR	IR	新聞	KEIB010	K	1994	66,146	74MB	CtKJE	30
			CIRB011	Ct	1998-1999	132,173	870MB	CtKJE	50
			CIRB020			249,508			
			Mainichi	J		220,078			
			EIRB010	E		10,204			
			Mainichi Daily			12,723			
NTCIR-3 PATENT	IR	專利全文	kkh *3	J	1998-1999	697,262	18GB	CtCsKJE	31
		專利抄錄	jsh *3		1995-1999	1,706,154	1,883MB		
			paj *3	E		1,701,339	2,711MB		
NTCIR-3 QAC	QA	新聞	Mainichi	J	1998-1999	220,078	282MB	J*	1200
NTCIR-3 WEB	IR	網路 html/text	NW100G-01	multiple*4	crawled in 2001	11,038,720	100GB	J*	47
NTCIR-4 CLIR	IR	新聞	CIRB011	Ct	1998-1999	132,173	173.9MB	CtKJE	50
			CIRB020			249,508	285.3MB		
			Mainichi	J		220,078	282.7MB		
			Yomiuri	J		375,980	494.6MB		
			Hankookilbo	K		149,921	225.4MB		
			Chosunilbo			104,517	171.2MB		
			EIRB010	E		10,204	24.6MB		
			Mainichi Daily News (Japan)			12,723	33.3MB		
			Korea Times			19,599	55.9MB		
			Xinhua (AQUAINT)			208,168	290.2MB		
			Hong Kong			96,856	253.2MB		

11

			Standard						
NTCIR-4 PATENT	IR	專利全文	kkh *3	J	1993-2002	3,496,252	94.5GB	CtCsKJE	101
		專利抄錄	paj *3	E	1993-2002	3,496,252	5,482MB		
NTCIR-4 QAC	QA	新聞	Mainichi	J	1998-1999	220,078	278MB	J	queries
			Yomiuri			370,000	495MB		
NTCIR-4 WEB	IR	網路 html/text	NW100G-01	多語	crawled in 2001	11,038,720	100GB	J	47

註：J:日文, E:英文, C:中文 (Ct:繁中, Cs: 簡中), K:韓文

資料來源：. “Test Collections – DATA,” by The NTCIR 6 Workshop, 2006, Retrieved MAY 4, 2006, from http://research.nii.ac.jp/ntcir/permission/data-en.htm

在 NTCIR5 之後 CLIR 中文測試集已經增到第四版(CIRB040)，共有 901447 筆供實驗研究，都以 XML 方式標記資料，如下所示：

```
<DOC>
<DOCNO> edn_xxx_20000101_0098665 </DOCNO>
<LANG>CH</LANG>
<HEADLINE> 國人投保率已逾104％ </HEADLINE>
<DATE> 2000-01-01</DATE>
<TEXT>
<P>記者陳令軒／台北報導</P>
<P>保險業者歸納今年的保險趨勢，認爲重要議題將圍繞在兒童保單、產壽險兼營保險及年金險上，其中又以兒童保單影響最深。</P>
<P>壽險公會昨 (31)日舉行座談會，壽險公會秘書長張仲源、怡富投信副總經理蕭碧華、逢甲大學保險研究所所長袁國寧均就保險及理財相關問題，向消費者提出建議。</P>
<P>張仲源表示，保險法修正草案送審後，千禧年的保險變革將接踵而來，其中影響國人最深的莫過於兒童保單死亡險是否停售的問題。</P>
<P>他認爲若欲防止道德風險，恢復保險法第 107 條是不夠的，第 135 條也應限定兒童不得投保傷害險，如此一來，家庭保險可能面臨瓦解。</P>
<P>此外，張仲源也公布，至 88 年上半年爲止，國人壽險投保率達 104％，較 87 年底增加約 5 個百分點，其中兒童保單貢獻良多。</P>
</TEXT>
</DOC>
```

圖 2-1：測試文件集資料標示

NTCIR5 所使用的文件集和 NTCIR6 相同，資料來源是由下表 2-3 所示之新聞

文件收集而成：(陳信希、陳光華，2005)

表 2-3：

NTCIR5 所使用文件集

語言	來源		數量		
			2000	2001	總計
中文	CIRB040r (581.7 MB)	聯合報 (udn)	244,038	222,526	466,564
		United Express (ude)	40,445	51,851	92,296
		民新日報 (mhn)	84,437	85,302	169,739
		經濟日報 (edn)	79,380	93,467	172,847
		總計	448,300	453,146	901,446
日文	每日新聞	2000-2001 (118.8 MB)	99,207	100,474	199,681
	讀賣新聞	2000-2001 (343.3 MB)	306,709	352,010	658,719
	總計		405,916	452,484	858,400
韓文	韓國時報	2000-2001 (52.1 MB)	40,306	44,944	85,250
	朝鮮日報	2000-2001 (88.7 MB)	67,711	67,413	135,124
	總計		108,017	112,357	220,374
英文	每日新聞	2000-2001 (9.9 MB)	6,608	5,547	12,155
	韓國時報	2000-2001(25.3 MB)	16,461	14,069	30,530
	讀賣新聞	2000-2001(22.9 MB)	9,082	8,660	17,742
	新華社	2000-2001(來源：LDC)	107,956	90,668	198,624
	總計		140,107	118,944	259,051

資料來源：「跨語言資訊檢索與擷取測試集」，陳信希、陳光華，2005，民國九十六年六月四日，取自：http://www.csie.ntu.edu.tw/~ciet/form/paper/1.doc

(二) 測試問題(Queries)

本研究中，主題檢索之檢索方式非一般以標題或是分類號的方式描述需求，

13

而是採用類似 TREC 所創的多欄位的「查詢主題」(Topic rather than Query)，藉以多面向的欄位以描述需求。一般的主題檢索對需求之描述過短，其中包含的訊息可能不足，因而本研究採類似 TREC 的方式，對需求以多元的方式描述以及對檢索目的、檢索背景等其他層面之敘述以表達使用者需求，而 NTCIR 提供了中文主題檢索所需要的資料。(蔡育欽，2005)

以下爲 NTCIR5 的題目範例：

<TOPIC>
<NUM>001</NUM>
<SLANG>CH</SLANG>
<TLANG>CH</TLANG>
<TITLE>時代華納，美國線上，合併案，後續影響</TITLE>
<DESC> 查詢時代華納與美國線上合併案的後續影響。</DESC>
<NARR>
<BACK>時代華納與美國線上於 2000 年 1 月 10 日宣佈合併，總市值估計爲 3500 億美元，爲當時美國最大宗合併案。</BACK>
<REL>評論時代華納與美國線上的合併對於網路與娛樂媒體事業產生的影響爲相關。敘述時代華納與美國線上合併案的發展過程爲部分相關。內容僅提及合併的金額與股權結構轉換則爲不相關。</REL>
</NARR>
<CONC>時代華納，美國線上，李文，Gerald Levin，合併案，合併及採購，媒體業，娛樂事業</CONC>
</TOPIC>

圖 2-2：NTCIR5 題目範例

在 NTCIR 問題集由於下部分所組成：

(1) SLANG：問題集是由日本、韓國、台灣、以及 TREC 共同製作的，因此使用<SLANG>標記表明該問題的製作國家或機構。

(2) TLANG：標記則用於表明該問題目前的呈現語言。(陳光華、陳信希，2004)

(3) TITLE：描述查詢需求之主題，通常爲名詞或是名詞片語，並對查詢主題做簡單描述。爲 NTCIR 會議必須送出的結果。

(4) DESC(description)：用簡單的一至二個句子描述需求的主要內容。爲 NTCIR

會議必須送出的結果。

(5) NARR(narrative)：用數個句子描述查詢問題與專有名詞的定義，也描述了相關與不相關之範疇或其他特殊限制。

(6) CONC(concept)：用數個詞彙描述查詢主題中各層次相關的詞彙。(蔡育欽，2005)

(三) 答案集

當檢索系統從查詢問題在文件集找到相關文件後，答案集爲判斷檢索系統成效的最後準則，而答案集是由人工方式，從文件集判斷出相關文件。一般文件的相關判斷會被分爲四個層級：非常相關、相關、部分相關、與不相關，每一個層級都會設定代表的符號與數值。(陳光華、陳信希，2004)

表 2-4：

答案集相關判斷層級

相關層級	符號	分數
非常相關（Highly Relevant）	S	3
相關（Relevant）	A	2
部分相關（Partially Relevant）	B	1
不相關（Irrelevant）	C	0

資料來源：「CIRB030 資訊檢索測試集簡介」，陳光華、陳信希，2004，*中華民國計算語言學學會通訊*。

在 NTCIR 整理出兩組答案，一組爲嚴謹(Rigid)相關，也就是非常相關與相關視爲相關；一組爲鬆散(Relax)相關，也就是非常相關、相關、部分相關皆視爲相關(陳光華、陳信希，2004)。

15

表 2-5：

答案集的範例

題號	相關判斷	文件代號
001	S	udn_xxx_19990913_0332
001	S	udn_xxx_19990914_0208
001	C	udn_xxx_19990914_0679
001	C	udn_xxx_19990915_0244
001	S	udn_xxx_19990915_0246
001	C	udn_xxx_19990915_0633
001	C	udn_xxx_19990917_0516
001	C	udn_xxx_19990917_0676
001	A	udn_xxx_19990918_0246
…	…	…

三、檢索成效評估

檢索成效評估基礎是以精確率(precision ratio)與回收率(recall ratio)來做判斷，分子部分皆爲「檢索出之相關文件」，差異於分母間的關係。一般而言對此關係解釋意義如下：

(1) 精確率(precision ratio)：「檢索出的相關文件」佔「檢索出的所有文件」中的比例。

(2) 回收率(recall ratio)：或稱爲召回率，爲「檢索出的相關文件」佔「資料庫內所有相關文件」的比例。

表 2-6：

精確率、回收率和雜訊比之２乘２表格

	相關	不相關	總計
被檢索出	TP	FP	TP+FP
未被檢索出	FN	TN	FN+TN
總計	TP+FN	FP+TN	TP+FP+FN+TN

資料來源：” The Parametric Description of Retrieval Test,” by S. E. Robertson, 1969, *The Journal of Documentation,* 25:1, p.3.

(1) TP（True Positive）：代表相關文件被檢出的筆數。

(2) FP（False Positive）：代表不相關文件被檢出的筆數。

(3) FN（False Negative）：代表未被檢出之相關文件筆數。

(4) TN（True Negative）：代表正確回絕之不相關文件筆數。

又能用以下數學式描述之：

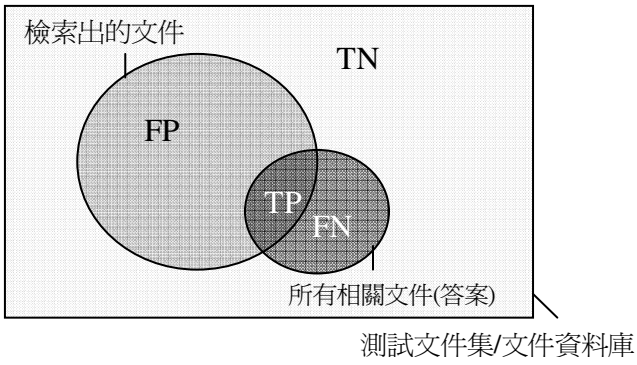


圖 2-3：精確率及回收率關係

$$\text{精確率（P）} = \frac{TP}{TP+FP} = \frac{\text{檢索出之相關文件 筆數}}{\text{檢索所得之所有文件 筆數}}$$

$$\text{回收率（R）} = \frac{TP}{TP+FN} = \frac{\text{檢索出之相關文件 筆數}}{\text{資料庫中所有相關文件 筆數}}$$

精確率與回收率常呈現反比的關係，爲了能有效的評估檢索系統的成效，結合以上兩種數據作爲準則成爲 F-VALUE，如下所示：

$$F-VALUE = \frac{2PR}{P+R} \qquad (2.1)$$

但現今的檢索系統除了檢索到的資料外，更重視相關資料排序，因而使用 NAP(Non-Interpolated Average Precision Rate)評鑑方式(Ricardo，1999)。TREC 使用 Buckley 發展的 TREC_EVAL 程式，用以更精確的評估排序的搜尋結果。TREC_EVAL 資訊檢索系統常用的評估工具(陳光華，2004)，也同爲 NTCIR 所使用。

TREC_EVAL 的執行畫面如下所示：

17

```
Queryid (Num):          42
Total number of documents over all queries
        Retrieved:      42000
        Relevant:        3284
        Rel_ret:         2547
Interpolated Recall - Precision Averages:
        at 0.00          0.7712
        at 0.10          0.5676
        at 0.20          0.5018
        at 0.30          0.4545
        at 0.40          0.4182
        at 0.50          0.3692
        at 0.60          0.3092
        at 0.70          0.2721
        at 0.80          0.1965
        at 0.90          0.1268
        at 1.00          0.0404
Average precision (non-interpolated) for all rel docs(averaged over queries)
                         0.3492
Precision:
    At       5 docs:     0.5476
    At      10 docs:     0.5048
    At      15 docs:     0.4683
    At      20 docs:     0.4393
    At      30 docs:     0.3921
    At     100 docs:     0.2557
    At     200 docs:     0.1819
    At     500 docs:     0.1041
    At    1000 docs:     0.0606
R-Precision (precision after R (= num_rel for a query) docs retrieved):
        Exact:           0.3734
```

圖 2-4：TREC_EVAL 執行結果範例畫面

數據意義如下所示：

(1) Queryid (Num)： 共檢索 42 題。

(2) Total number of documents over all queries：整個查詢中所有的文件。

Retrieved：檢索系統檢索出的 42000 筆文件。

Relevant：與查詢主題相關的有 3284 筆文件。

Rel_ret：系統檢索出與主題相關的文件有 2547 筆。

(3) Interpolated Recall - Precision Averages：以內差法的方式估計在固定的 Recall

下其相對的 Precision 值，爲所謂 11-point Precision，產生結果如下。

at 0.00： Recall 爲 0 時，Precision 爲 0.7712。

at 1.00： Recall 爲 1 時，Precision 爲 0.0404。

(4) Average precision (non-interpolated) for all rel docs(averaged over queries)：平均精確率，爲 NTCIR 評比檢索系統主要的依據，爲平均每篇相關文件被檢索出時的 Precision 值，其公式如下： (陳光華，2004)

$$Average\,\Pr ecision = AP(j) = \frac{\sum_{i=1}^{r_j} \frac{i}{\#Doc_j(i)}}{r_j} = NAP \qquad (2.2)$$

- r_j：系統在查詢主題題號 j 所檢索出相關文件數
- $\#Doc_j(i)$：系統對查詢主題編號 j 中，在第 i 篇相關文件被檢索出時，總共被檢索出的文件數

(5) Precision：精確率共有以下的狀況。

At 5 docs: 檢索出 5 篇文件時的 Precision 爲 0.5476。

At 1000 docs: 檢索出 1000 篇文件時的 Precision 爲 0.0606。

(6) R-Precision (precision after R (= num_rel for a query) docs retrieved): 表示檢索出 R 篇文件時的 Precision；R 是真正相關的文件數。

Exact：前 3284 篇時的 Precision 值爲 0.3734

19

第二節 FJUIR 於 NTCIR 使用的檢索機制

FJUIR 爲「輔仁大學圖書資訊學系資訊檢索研究室」，至今參與日本 NTCIR3、4、5 屆的單語檢索(SLIR)任務，及 NTCIR5 之專利文件檢索任務(PATENT)，FJUIR 團隊爲曾元顯教授所領導。

本論文的目的在於研究檢索文件，以 FJUIR 團隊檢索系統爲藍本，補強其檢索機制，使成效提升。FJUIR 在過去三屆 NTCIR 會議中，增加或改善了一些檢索方式，因此，以下探討 NTCIR 中，FJUIR 所使用最佳效能的檢索機制。

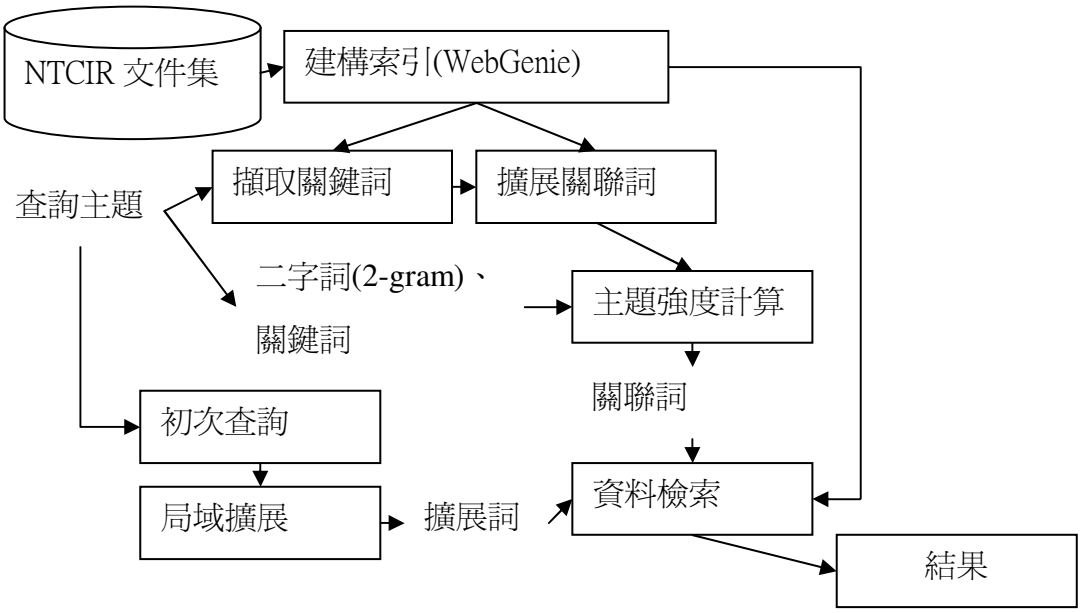


圖 2-5：FJUIR 於 NTCIR5(CLIR)檢索流程圖

一、建立索引與關鍵詞擷取

索引是資訊檢索最重要的工作，將文件、詞彙進行分析、轉換、組織，再進行有效率的運用，而 FJUIR 所用的檢索方式爲 N-grams 方式，配合停用詞(stop words)庫，過濾無意義的詞彙，用以保留有意義之詞。步驟如下：

(一)停用詞(stop words)：除去無意義的詞。例如：的、是、可以等…。

(二)n-grams：做完停用字處理後，再做 1-grams、2-grams 的處理，將文件內容依照字元分段成任意 n 個連續字元(2004，曾元顯)，將任意相連的字元都視爲可檢索的字串，例如：「資訊檢索」一詞中，做 1-grams、2-grams 的處理後，會產生「資」、「訊」、「檢」、「索」、「資訊」、「訊檢」、「檢索」七個索引詞。

(三)關鍵詞擷取：使用曾元顯教授的關鍵詞擷取演算法處理。之後將擷取出的關鍵詞也建置於反向索引檔之中。

```
1. Convert the input into a LIST.
2. Do Loop
2.1 Set MergeList to empty.
2.2 Put a separator to the end of LIST as a sentinel and
      set the occurring frequency of the separator to 0.
2.3 For I from 1 to NumOf(LIST) – 1 step 1, do
      If LIST[ I ] is the separator, Go to Label 2.3.
      If Freq(LIST[ I ]) > threshold and
            Freq(LIST[ I+1]) > threshold, then
         Merge LIST[ I ] and LIST[ I +1] into Z.
         Put Z to the end of MergeList.
      Else
         If Freq(LIST[ I ]) > threshold and LIST[ I ]
               did not merge with LIST[ I - 1], then
            Save LIST[ I ] in FinalList.
         If the last element of MergeList is not the separator, then
            Put the separator to the end of MergeList.
   End of For loop
2.4 Set LIST to MergeList.
  Until NumOf(LIST) < 2.
```

圖 2-6：關鍵詞擷取演算法

資料來源：”Uniform Indexing and Retrieval Scheme for Chinese, Japanese, and Korean,” by Y. H. Tseng, D. W. Juang, 2003, *NTCIR 3 Workshop*.

下圖爲字串經由演算法處理後，所產生的關鍵詞。

範例： 輸入字串: BACDBCDABACD.
假設 threshold (門檻值)=1, 分隔符號=x.
步驟一： 建立單一符號(tokens)串列
LIST = (B:3, A:3, C:3, D:3, B:3, C:3, D:3, A:3, B:3, A:3, C:3, D:3)

21

步驟二：經第一次程式流程

MergeList = (BA:2, AC:2, CD:3, DB:1, BC:1, CD:3,
DA:1, AB:1, BA:2, AC:2, CD:3)

FinalList = ( )

經第二次程式流程：

MergeList = (BAC:2, ACD:2, x, BAC:2, ACD:2)

FinalList = (CD:3)

經第三次程式流程：

MergeList = (BACD:2, x, BACD:2)

FinalList = (CD:3)

經第四次程式流程：

MergeList = (x)

FinalList = (CD:3, BACD:2)

圖 2-7：關鍵詞演算法執行流程

資料來源：”Uniform Indexing and Retrieval Scheme for Chinese, Japanese, and Korean,” by Y.-H. Tseng, D. W. Juang, 2003, *NTCIR 3 Workshop*.

(四) 反向索引檔(inverted file)：或稱倒置檔，記錄每個索引詞及其出現文件的編號，可從此索引檔直接取得包含某索引詞的所有文件(曾元顯，2004)。如下面範例：

詞＼文件	01	02	03	04	05	06	07	08
台北	1	1	1	1	0	1	0	1
台北市	0	1	0	1	0	0	0	1
執行	0	0	1	1	0	0	1	1
執行法條	0	0	0	1	0	0	0	0

1：爲該詞有出在文件編號 x；0 則表不出現

圖 2-8：反向索引檔(inverted file)範例

二、檢索模式

檢索模式爲檢索系統比對檢索條件與相關文件的依據，FJUIR 所使用的檢索模

式有許多種，以下介紹在 FJUIR 中所使用的檢索模式。(曾元顯，2005)

(一) 向量模式(Vector Space Model)

向量模式的檢索模型是由 Salton 等人在 1971 年所提出，爲提升檢索系統之效能及解決布林模式的諸多限制，其作法爲：（Ricardo，1999）

1. 將使用者的詢問句及資料庫中的文件轉換成維度（Dimension）相同的向量表示法。

   q=[W1,Q,W2,Q,…Wn,Q]，代表查詢向量

   dj=[W1,dj,W2,dj,…Wn,dj]，代表文件向量

2. 以 Cosine 來計算兩向量的夾角，其值介於 0 到 1 之間

將文件與查詢句轉換爲向量之後，就可以用量化方式處理，並計算其相似度。

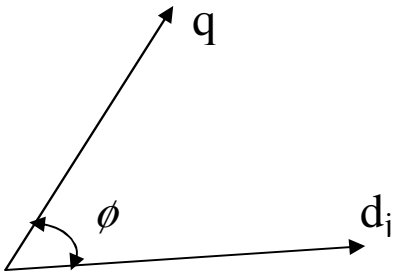


圖 2-9：向量模式示意圖

資料來源：”Modern Information Retrieval,” by B.Y. Ricardo & R.N. Berthier, 1999.

$$sim(q,d_j)=\frac{d_j \bullet q}{|d_j|\times|q|}=\frac{\sum_{i=1}^{t} w_{j,dj}\times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,dj}^2}\times\sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \qquad (2.3)$$

(1) ByteSize Normalization

ByteSize Normalization 檢索模式，是依照詞彙向量關係所延伸出來的計算方式。而 ByteSize 是依照向量關係所延伸出來的檢索模式，

23

$$sim(d_i, q_j) = \frac{\sum_{k=1}^{t} d_{i,k} q_{j,k}}{(bytesize_{di})^{0.375} \sqrt{\sum_{k=1}^{t} q_{j,k}^2}}$$

- 文件 d_i 與 查詢 q_j
- T : 查詢詞彙的個數
- bytesize : 文件的長度，byte 個數

(2) Pivoted Normalization Method

Singhal 等人於 1996 SIGIR 會議提出改進 Bytesize 的公式，爲目前最佳的向量模式。

$$Pivot(d_i, q_j) = \sum_{k=1}^{T} tf_{j,k} \log\left(\frac{n+1}{df_k}\right) \left(\frac{1+\log(1+\log(tf_{i,k}))}{(1-s)+s\frac{dl_i}{Avgdl}}\right)$$

- T ：查詢主題 qj 的查詢詞個數
- $tf_{j,k}$：某個查詢詞 k 在查詢主題 q_j 的詞頻
- $tf_{i,k}$：查詢詞 k 在文件 i 的詞頻
- df_k：查詢詞 k 出現的文件篇數
- dl_i：文件 i 的長度（任意單位，如 byte size）
- Avgdl ：所有文件的平均長度（與上面同單位）
- n ：文件總篇數

(二) 機率模式(Probability Model)

以機率的概念處理資訊檢索的問題。先建立一組相關文件作爲理想解答來當

作訓練文件，透過遞迴的方式計算資料間分布的關係，使得查詢句對應到文件的機率計算能達到最大，最後得到一組收斂的參數。當資料集夠大時，該參數將可代表一般情況下理想的分布情形。因此機率模型一開始是以猜測的方式取得初步的理想解答，系統再利用此資訊改善理想解答，經重複多次處理後接近實際的理想解答（Ricardo, 1999）。機率模式以 OKAPI 權重模式爲最佳的檢索模式。

(1)BM11

BM11 在 NTCIR 的檢索評比有很好的成效(Robertson, 1994)

$$BM11(d_i, q_j) = \sum_{k=1}^{T} tf_{j,k} \log\left(\frac{n - df_k + 0.5}{df_k + 0.5}\right)\left(\frac{tf_{i,k}}{tf_{i,k} + \frac{dl_i}{Avgdl}}\right)$$

- T 是查詢主題 qj 的查詢詞個數
- tfj,k 是某個查詢詞 k 在查詢主題 qj 的詞頻
- tfi,k 是查詢詞 k 在文件 i 的詞頻
- dfk 是查詢詞 k 出現的文件篇數
- dli 是文件 i 的長度（任意單位，如 byte size）
- Avgdl 是所有文件的平均長度（與上面同單位）
- n 是文件總篇數

(2)BM25

BM25 在 TREC 的檢索評比有很好的成效(Fang, 2004)

$$BM25(d_i, q_j) = \sum_{k=1}^{T} \frac{(k_3+1)tf_{j,k}}{k_3 + tf_{j,k}} \log\left(\frac{n - df_k + 0.5}{df_k + 0.5}\right)\left(\frac{(k_1+1)tf_{i,k}}{tf_{i,k} + k_1((1-b) + b\frac{dl_i}{Avgdl})}\right)$$

25

- T：查詢主題 q_j 的查詢詞個數
- $tf_{j,k}$：某個查詢詞 k 在查詢主題 q_j 的詞頻
- $tf_{i,k}$：查詢詞 k 在文件 i 的詞頻
- df_k：查詢詞 k 出現的文件篇數
- dl_i：文件 i 的長度（任意單位，如 byte size）
- Avgdl：所有文件的平均長度
- N：文件總篇數

(3)BM25m

修改自 BM25，將$(n-df_k+0.5)$改成 n+0.5，因爲$(n-df_k+0.5)$有可能出現爲 0，此情況將無法進行計算

$$BM25m(d_i,q_j)=\sum_{k=1}^{T}\frac{(k_3+1)tf_{j,k}}{k_3+tf_{j,k}}\log\left(\frac{n+0.5}{df_k+0.5}\right)\left(\frac{(k_1+1)tf_{i,k}}{tf_{i,k}+k_1((1-b)+b\frac{dl_i}{Avgdl})}\right)$$

- T：查詢主題 q_j 的查詢詞個數
- $tf_{j,k}$：某個查詢詞 k 在查詢主題 q_j 的詞頻
- $tf_{i,k}$：查詢詞 k 在文件 i 的詞頻
- df_k：查詢詞 k 出現的文件篇數
- dl_i：文件 i 的長度（任意單位，如 byte size）
- Avgdl：所有文件的平均長度
- N：文件總篇數

(三)語言模式 Language Model

其目前最佳的語言模式之公式爲 Dirichlet Prior Method，其公式如下

$$LM(d_i, q_j) = \sum_{k=1}^{T} |q_j| \log(\frac{\mu}{dl_i}) + tf_{j,k} \log\left(1 + \frac{tf_{i,k}}{\mu \frac{df_k}{n}}\right)$$

- T：查詢主題 qj 的查詢詞個數
- $tf_{j,k}$：某個查詢詞 k 在查詢主題 qj 的詞頻
- $tf_{i,k}$：查詢詞 k 在文件 i 的詞頻
- df_k：查詢詞 k 出現的文件篇數
- dl_i：文件 i 的長度（任意單位，如 byte size）
- n：文件總篇數
- $|q_j|$： 查詢長度

三、查詢擴展

由於檢索系統無法判斷詞彙相關意義，因此常常會找不到具有相關意義；但運用不同詞彙表達的文件，資訊檢索系統常利用查詢擴展（Query Expansion）的方式補足使用者所提供查詢詞的不足，以達到檢索成效的提升(蔡育欽，2005)。

(一)局域擴展（Local Expansion）

在本文中是指將第一次檢索結果的前 N 篇經過相關排序之後，抽取出前 M 個關鍵詞作爲擴充詞彙加入原先的查詢。也就是擴充詞的來源爲部份文件之集合(蔡育欽，2005)。又可稱爲 BRF(Blind Relevance Feedback)或 PRF(Pseudo Relevance Feedback)，在 FJUIR 的局域擴展方式爲「取檢索結果前 6 篇文件最佳 30 個關鍵詞」

27

送回查詢。(Tseng, Juang & Chen, 2004)

(二) 全域擴展（Global Expansion）

運用所有資料集之文件，建立一索引典之架構以定義出詞與詞之間的關係，並找出與查詢詞有關的詞作爲擴充詞彙，加入初始查詢(蔡育欽，2005)。

由於局域擴展運用檢索排序前幾篇的文件，由於那些文件本身具有相當程度的與主題相關性，因此擴展出來的詞彙也能穩定的控制，因此該詞彙的確能有相當程度的幫助。使用全域擴展，其擴展的詞彙量大，容易造成的主題偏移，在 NTCIR4 實驗中，使用全域擴展提升程度不大，而使用局域擴展反而都有顯著的成效提升。下表爲 NTCIR4 中 FJUIR 全域擴展及局域擴展實驗於 NTCIR3 的成績。

表 2-7：

NTCIR4 全域擴展及局域擴展的實驗成績

NTCIR3				
RunID	Rigid		Relax	
	MAP	% imp	MAP	% imp
C-C-D	0.1858	-	0.2281	-
*C-C-D+AT	0.1894	1.94	0.2432	6.62
*C-C-D+BRF	0.2246	20.88	0.2796	22.58
*C-C-D+BRF(p)	0.2474	33.15	0.3009	31.92
Max of C-C-D	0.3933		0.4990	
Avg of C-C-D	0.2130		0.2670	
Min of C-C-D	0.0347		0.0443	

註：爲送出的成績；AT 爲全域擴展；BRF 爲局域擴展；(P)爲使用機率模式

資料來源：”Global and Local Term Expansion for Text Retrieval,“ by Y. H. Tseng, D. W. Juang, & S. H. Chen, 2004, Working Notes of NTCIR-4, Tokyo.

在 NTCIR5 中實驗了數個策略來找尋較佳的查詢詞彙，嘗試使用一些限制策略避免主題偏移(topic drift)，而實驗後的結果以下述策略成效較佳，「以關聯詞與主題之強度作爲排序篩選的依據，即計算出關聯詞對整個查詢主題之強度，而非對個別關鍵詞之強度。並嘗試以多種公式計算強度取檢索結果最佳者」(蔡育欽，

2005)，使用該策略過濾選出來的詞再結合局域擴展的詞，送入查詢，在 NTCIR5 中達到 FJUIR 較佳的成效，但最後還是未能達到理想的結果 (Tseng, et al., 2005) 。

四、系統效能評估

從 NTCIR3、4、5 屆中，FJUIR 團隊參與 CLIR 之單語檢索任務(SLIR task)，吸收其他優秀團隊的研究技術及經驗，在檢索系統上，研究技術的方向如下所示：

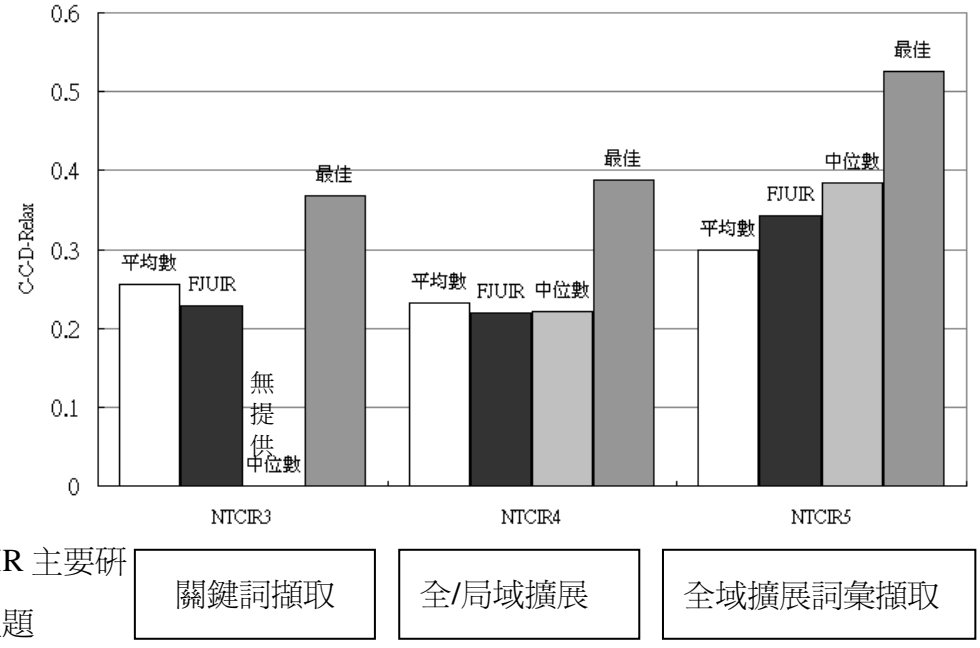


圖 2-10：FJUIR 歷年檢索技術方向

由於每年的題目不同，因此上面分數供參考，但是可以看出各隊參賽隊伍分數差距開始拉大，因此在如此競爭環境下，「不進則退」，因此希望能先達到一定的水準之前，先吸收成效佳團隊的經驗，將成效提升到一定的水準，進而再向最佳成效邁進。

檢索系統面對各式主題查詢時，會因爲查詢問題的難易、表達方式，文件集使用的文件特性、檢索系統索引方式及檢索策略等…因素，而使檢索系統在處理文件時，較無法搜尋出相關文件。對於本研究所實驗的檢索系統，在過去的實驗中(蔡育欽，2005)，分析出系統對於各式查詢詞檢索成效的特性，雖然查詢詞在真

29

實情況，使用者使用的詞較難歸納出明顯的特性，但參考檢索會議所提供的環境，不外乎著重於

a. 查詢詞彙擴展方式

b. 文件排序再調整

對於本實驗系統來說，在「查詢詞彙擴展」使用 PRF 已經達到一定的成效，雖然此部分也許還有需要再研究的部分，但相較於「文件排序再調整」來說，這比較本研究系統系統最缺少的部分。因此本研究著重的方向也會偏於得到一個較佳的「文件排序再調整」機制。

第三節 文件排序研究

本研究爲了提升 FJU 團隊的主題研究成效，先針對 FJU 團隊所使用的技術爲基礎，參考分析其他效能較高的團隊所使用的檢索技術，進而增加 FJU 團隊的檢索成效。

成效最佳的前幾名團隊，是參考 NTCIR3、4、5 屆會議工作報告，所整理出來的隊伍，而整理方式及排序如下：

(1) NTCIR3 屆：大會提供的資料，是只以成效最高的分數做排序，例如：C、TDNC、DN、D(Title、Description、Narrative、Concept)綜合排名。因此在表 2-8 中，NTCIR3 的排列方式是以參考大會提供的「綜合排名」爲主。

(2) NTCIR4、5 屆：由於 NTCIR3 排名在前的團隊的多以送 TDNC 的分數最好(因爲送入檢索的查詢詞彙最多)。NTCIR4、5 大會改以提供最佳團隊的 D(Description)-run 成績，因此而 NTCIR4、5 改以 D(Description)-run 成績爲主。因此在表 2-8，NTCIR4、5 屆的排列方式是參考大會提供的 Rigid-D(Description)-run 爲主。

從技術文件也不難看出，各研究團隊在單語檢索領域中，無不致力研究以下兩個主題：

(1) 如何得到更有效的查詢詞的擴展方式

(2) 再初次排序之後，如何調整更佳的文件排序

如下表所示，爲 NTCIR3、4、5 屆成效最佳研究團隊：

表 2-8：NTCIR3、4、5 屆成效最佳參賽團隊

屆別	團隊代號	團隊名稱	國家	成效較佳的結果	
				Rigid	Relax
NTCIR3	pircs	Queens College City University of New York	USA	TDNC、D	TDNC、D
	CRL	Communications Research Laboratory	Japan	TDNC、TC、D	TDNC、TC、D

31

	APL	Applied Physics Laboratory Johns Hopkins University	USA	TDNC	TDNC
	BRKLY	University of California at Berkeley	USA	D	D
	HKPU	Hong Kong Polytechnic University	Hong Kong	TDN、CT	TDN
	CMU	Language Technologies Institute Carnegie Mellon University	USA	TDNC	TDNC
	HUM	Hummingbird	Canada	TDNC	
NTCIR4	I2R	Natural Language Processing Lab; Institute of Inforcomm Research	Singapore	D、T	D、T
	OKI	Oki Electric Industry	Japan	D、TDNC、T	D、TDNC、T
	pircs	Queens College City University of New York	USA	D、T	D、T
	RCUNA	Ubiquitous Solution Lab; Software R&D Group; RICOH COMPANY; LTD	Japan	D、T	D、T
	UniNE	University of Neuchatel	Switzerland	TDNC、D、T	D、T
	KLE	Knowledge & Language Engineering Lab.; Pohang University of Science & Technology	Korea	TDNC、D、T	D、TDNC、T
	IFLAB	University of Tsukuba; Ishikawa-Fujii Laboratory	Japan	DN、D、T	D、DN、T
	JSCCC	Clairvoyance Corporation	USA	DN、D、T	D、DN、T
NTCIR5	I2R	Natural Language Processing Lab; Institute of Inforcomm Research	Singapore	D、T	D、T
	UniNE	University of Neuchatel	Switzerland	DN、T	D、T
	ISCAS	Institute of Software; Chinese Academy of Sciences	China	D、T	D、T
	pircs	Queens College City University of New York	USA	D、DN、T	D、DN、T
	CCNU	Central China Normal University	China	D	D
	OKI	Oki Electric Industry	Japan	D、TDNC、T	D、TDNC、T
	HKPU	The Hong Kong Polytechnic University	Hong Kong	D、TDNC、T	D、TDNC
	UNTIR	University of North Texas	USA	D、TDNC、T	D、DN、T

註：代號 T、D、N、C，分別代表檢索問題的敘述欄位 Title、Description、Concept、Narrative。

本研究將從上述「最佳參賽團隊」所提供的會議文件，分析出能有效提升檢索成效的機制，並且整理參考其他相關文獻來判斷技術可行性。

(1) 優先參考成效最高的參賽團隊檢索機制。

(2) 基礎檢索部分由於差異不大，暫時不考慮。

(3) 查詢擴展的方式有網路詞彙擴展(He, Qu, Tu & Ji, 2004)、ontology 廣域詞彙擴展(SYang, Ji & Tang, 2004)等…，考慮到「處理時間效能」及「該團隊未在之後幾屆使用」的情況下，便不再考慮實作。

綜合以上文獻整理後的特性，本研究將實驗及文獻整理方向設定以「排序研究」爲主，這個方向也是過去未曾致力的方向。從各參加團隊技術文件歸納，在基礎檢索部分，各團隊成效大同小異，索引部分以 2-gram 爲主，檢索模組以 Okapi 的 BM25 機率模式爲主，也由各團隊驗證，得到不錯的成效。

CLIR 任務文件二次排序(Document Re-ranking)技術在過去就有參賽團隊在實作，NTCIR5 會議以新加坡 I2R（Natural Language Processing Lab; Institute of Inforcomm Research）團隊的運用方式，將其檢索成效顯著提升，如下表所示：

表 2-8：

NTCIR3-5 屆 CLIR 任務使用 Re-ranking 技術的團隊

<table>
<tr><th rowspan="2">屆別</th><th rowspan="2">團隊
代號</th><th rowspan="2">團隊名稱</th><th colspan="2">結果項目</th><th rowspan="2">該項目平均分數</th></tr>
<tr><th>rigid/relax</th><th>分數</th></tr>
<tr><td rowspan="4">3</td><td rowspan="2">ULIS</td><td rowspan="2">University of Library and Information Science</td><td colspan="2">J-J-d</td><td rowspan="2">0.2633</td></tr>
<tr><td>Relax</td><td>0.3427</td></tr>
<tr><td rowspan="2">TSB</td><td rowspan="2">Knowledge Media Laboratory, Toshiba Corporate R&D Center</td><td colspan="2">J-J-D</td><td rowspan="2">0.2633</td></tr>
<tr><td>Relax</td><td>0.3910</td></tr>
<tr><td rowspan="4">4</td><td rowspan="2">I2R</td><td rowspan="2">Natural Language Processing Lab; Institute Inforcomm Research</td><td colspan="2">C-C-D</td><td rowspan="2">0.1826</td></tr>
<tr><td>Rigid</td><td>0.3255</td></tr>
<tr><td>PolyU</td><td>The Hong Kong Polytechnic, department of computing</td><td colspan="3">以 ntcir3 做實驗，於 NTCIR5 中更名爲 HKPU</td></tr>
<tr><td rowspan="6">5</td><td rowspan="2">I2R</td><td rowspan="2">Natural Language Processing Lab; Institute Inforcomm Research</td><td colspan="2">C-C-D</td><td rowspan="2">0.2986</td></tr>
<tr><td>Rigid</td><td>0.4826</td></tr>
<tr><td rowspan="2">CCNU</td><td rowspan="2">Central China Normal University</td><td colspan="2">C-C-D</td><td rowspan="2">0.2986</td></tr>
<tr><td>Rigid</td><td>0.3441</td></tr>
<tr><td rowspan="2">HKPU</td><td rowspan="2">The Hong Kong Polytechnic, department of computing</td><td colspan="2">C-C-D-01</td><td rowspan="2">0.2986</td></tr>
<tr><td>Rigid</td><td>0.3330</td></tr>
</table>

33

	RYU	Ryukoku University	JJT		0.2954
			Rigid	0.1802	

註：團隊名稱**粗**體表示有進入中文單語前八名。

由於 TREC_EVAL 評審方式，將文件的順序納入評比的主要因素，因此排序研究在資訊檢索領域興起，大部分團隊使用文件二次排序技術後，其成效皆有明確的提升，而每一個團隊重新排序的狀況大不相同，其中以 I2R 及 HKPU 團隊的研究較爲明確，實驗結果有很好的成效，以下分析 I2R 團隊的實作方式。

(一) I2R 文件二次排序

依查詢詞調整文件權重的方式，爲前 1000 篇文件計算「文件頻率」、「詞彙長度(term length)」、「文件排序位置」根據新的權重重新計算文件分數。在 NTCIR4，I2R 的 Re-ranking 檢索系統流程如下圖所示：

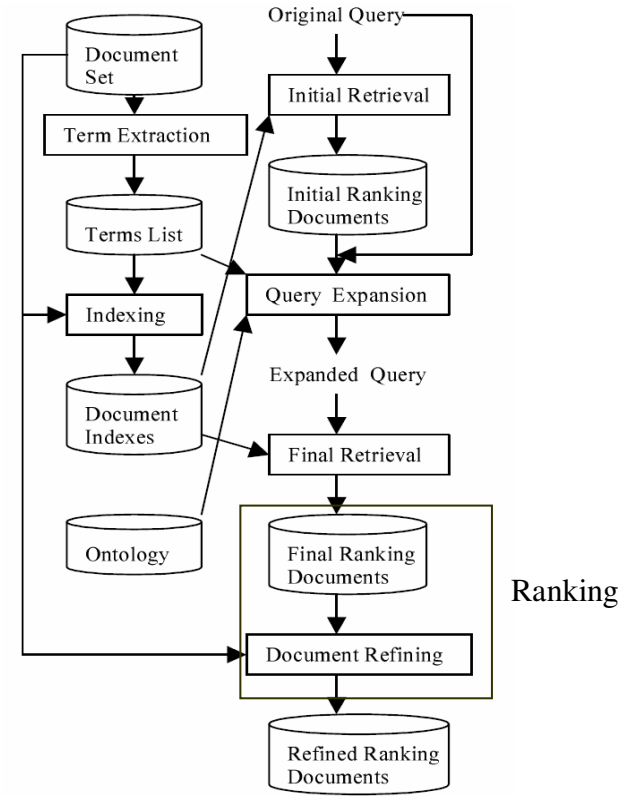


圖 2-11：I2R 團隊檢索系統流程圖

資料來源：”Chinese information retrieval based on terms and ontology,” L. Yang, D.Ji,

& L. Tang, 2004, Working Notes of the Fourth NTCIR Workshop Meeting, Cross-Lingual Information Retrieval Task.

NTCIR5 會議中，I2R 將區域擴展置於 Re-ranking 步驟之後，從排序更佳的文件中，擴展到更相關的詞。由 I2R 在 NTCIR4 發表的會議文件，可以了解整個二次排序將權重重新分配的流程，大致分成兩個步驟，爲「文件二次排序」 及「文件微調(Document Refining)」，

文件二次排序是預期有更相關的文件出現在較高的排序之中，以提升「查詢擴展」成效(Yang & Ji, 2005)，從「相關文件頻率」、「文件位置」及「詞彙長度」來協助文件二次排序。前 1000 篇文件中擷取關鍵字，假設這些關鍵字對「二次排序」有幫助，而且只去分析出現在查詢詞裡的詞彙。這表示此步驟不用再作任何查詢擴展，因此擷取出的詞也可視爲查詢詞，而給予權重的方式如下：

(1)相關分佈：前 1000 相關文件頻率對整個文件集文件頻率。愈是出現在前 1000 篇文件的詞相對於整個文件集，愈加重要。

(2)詞彙長度：詞彙字數，詞彙長度愈長對該詞彙運算的精確度愈有幫助。

(3)文獻排序位置：用一個連續數字記錄前 1000 篇文件的位置。

給予詞彙 t 權重方式公式如下：(Yang et al., 2005)

$$\sqrt{\frac{\left(\sum_{i=1}^{1000} df\ (t, d_i) \times f(i)\right) / 1000}{DF\ (t, C) / R}} \times \sqrt{|t|}$$

$$df\ (t, d_i) = \begin{cases} 1 & t \in d_i \\ 0 & t \notin d_i \end{cases}$$

- df(t,d) 爲 di 文件是否出現 t
- DF(t, C) 爲該文件集 C 中出現 t 的文件數
- d_i=第 i 篇文件
- R=文件集 C 中的全部文件數

35

- |t|=關鍵字長度
- f(i)=1+1/Sqrt(i),爲 d_i 的權重

如果一個詞彙在一個文件中出現頻率低，且該文件排序較低；以及一個高頻詞，並且出現在排序較高的文件，相較之下，通常處理文件頻率的方式都計算爲 1 次(count)，無論文件於何處，都視爲較不重要。

在前 1000 篇文件，首先找出查詢詞出現在那一篇文件，並對這些查詢詞加上權重，之後聚集這些值，並且在文件及查詢詞間計算新的排序成績，最後，使用新的排序成績，重新排序前 1000 篇文件。演算法如下圖所示：

步驟 1：在查詢詞 q 中取得詞彙和其權重

1.1 檢索後的前 1000 篇中的每一篇文件 d 中擷取詞；實際上，詞彙擷取方式，只須做一次，而此步驟可做爲索引的一部分。

1.2 收集出現在 q 的詞彙，並且用上述公式計算權重

步驟 2：重新排列 1000 篇文件

2.1 前 1000 篇中每一篇文件 d_i 用下列公式計算其新的排序值 Si，Ri 爲初始相似度。

$$S_i = R_i \times (1 + \sum_{t_j \in q, d_i} W(t_j))$$

2.2 重新排列前 1000 篇文件，依照{S_1,S_2,.....S_i...,S_{1000}}

圖 2-12：文件二次排序演算法

資料來源：”Chinese information retrieval based on terms and ontology,” by L. Yang,D. Ji & L. Tang, 2004, Working Notes of the Fourth NTCIR Workshop Meeting, Cross-Lingual Information Retrieval Task.

I2R 團隊在 NTCIR 第四屆及第五屆爲最佳成效團隊，從 NTCIR3 中參賽開始，便設定了未來發展的方向及基礎 (Ji, Yang & Nie, 2003)，最後在 NTCIR4、5 完成

進而達到最佳成績。該隊在 NTCIR5 中實驗證實「文件二次排序」(document re-ranking)方式成效相當良好。如下表所示：

表 2-9：

I2R 團隊於 NTCIR5 之成績

	C-C-T		C-C-D	
	Rigid	Relax	Rigid	Relax
初始檢索	0.2928	0.3551	0.2685	0.3326
檢索後再次排序	0.3473	0.4251	0.3184	0.3725
二次排序後查詢擴展	0.5047	0.5441	0.4826	0.5249
FJUIR(NTCIR5)	0.2604	0.3217	0.2820	0.3425

資料來源：”I2R at NTCIR5,” by Yang & Ji, 2005, Proceedings of NTCIR-5 Workshop Meeting.

FJUIR 相較之下和 I2R 初始檢索的成績不相上下，因此可以看出實作「文件再次排序」之後，或是「再次查詢擴展」，可能會達到相當的成效。本研究實驗部分將實作「文件二次排序」(document re-ranking)的方式，並參照該隊的作法。

但假設實作 I2R 團隊所得的結果無法提升成效，將改以實作 HKPU 團隊的重新排序方式，實驗是否能提升成效，若還是無法提升，盡可能的因素分析，了解問題及原因。

另外 I2R 研究團隊也於 NTCIR4 提出，文件排序微調(Refining)技術，文件排序微調(Refining)是，在最後的排序中對前 M 篇文件(M<2000)做「微排序」(micro-rank)，排序方式是用「詞彙涵蓋率」及「事件偵測」(event detection)。

事件偵測是試著找到「查詢和事件是否有關」，如果一個查詢詞偵測出是和事件相關的文件，其發佈日期落在這段時間文件，將視爲事件相關文件，運用<date></date>標籤欄位，在「文件二次排序」之後的前 N 篇，去偵測該查詢詞是否與事件相關。如果文件是落在事件的日期範圍，將會加強其權重。例如查詢問題第二題：(Yang, et al., 2004)

37

<TITLE>約翰走路，菁英高爾夫球慈善賽，台灣</TITLE>

<DESC>查詢 1999 年來台參加約翰走路菁英高爾夫慈善賽的國際高爾夫球星及相關活動的內容</DESC>

圖 2-13：NTCIR4 查詢問題第二題

此題系統從 1999 年 11 月 09 日和 1999 年 11 月 14 日中找出許多相關文件，所以系統判斷此查詢是和一個事件有關，因此落於 1999 年 11 月 09 日及 1999 年 11 月 14 日的文件給於更多的權重。另一個例子如查詢 13 所示(Yang, et al., 2004)：

<TITLE>日本，首相，小淵惠三，訪問，美國</TITLE>

<DESC>查詢日本首相小淵惠三訪美相關內容</DESC>

<NARR><BACK>日本首相小淵惠三於 1998 年四月二十九日啓程訪美，這是日本首相十二年來首次正式訪美。小淵惠三此次的美國行帶給美國諸多見面禮，包括雙方經濟合作對策、國際金援以及美日防衛合作方針配套法案，使得美日安保體制進入新階段。請查詢此次訪問活動的內容，包括日本所提出的支援、美日的合作協議等。<BACK>

<REL>相關資料爲小淵惠三訪美的內容報導，各國的看法或意見則爲無關。<REL>

</NARR>

圖 2-14：NTCIR4 查詢問題第 13 題

此題，雖然背景敘述拜訪行程在 1998 年 04 月 29 日，而系統卻在 1998 年 4 月 27 日及 1998 年 05 月 06 日找出大部分的相關文件，並非在 1998 年 04 月 29 日，所以系統還是判斷此題和事件相關，因此落於 1998 年 04 月 27 日及 1998 年 05 月 06 日發佈的新聞加重其權重(Yang, et al., 2004)。

文件涵蓋率(coverage)排序方式，主要做法爲如果該文件涵蓋很多「查詢詞彙」，將其權重加重。除此之外，還可以因詞彙涵蓋數，及詞彙涵蓋文件長度(total

length of terms coved)，等…(Yang, et al., 2004)。

(二) 標題二次排序

HKPU 二次排序的方式是依查詢詞和文件集標題作相似度計算，再重新排序。依文件寫作習慣來說，文件標題會較代表文件內容，因此在文件標題上的詞可以加強權重，表示該詞更具代表性，且該文件也爲重要的文件。以下簡介 HKPU 研究團隊的文件二次排序做法。

HKPU 文件二次排序技術公式如下：

$$sim'(q,d_i) = \left(sim(q,d_i) - m\right) \times M(q_t, t(d_i)) + m$$

- sim(q,di)：原始相似分數
- m：前 n 篇文件中最小的原始相似分數
- t(di)：第 i 篇文件的 title
- qt：符合查詢 q 的標題(corresponding title query of q)
- M(.)：查詢標題(title query)和文件集標題符合(matched)的詞數

重新排序之後，HKPU 使用局域擴展(PBF)來增加成效，而 HKPU 做了多個 PBF 實驗，以計算最佳的回饋文件數及詞數。由於局域擴展處理的文件數較少，因此 HKPU 也實驗了三字詞(trigram)回饋方式 (Xiao, Luk, Wong & Kwok, 2005) 。雖然局域擴展的研究所得出來的數據較不具代表性(很多團隊都有各自最佳回饋方式)，但由 HKPU 團隊的研究方法值得借鏡。

(三) 以歸類方式二次排序

在 I2R 於 NTCIR5 提出「文件二次排序」的方式後，於 2006 年提出另一個調整排序的方式(Yang, Ji, Zhou, Nie & Xio, 2006)，使用文件歸類的方式來調整修正，將初次檢索依相似度排序的結果，分成三個類別：

39

(1) 排序於前的文件，假設爲相關文件。

(2) 排序於後的文件，假設爲不相關文件。

(3) 其他排序於中間的文件，待歸類的文件。

因此可以使用歸類的方式，將初次排序後的文件，再依特性分類，可以成爲調整相似度的依據，歸類模組可以使用 KNN、VSM 或 Label Propagation 方式，而 Label Propagation 也可以視爲 KNN 的另一種方式。

在 I2R 研究團隊(Yang, et al., 2006)的研究中，使用 Label Propagation 的歸類方式進行文件二次排序，方法如下所示：

將初次結果參照上面的模式分成三個資料集：

(1) M：未標記(label)文件。

(2) R：假設爲相關文件，視爲已標記文件。

(3) N：假設爲不相關文件，視爲已標記文件。

而處理過程也是依 R、N 的文件特性將 M 歸類完成。演算法如下圖 2-15：

輸入：

q:查詢詞

M:需要被再次排序的文件組

R:前 k 篇文件假設爲相關的文件組

N:後面幾篇文件假設爲不相關的文件組

演算法: Label Propagation(q, M, R, N)

開始

設定變數 *t=0 ;*

BEGIN DO Loop

用 $Y^{t+1} = \overline{T}Y^{t}$ 繁殖標記;

Y^{t+1} 的上層 l 列取代爲 Y_L^o ，用以縮減(clamp)標記資料;

END DO Loop 當 Y_t 聚合時;

根據 Y_{h1}(爲相關文件的機率)再次排序文件 x_h (l+1≤h≤l+M) 結束

圖 2-15：Label Propagation 歸類調整文件排序

資料來源：”Document Re-ranking Using Cluster Validation and Label Propagation, ” by Lingpeng Yang, Donghong Ji, Guodong Zhou, Yu Nie & Guozheng Xiao, 2006, Proceedings of the 15th ACM international conference on Information and knowledge management CIKM '06, pp. 690 – 697.

其中(Yang et al., 2006)：

(1) $X = \{x_i\}$ $(1 \le i \le R+N+M)$：文件群集，例如：

a. x_i $(1 \le i \le R)$ 表示 R 相關標記文件 $\{r_j\}$ $(1 \le j \le R)$, 。

b. x_i $(R+1 \le i \le R+N)$ 表示 N 不相關標記文件 $\{n_j\}$ $(1 \le j \le N)$ 。

c. x_i(R+N+1≤ i ≤R+N+M)表示 M 檢索後文件$\{m_j\}$ $(1 \le j \le M)$ ，需被調整排序。

(2) $C = \{c_j\}$ $(1 \le j \le 2)$：歸類後的一個叢集(class)，c1 表示文件相關而 c2 表示文件不相關

(3) $Y^0 \in H^{s \times 2}$ (s=R+N+M)：，如果 y_i 爲 c_j 和 0 則 $Y_{ij}^0 = 1$，也表示標記(label)到達一個頂點，否則，讓 Y_L^0 等於 Y^0 前 l=R+N 列，這相當於被已被標記的資料；而 Y_U^0 保留 u=M 列，這相當於被未被標記的資料； Y_U^0 初始值根據文件相似度。

文件 x_i 和文件 x_j 間的權重，可以變爲一個機率表示方式：

$$t_{ij} = p(j \to i) = w_{ij} / \sum_{k=1}^{s} w_{kj}$$

t_{ij} 爲文件 x_j 和文件 x_i 的增殖(propagate)標記的機率。理論上，可以預期兩個文件跨不同叢集(class)的 w_{ij} 值盡可能的較小，而且同叢集 w_{ij} 值盡可能的較大。這個演算法，會對 u=M 和 l=R+N，計算聚合出一個解：

41

$$Y'_U = \lim_{t\to\infty} Y_U^T = (I - \overline{T}_{uu})^{-1}\overline{T}_{ul}Y_L^0$$

I 爲 $u \times u$ 矩陣，$\overline{T}_{uu}$ 和 $\overline{T}_{ul}$ 爲 $\overline{T}$ 矩陣的分割，如下：

$$\overline{T} = \begin{bmatrix} \overline{T}_{ll} & \overline{T}_{lu} \\ \overline{T}_{ul} & \overline{T}_{uu} \end{bmatrix}$$

因此 Y_U^0 的初始值不是這麼重要，而 Y_U^0 也不會對建立 Y'_U 有這麼重要(Yang et al., 2006)。

計算情形如下圖所示(Tseng, Tsai & Chuang, 2007)：

$$\text{Y(t+1) = T * Y(t) =>} \quad \begin{matrix} \\ d_1 \\ d_2 \\ \\ d_m \end{matrix} \begin{matrix} L_R & L_N \end{matrix} \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ M & M \\ y_{m1} & y_{m2} \end{bmatrix} = \begin{matrix} d1 & d2 & & dm \end{matrix} \begin{bmatrix} t_{11} & t_{12} & \Lambda & t_{1m} \\ t_{21} & t_{22} & \Lambda & t_{2m} \\ M & M & O & M \\ t_{m1} & t_{m2} & \Lambda & t_{mm} \end{bmatrix} \times \begin{matrix} L_R & L_N \end{matrix} \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ M & M \\ y_{m1} & y_{m2} \end{bmatrix} \begin{matrix} \} I_{R+N} \\ \\ \\ \} I_M \end{matrix}$$

I_{R+N}：相關和不相關文件列

I_M：待歸類文件列

L_R：相關

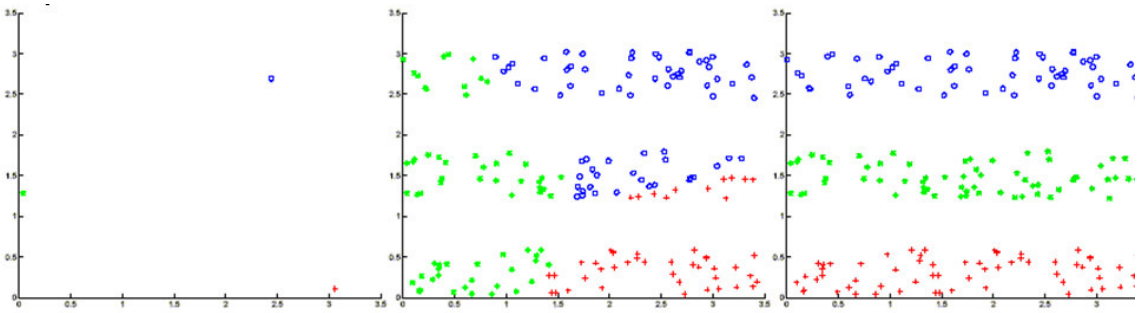
L_N：不相關

圖 2-16：Label Propagation 計算情況示意圖

資料來源：”On the Robustness of Document Re-Ranking Techniques: A Comparison of Label Propagation, KNN, and Relevance Feedback,” by Y. H. Tseng, C. Y. Tsai, and Z. J. Chuang, 2007, Proceedings of the Sixth NTCIR Workshop on Research in Information Access Technologies - Cross-Lingual Information Access.

簡單的說，此爲以 KNN(K-Nearest Neighbor)歸類法爲基礎所延伸的模式，其主要的差異在於迴圈計算，讓相對少的文件更加聚合(Tseng et al., 2007)。

而文件經由 Label Propagation 和 KNN 計算，如下圖所示可以較明確看出差異處(Zhu & Ghahramani, 2002)：



a. 資料　　b. KNN　　c. Label Propagation

圖 2-17：KNN 及 Label Propagation 歸類方式示意圖

資料來源：" Learning from Labeledand Unlabeled Data with Label Propagation," by X. Zhu & Z, Ghahramani, 2002, *CMU CALD technical report* CMU-CALD-02-107.

43

第三章 研究方法與設計

第一節 研究方法

一、實驗研究法：

實驗研究法（Experimental Method）是指研究者在妥善控制一切無關變項的情況下，操弄實驗變項，而觀察實驗變項的變化對依變項所產生的影響效果。

本研究使用「實驗研究法」進行，藉助實驗程序以發現因果關係或比較各個變項（Variables）之結果，進行多次的實驗並觀察實驗結果，尋求現象的因果關係。其基本原則爲藉著操控（X）變項以瞭解（Y）變項的改變情形，其基本原則如下：（葉至誠，2000）

（1） 當一變項（X）改變時，另一變項（Y）是否也會跟著變動。
在本研究的意義爲：「當較佳檢索機制運作時，而檢索成效是否提升」

（2） 及是否只有變項（X）改變時，才會造成（Y）變項的改變。
在本研究的意義爲：「若變更檢索成效如預期提升時，是否控制有其他變數」

二、實驗的控制方法

(一) 物理的控制

實驗者在實驗期間除需要注意使各組在實驗程序、實驗說明和實驗態度等方面相同外，還須考慮實驗情境的物理條件是否保持恆定、刺激的呈現是否標準以及反應的記錄是否客觀一致等。

(二) 選擇的控制

（1） 排除變項法

（2） 因子設計法

（3） 配對法

（4） 隨機法

(三) 統計的控制

當無關變項的控制不適用於使用上述的控制時，就必須採用統計的控制。通常採用共變數分析。統計資料常常受到各種因素的影響，而使個別個體的某些特徵發生變化，對這種影響因素所造成的變異的觀察與檢定的統計方法就稱爲變異數分析(Analysis of variance, ANOVA)。常用到的變異數分析方法有一因子完全隨機實驗、二因子未重複實驗、二因子重複實驗跟拉丁方格等方法。

三、實驗法之特徵：

（1） 實驗本身提供有意義的評估結果

（2） 同時運用必要的程序以控制已知的變異來源。

（3） 根據實驗方式以進行統計分析。

（4） 同一時間上實驗許多因素。

因此本研究將以實驗法進行，將實作各個檢索機制，對各個變項的變異方向與變異量進行瞭解，再實作於 NTCIR5 的測試集，收集各個變項的變化對測試結果所發生的效應。

45

第二節 研究流程與架構

本研究流程如圖所示：

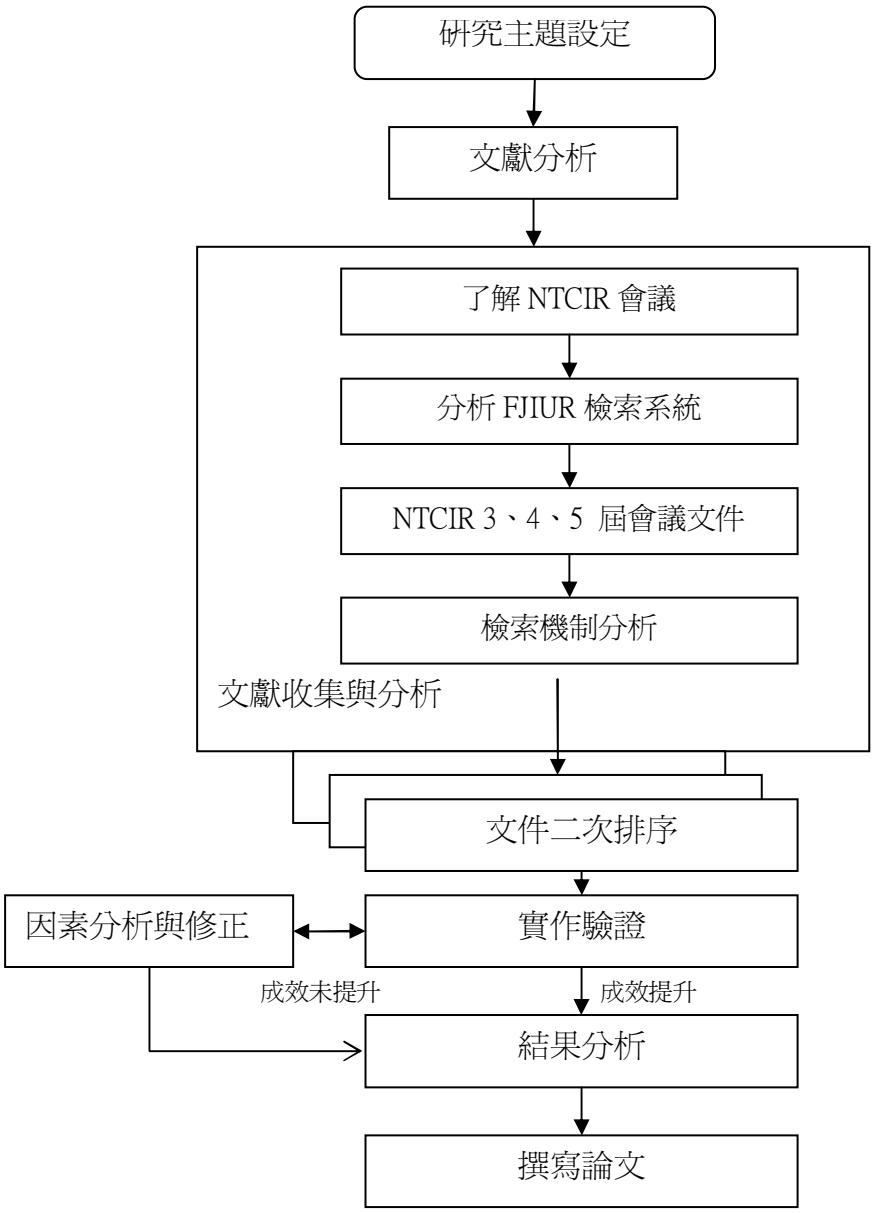


圖 3-1：研究流程圖

一、研究主題設定

此階段包含如何將研究主題確立及初步可行性分析，並確定文獻分析範圍及研究成果評估等…，在第一章節之中已描述。

二、文獻分析

文獻分析包含以下四個步驟：了解 NTCIR 會議、分析 FJUIR 檢索系統、NTCIR 3、4、5 屆會議文件

(1) 了解 NTCIR 會議

從 NTCIR 檢索會議的環境開始了解，例如參加的任務、使用的測試集、評估系統的方式等…。進而了解整個會議運作機制。

(2) 分析 FJIUR 檢索系統

爲了能達到本研究目的，更需要了解自身檢索系統的優缺點、實作方式等…。將從 FJUIR 的資訊處理流程探討：索引方式、檢索模式、查詢擴展等…。

(3) NTCIR 3、4、5 屆會議文件

收集 NTCIR 3、4、5 屆的會議論文，從前八名的優秀團隊使用的技術爲主要的研究對象，並從中擴充相關文獻。

(4) 檢索機制分析

最後綜合上面三點文獻，設計實驗問題，及初步實驗架構，將文獻整理成更適合於本研究環境。

47

三、文件二次排序

綜合文獻分析的結果後，本研究將研究焦點制定於文件二次排序的研究，希望能研究該機制是否能如其他研究一般，有效的提升成效。

四、實作驗證

了解文件二次排序的架構，將成效較佳的二次排序方式實作於本研究架構之中，便於實驗分析。

五、因素分析與修正

當實驗失敗時會有可能出現的因素需要控制，例如：「程式撰寫錯誤」、「公式參數錯誤」、「該機制不適用」或「需要先行完成某些處理」等…，待了解問題後，將進行修正，如果修正失敗、或是「該機制不適用」時，應更換「其他檢索機制」提升成效。

六、更換其他檢索機制

其他檢索機制定義如下：

(1) 該檢索機制具有研究價值。

(2) 方向在於二次排序技術的比較。

(3) 該機制已由其他團隊所實作，且實作成效爲較佳。

因此嘗試更換其他檢索機制，以完成提升檢索成效之目的。

七、結果分析

在出現預期結果後，將該結果細部分析，以便撰寫成論文。但如果出現非預期的結果，也將結果分享，探討那環節出現問題。

49

第三節 研究設計

一、自動建構索引

選擇新聞文件作爲評估的資料集，主要是取決於新聞文件內容主題豐富且符合現實生活的環境，其評估出來的結果可信度較高。

本次實驗主要採用的訓練資料集爲日本 NTCIR 會議，由其中的單語檢索(CLIR 的中文資訊檢索測試集 1.1 版（CIRB011, Chinese Information Retrieval Benchmark version 1.1）和 2.0 版（CIRB020, Chinese Information Retrieval Benchmark version 2.0）；而 CIRB011 皆下載自五個新聞網站於 1998 年 5 月至 1999 年 5 月間的報導，這些新聞包括：中國時報、工商時報、中時晚報、中央日報以及中華日報。CIRB020 則下載自聯合新聞網站於 1998 年 1 月至 1999 年 12 月底的報導。而到了 NTCIR5 使用 4.0 版(CIRB040r)，爲 2000 年到 2001 年的聯合報、經濟日報等…，如下表所示：

表 3-1：

CIRB040r 文件集收錄資料來源

來源	數量		
CIRB040r (581.7 MB)	2000	2001	總計
聯合報 (udn)	244038	222526	466564
United Express (ude)	40445	51851	92296
民新日報 (mhn)	84437	85302	169739
經濟日報 (edn)	79380	93467	172847
總計	448300	453146	901446

資料來源：「跨語言資訊檢索與擷取測試集」，陳信希、陳光華，2005，民國九十六年六月四日，取自：http://www.csie.ntu.edu.tw/~ciet/form/paper/1.doc

FJUIR 團隊建構索引的方式爲，威知資訊公司所提供之 WebGenie 試用版軟體建立索引，再以 Perl 撰寫程式匯出關聯詞。

首先必須先將測試文件集轉換成資料庫之格式，由於測試文件集本身是 XML 標記格式，無法直接使用 WebGenie 來建立索引，因此必須先撰寫轉檔程式，將 XML 以 Regular Expansion 的方式轉換成 CSV 檔案格式並匯入資料庫，其欄位可分爲(蔡育欽，2005)：

1. DocID 代表文件編號
2. DocDate 代表文件日期
3. DocTitle 代表文件標題
4. DocContent 代表文件內容

二、控制組檢索系統設計

設計本實驗的控制組，有效的提升實驗控制組的成效爲本研究目標。而實驗控制組系統也參考過去使用的基礎檢索系統。使用威知公司(WebGenie)開發的系統建置索引，索引將一字詞、二字詞和關鍵字存於反向索引檔中，並計算詞頻、文件平均長度等…，供 BM25 檢索模組計算文件和查詢詞的相似度，架構如下所示：

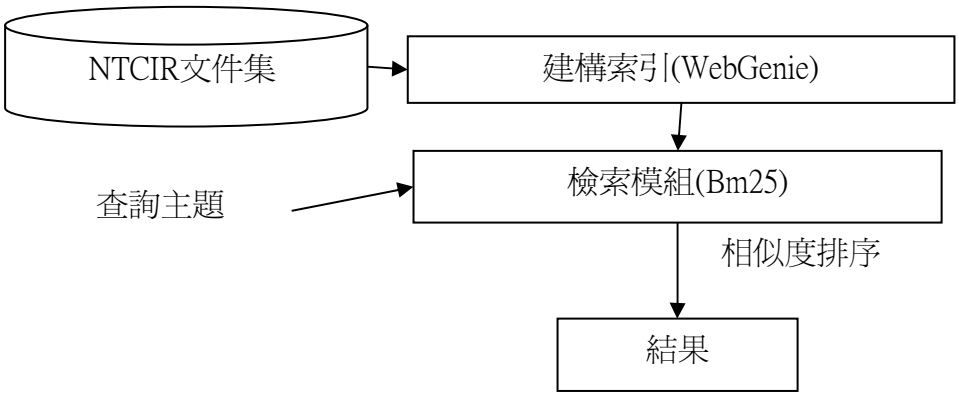


圖 3-2：控制組檢索系統

51

三、實作文件二次排序

I2R 團隊所提出的「文件二次排序」機制及「歸類二次排序」機制，爲本實驗最主要研究的對像，實驗的機制主要有兩種，「I2R 文件二次排序」及「HKPU 標題二次排序」。

(一) I2R 文件二次排序方式

（1） 查詢詞 q 中取得詞彙，和其權重

（2） 重新建構上述 1000 篇文件索引

（3） 收集出現在 q 的詞彙，並且用下列公式計算詞彙 t 權重

$$\sqrt{\frac{(\sum_{i=1}^{1000} df\ (t,d_i)\times f(i))/1000}{DF\ (t,C)/R}}\times\sqrt{|t|}$$

$$df\ (t,d_i)=\begin{cases}1 & t\in d_i\\ 0 & t\notin d_i\end{cases}$$

- d_i=第 i 篇文件
- R=文件集 C 中的全部文件數
- df(t,d) ,df(t,C) =爲 d,C 文件出現與否
- |t|=關鍵字長度
- f(i)=1+1/Sqrt(i),爲 d_i 的權重

（4） 前 1000 篇中每一篇文件 d_i 用下列公式計算其新的排序值 Si，Ri 爲初始相似度。

$$S_i = R_i\times(1+\sum_{t_j\in q,d_i} W(t_j))$$

（5） 重新排列前 1000 篇文件，依照$\{S_1,S_2,.....S_i...,S_{1000}\}$

架構如下：

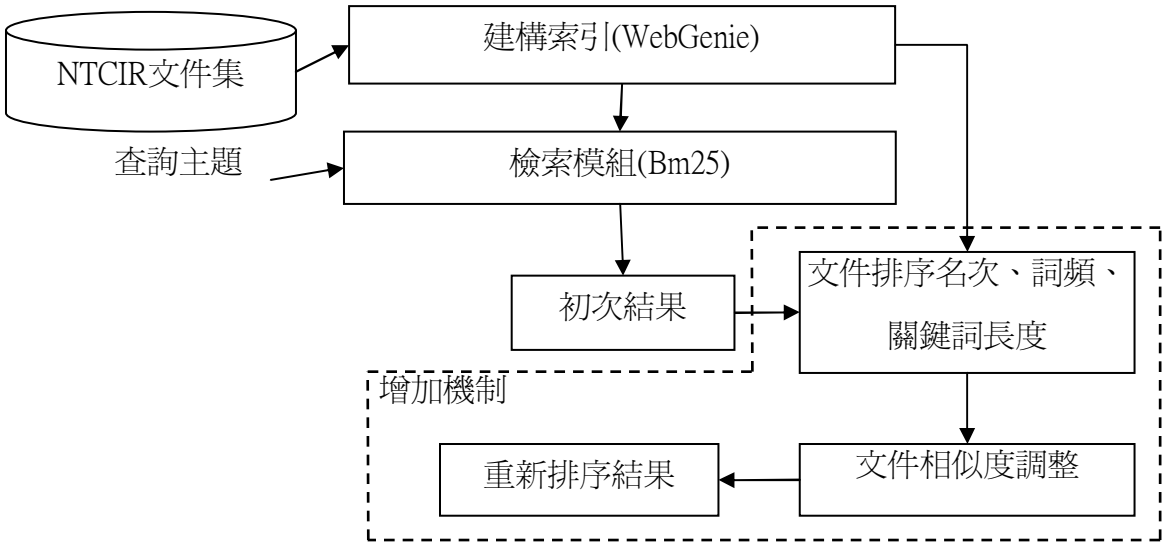


圖 3-3：I2R 文件二次排序架構圖

而 I2R 研究團隊於 NTCIR4 所提出的「文件涵蓋率(coverage)」、「事件偵測(event detection)」方式，考慮到：

（1） 文件涵蓋率：I2R 研究團隊未提出明確做法

（2） 事件偵測：針對「新聞」或具有「時間代表性」的文件較有意義，較非一般文件集所有，比較不易跨其他文件集。

所以未加以實作，還是以主要的二次排序方式爲本研究實作的目標，假使該團隊所提文件二次排序方式實驗有達顯著的成效提升，便可以加以研究這部分的理論。

(二) HKPU 標題二次排序

本實驗完全參考 HKPU 的方式，以提升研究成效，

文件二次排序技術公式如下：

$$sim'(q,d_i) = \left(sim(q,d_i) - m\right) \times M(q_t, t(d_i)) + m$$

- sim(q,d_i)：相似性計算方式以 BM25 計算後的相似度運算。
- m：前 n 篇文件中最小的原始相似分數

53

- $t(d_i)$：第 i 篇文件的標題
- q_t：符合查詢 q 的文件標題
- M(.)：查詢標題和文件標題符合的詞數

實驗架構如所示：

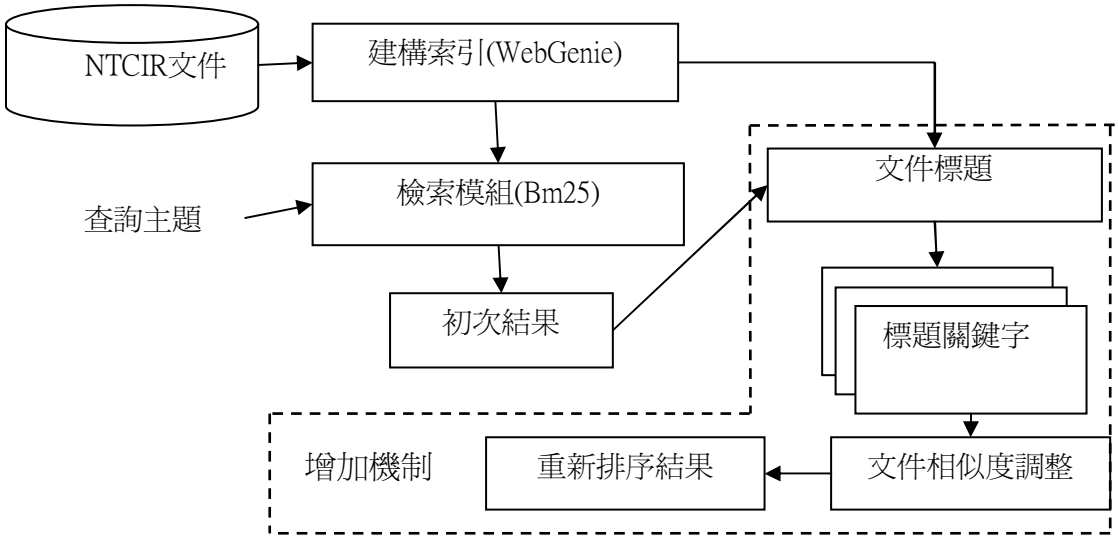


圖 3-4：標題文件二次排序架構圖

四、歸類方式調整排序

試著了解歸類方式是否能有效的提升成效，如 I2R 研究團隊所實驗的情況一般，因此以「Label Propagation 歸類模組」爲實作模組，並也實作「KNN 歸類模組」試著比較成效。

(一) Label Propagation 歸類模組

由於 KNN 需要較多的文件資訊才能有較精準的分類特性，因此使用 Label Propagation 來進行少量文件的歸類，Label Propagation 爲 KNN 模組再進行多次計算而加強特徵的方式，因此會和 KNN 實作的方式差異不大，以下爲 Label Propagation 演算法：

輸入：T, Y(t=0), *l* 和 Error_threshold
Repeat
（1） Y(t+1) = T * Y(t)
（2） 正規化 Y(t+1)列，保留機率分布情況，

例如：

$$y_{ij} = \frac{y_{ij}}{sy_i}, sy_i = \sum_{k=1}^{n} y_{ik}$$

（3） Error = |Y(t+1) - Y(t)|
（4） Y(t) = Y(t+1) 讓第一個 Y(t)的 *l* 列成爲 Y(0)
直到 Error < Error_threshold
輸出： Y(t) //最後 *m-l* 列值爲結果

圖 3-5：Label Propagation 歸類演算法

設計本實驗的實驗架構如下：

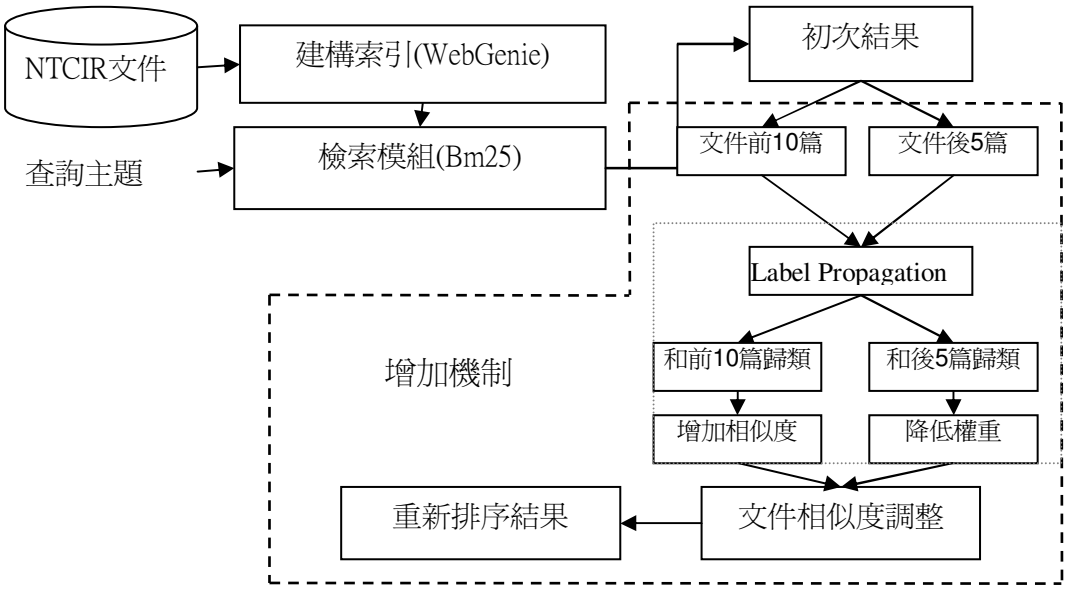


圖 3-6：Label Propagation 歸類模組進行文件二次排序

(二) KNN 歸類模組

在初次排序之後，將排序於前最佳前 10 篇設定爲「相關訓練文件組」，而後 5 篇文件設定爲「不相關訓練文件組」，將其他的文件進行 KNN 歸類，和相關文件歸類在一起的文件，也可以視爲相關，而和不相關文件歸類一起的文件，也可以視爲不相關。依此假設調整相似度。

本實驗的演算法如下所示：

55

輸入： T, Y　　// m-1 列 Y 的最小值設爲 0

執行：

1. Y(t+1) = T * Y(t)
2. 正規化 Y(t+1)列，保留機率分布情況,例如：

$$y_{ij} = \frac{y_{ij}}{sy_i}, sy_i = \sum_{k=1}^{n} y_{ik}$$

輸出： Y(t)　//最後 *m-1* 列值爲結果

圖 3-7：KNN 歸類演算法

設計系統架構圖如下所示：

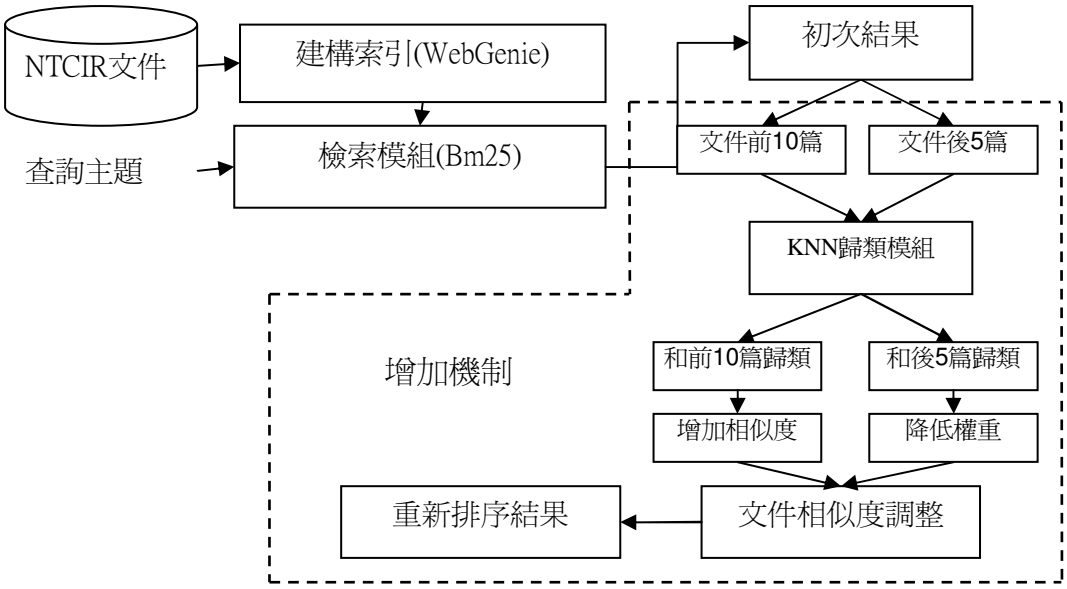


圖 3-8：KNN 歸類模組進行文件二次排序

實驗結果和 Label Propagation 相互比較，以了解 Label Propagation 是否能較 KNN 佳，並且和 I2R 研究團隊實驗呼應。

五、PRF 查詢詞擴展

最後一部分，也將過去常使用「PRF 查詢擴展」，和其他實驗機制比較，PRF 查詢擴展本實驗以「初次查詢前六篇文件每篇十五個詞，再次送入查詢」，架構如下所示：

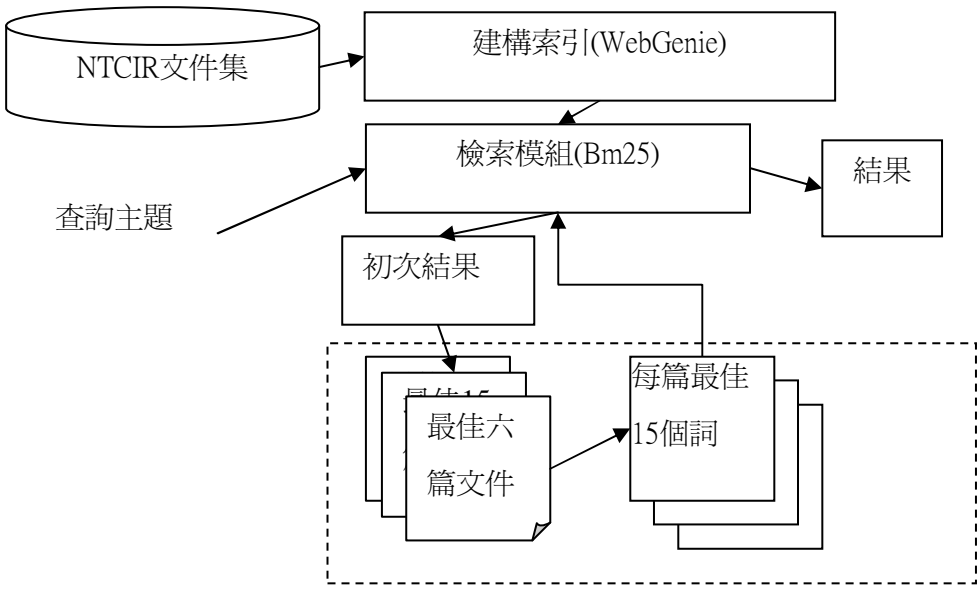


圖 3-9：PRF 查詢擴展實驗架構圖

六、成效評估與分析

成效評估方式爲將各式檢索機制組合，運用於 NTCIR5 測試集之中，最後取 TREC_EVAL 中的平均精確率(MAP)爲主要的評估依據。

將檢索系統查詢後的結果，重新排序，此階段可以獨立運作。由於送入 TREC_EVAL 有一定的格式，因此可以使用該格式，來重新排序，不必另外使用檢索系統。呈現方式如下顯示結果所示：

```
001 0 edn_xxx_20001109_0623629     0       310.8986 NTNU-C-C-D
001 0 edn_xxx_20011110_1179661     1       268.4390 NTNU-C-C-D
001 0 udn_xxx_20011110_1180027     2       262.6515 NTNU-C-C-D
001 0 udn_xxx_20000229_0187984     3       247.5701 NTNU-C-C-D
001 0 edn_xxx_20001109_0623629     4       230.2578 NTNU-C-C-D
001 0 udn_xxx_20000229_0187985     5       220.7096 NTNU-C-C-D
001 0 edn_xxx_20011112_1182048     6       202.5672 NTNU-C-C-D
001 0 mhn_xxx_20011110_1179654     7       199.8533 NTNU-C-C-D
001 0 udn_xxx_20001106_0618911     8       197.1457 NTNU-C-C-D
001 0 udn_xxx_20011107_1174932     9       194.3979 NTNU-C-C-D
001 0 udn_xxx_20011110_1180020     10      193.6735 NTNU-C-C-D
001 0 edn_xxx_20011110_1179661     11      191.3111 NTNU-C-C-D
001 0 udn_xxx_20011110_1180027     12      188.3212 NTNU-C-C-D
001 0 edn_xxx_20011109_1177963     13      186.0304 NTNU-C-C-D
```


57

```
001 0 mhn_xxx_20000310_0051029     14      181.7117 NTNU-C-C-D
001 0 udn_xxx_20011112_1182226     15      181.1063 NTNU-C-C-D
001 0 udn_xxx_20000516_0321853     16      178.7239 NTNU-C-C-D
001 0 udn_xxx_20010430_0880271     17      177.2269 NTNU-C-C-D
001 0 udn_xxx_20000519_0437490     18      174.6421 NTNU-C-C-D
001 0 udn_xxx_20000412_0095138     19      174.2473 NTNU-C-C-D
 ..............
```

圖 3-10：檢索系統查詢結果

爲了有效的了解成效提升的情況，評估計算方式使用 FJUIR 於 NTCIR4 的平均精確率增加百分比(Percentage improvements in MAP)計算方式，公式如下：

$$imp = \frac{b-a}{a} \times 100\%$$

a：爲控制組平均精確率，在此爲基礎檢索系統。

b：爲實驗組平均精確率，在此爲實作檢索機制後的檢索系統。

由於檢索系統之平均精確率差異性不大，在觀察許多實驗中，若平均精確率提升 5%左右，將可能達到顯著的差異，顯著差異的發生，也可以看出是否該方式能有效的提升成效。爲了更明顯了解效能是否有提升，使用上述 imp 公式精確控制。以現階段的實驗情況及經驗判斷，本階段實驗將 imp 差異於 5%定義爲顯著差異，分成以下可能：

（1） imp>5%：相關文獻研究結果成立，表示該檢索機制在一定的實驗環境及相同的檢索機制上，能有很好的成效表現，

（2） imp<-5%：成效下降，相關文獻研究結果不成立，在查證程式無誤後，將回顧相關文獻分析差異因素，探討分析問題主因。

（3） -5%<imp<5%：成績差異不大，視爲未提升，在查證程式無誤後，分析原因，認爲些檢索機制無法提升檢索成效。

第四章 實驗結果與分析

本實驗所使用的檢索系統，使用威知公司(WebGenie)開發的系統建置索引，索引將一字詞、二字詞和關鍵字存於反向索引檔中，並計算詞頻、文件平均長度等…，供 BM25 檢索模組計算文件和查詢詞的相似度，爲本實驗「控制組」的實驗操作方式

實驗文件集以日本 NTCIR 會議提供的文件集爲主，使用版本爲 NTCIR 第五屆的 CIRB040 文件集、檢索問題及答案集。文件集共有 901446 篇文件，文件平均長度爲 587.7 個字，檢索問題 50 題。查詢問題如附錄 A 所示。

使用 WebGenie 系統建立索引共花費時間 113 分 58 秒，建出 1160209 字，使用 BM25 檢索模組計算文件相似度後的結果，如下表所示：

表 4-1：

基礎檢索系統平均精確度

	Rigid(MAP)	Relax(MAP)
Title	0.2691	0.3229
Description	0.2379	0.2912

以下實驗依此架構，另外加上能增加成效的機制，第一節的設計是「使用查詢詞爲主，對初次結果文件進行二次排序」，第二節爲以「以排序後的文件相似情況進行二次排序」，最後進行綜合分析。

59

第一節 以查詢詞特徵調整相似度

本實驗依照 NTCIR 論文中，二次排序機制成效較佳的團隊，有 I2R 團隊將文件內容和查詢詞作符合度計算，HKPU 團隊將文件 TITLE 和查詢詞的符合度計算，本研究依該團隊於論文中所發表的公式進行實作：

一、I2R 文件二次排序

依照初次檢索結果；詞頻、詞長、來調整初次檢索結果的排序，實作於本系統中。以 TREC_EVAL 評估方式，計算文件平均精確率，成效如下：

表 4-2：

I2R 文件二次排序實驗結果

	欄位	Rigid	Relax
原架構	Title	0.2691	0.3229
	Description	0.2379	0.2912
使用後	Title	0.2486	0.3098
	imp	-7.6%	-4.1%
	Description	0.2027	0.2606
	imp	-14.8%	-10.5%

使用後公式重新計算權重後，成效沒有增加，反而降低。降低的程度以 Description 查詢欄位較爲嚴重，可以達到 10%以上。由於共有 50 題查詢問題，發生的問題的情況難以逐一分析，但查看每一查詢問題所 MAP 增加的情況，可大致觀察了解成效降低分布情況，如下圖所示：

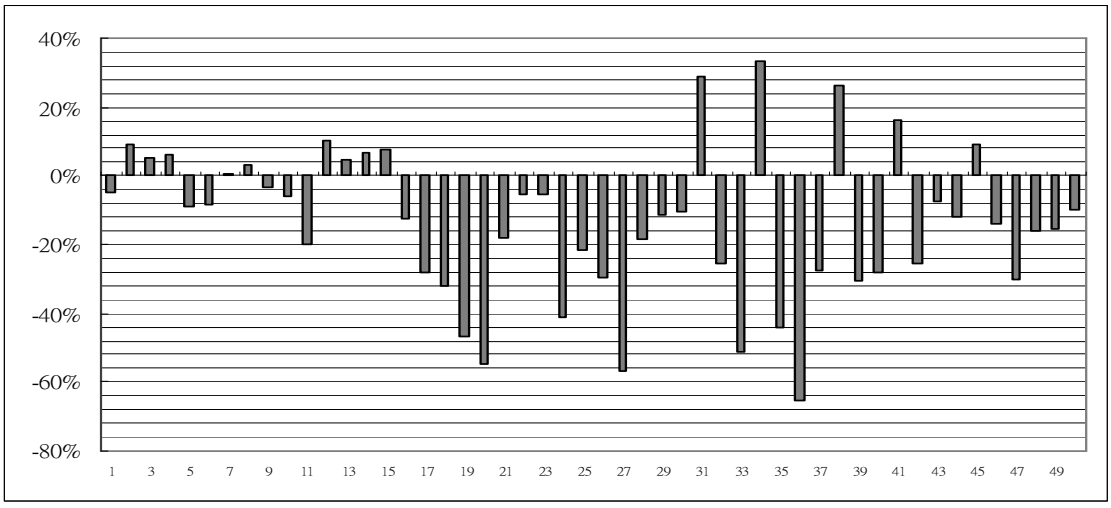


圖 4-1：文件二次排序各題查詢問題成效提升(imp)圖(Description-run rigid)

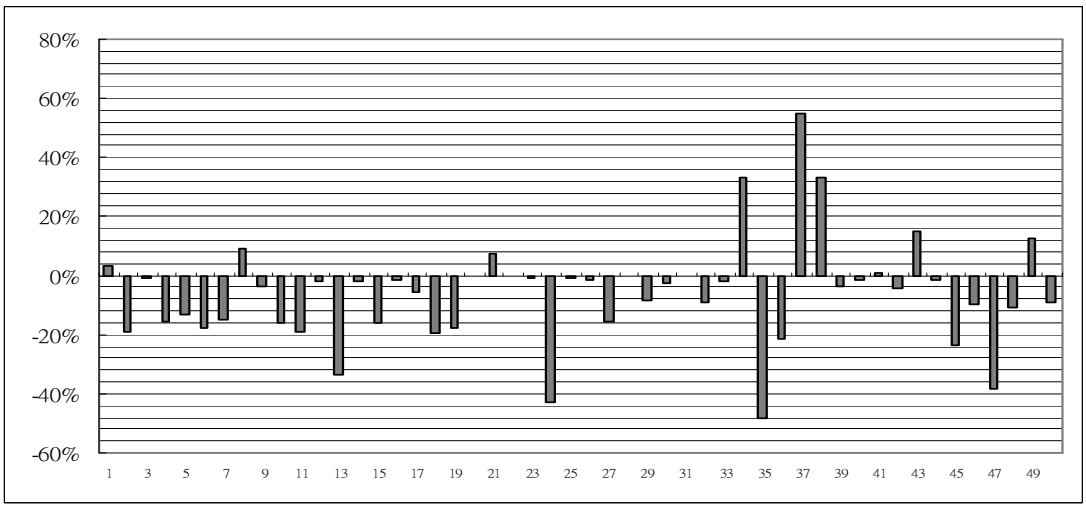


圖 4-2：文件二次排序各題查詢問題成效提升(imp)圖(Title-run rigid)

由上圖可以看出，查詢欄位 Description 有 14 篇文件成效有所提升，查詢欄位 Title 有 8 篇文件成效有所提升，占爲少數，大多數的題目使用此二次排序機制調整相似度後，都產生成效下降的情況。初步從公式上分析可能的問題：

$$W_t = \sqrt{\frac{(\sum_{i=1}^{1000} df(t,d_i) \times f(i))/1000)}{DF(t,C)/R}} \times \sqrt{|t|}$$

61

$$df(t,d_i)=\begin{cases}0 & t\in d_i\\ 1 & t\notin d_i\end{cases}$$

$$Sim_i'=Sim_i\times(1+(W_{t1}+W_{t2}+...W_{tn})$$

從文件排名的變化和調整後的權重來看，如下圖所示：

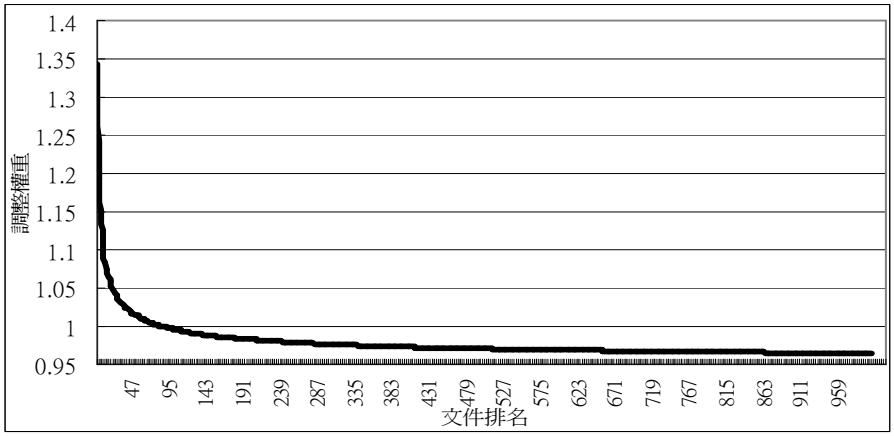


圖 4-3：文件排序 f(i)對權重的影響(假設該詞文件數爲 1000)

排序在前面的文件受到影響的範圍比排序於後的要大，大約排名於 50 名之後的文件就較沒有受到名次的影響。但相較於初次排序的名次，公式中 DF(t,C)(某查詢詞所能找到的文件數)的調整幅度就比較大，如下圖所示，名次的重要性相較於 DF(t,C)的重要性爲低，DF(t,C)約在 60 以上的詞。

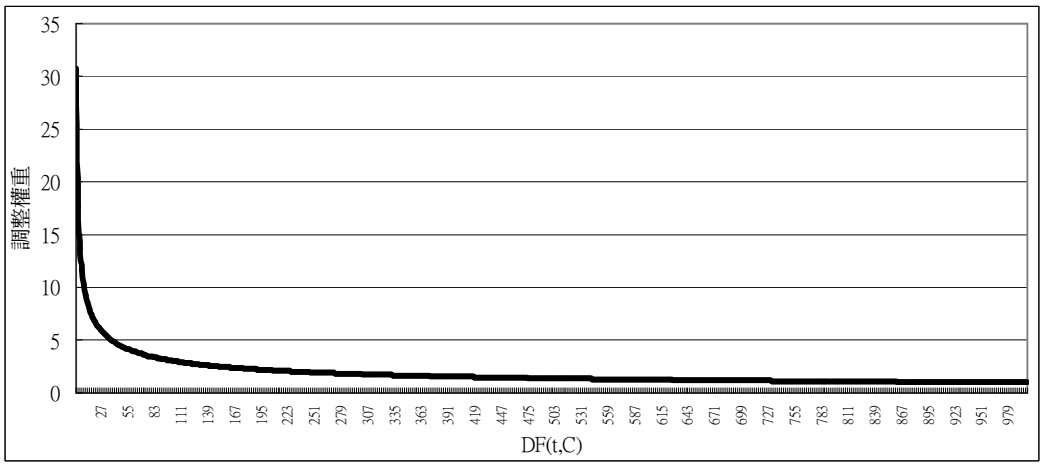


圖 4-4：DF(t,C)對權重的影響(假設排序於第 500 篇文件)

權重調整的公式可以看出，該調整方式以「長字串」、「排名於前 50 名之文件」、「DF(t,C)<100」的文件有非常高的權重。進一步觀察，相較於「文件排序」、「字串長度」而言，「DF(t,C)」的值對整個公式影響較大，可以差至 30 倍以上。雖然理論上，「DF(t,C)」和該詞的重要性多呈反比，但若產生特例情況，易使不相關文件會被大幅度提升，例如第 50 題「人類基因(DF=637)，解碼(DF=882)，醫藥業(DF=49){斷出另一詞：藥業(DF=4825)}」，「醫藥業」雖爲文件數較少，只有 49 及爲其爲 3 字詞，因此權重值較大，將許多有關「醫藥業」，不相關的文件提升，如二次排序後，非相關的文件標題爲「生技股受股舞」，相似度從到 64.4466 提升到 1186.5978，因此公式的計算方式可能需要進一步修正。

像 Title 欄位檢索第 19 題「超音速飛機，協和號，墜機」，很容易將有關「超音速飛機」的文件提升不少，但和此有關的非相似文件，例如一篇文件「超音速飛機 - 下月處女航 - X—43A 時速超過五千哩」的談論有關「超音速飛機」，其相似度計算從原本的 72.9674 加權到 4341.1301，在此也看出相似度的修正方式也缺少一個最大範圍值，使原本的相似結構受到大幅改變。因此詞的選擇，能大幅影響文件權重，反之能也能大幅降低成效，使得此調整方式過於敏感。

二、HKPU 標題二次排序

將有標題出現查詢問句關鍵詞的文件，調整該文件權重。實驗後，以 TREC_EVAL 評估方式，計算文件平均精確率，成效如下：

表 4-3：

HKPU 標題二次排序實驗結果

欄位		Rigid	Relax
原架構	Title	0.2691	0.3229
	Description	0.2379	0.2912
使用後	Title	0.2027	0.2606
	imp	-24.3%	-26.2%

63

	Description	0.1995	0.2512
	imp	-25.9%	-22.2%

重新計算後的排序，成效也是大幅度的降低，降低的程度較 I2R 文件二次排序嚴重，將每一查詢問題所增加的情況，展開分析；下圖爲每一個題目 imp 成長情況：

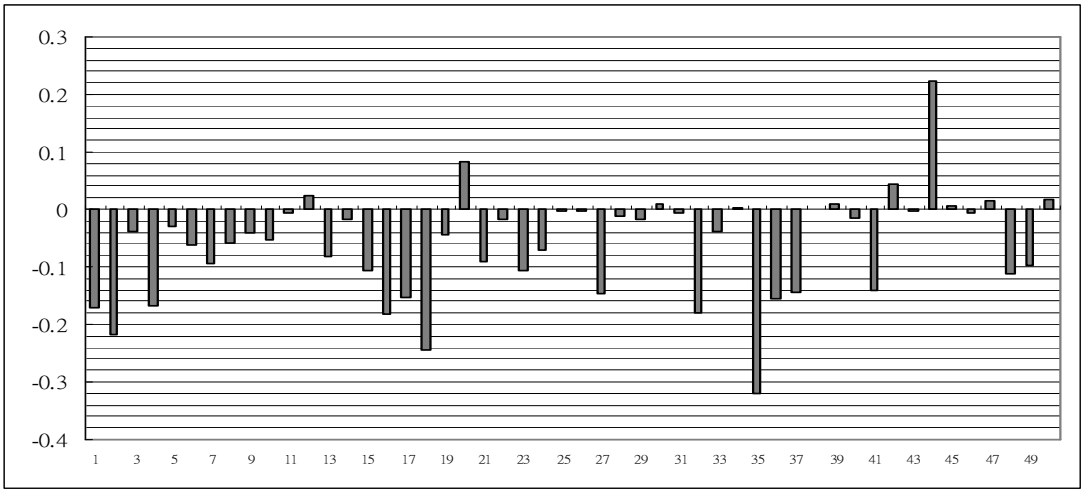


圖 4-5：標題二次排序各題查詢問題成效提升(imp)圖(Description-run/rigid)

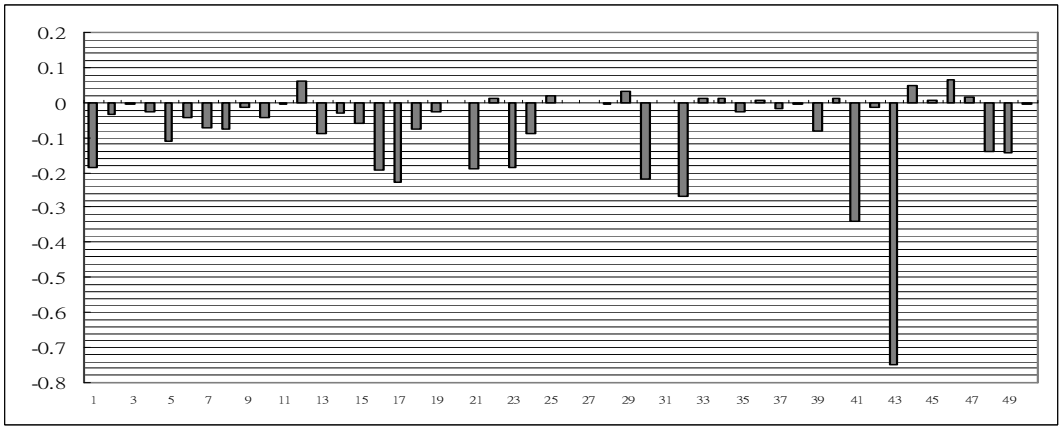


圖 4-6：標題二次排序各題查詢問題成效提升(imp)圖(Title-run/rigid)

從上圖所示，大多數的題目產生文件成效下降，由調整權重的公式探討：

$$sim'(q,d_i) = (sim(q,d_i) - m) \times M(q_t, t(d_i)) + m$$

在此調整方式中，文件直覺性的乘上「文件標題次數」來調整權重，也因此容易造成權重調整過當等情況，觀察排序前後的情況，有以下四個情況可以了解爲何會降低權重：

(1) 文件標題包含直接且明確的查詢詞，因此，假設一篇正確文件內將查詢詞以其他方式敘述，或是使用別的詞表示，甚至是縮寫詞，都不易被正確的加權調整。例如：在 Title 查詢第 1 題題目爲「時代華納，美國線上，合併案，後續影響」中，一篇正確答案的文件題目爲「雙劍合璧 有利拓展亞洲市場」沒有出現任何查詢詞，從原本的 29 名調整到 171 名。Description 查詢第 43 題「俄羅斯批准 STAR2 的背景。」，而一篇正確答案的標題爲「普亭：俄國願談判裁減軍備 - 新總統強調 努力說服下院通過第二階段限武條約 但也將提升核武嚇阻能力 - 烏拉山之行，討論核工業」也沒有出現查詢詞，從初次查詢第 40 名擠到排名 370 名。

(2) 除了包含查詢詞，且具有副標題的文件。某些新聞標題會下副標題，但不是全部的新聞都有下副標題的補述說明方式，常容易和查詢詞符合，但不代表正確相關的文件也都有副標題，在此情況下，正確相關的文件，常被其他文件擠到排名於後。

(3) 無法由符合個數來確定相關程度。即使文件標題和查詢詞符合，還是有出現許多偏差的提升情況產生。例如：Title 查詢 1 題「時代華納，美國線上，合併案，後續影響」，而不相關文件標題「嘉禾宏網共建亞洲最大娛樂入門網站 - 成立易看聽及子網站一按來 雙方期許媲美時代華納與美國線上合併案」反而從 55 提升到 7 名。

(4) 由標題較不易得知正確內文意思。即使文件標題和查詢詞符合，但實質內容非爲正確文件。例如：Description 查詢第 18 題「菸草商遭受控告並求賠償金之相關報導。」而一篇文件爲「賠償金談不攏 家屬抬棺抗議 - 負責公司派代表與地方民代再次協調 確定賠償金額後結束紛爭」也從第 83 名提升到第 4 名。

65

第二節 以初次查詢排序調整相似度

將初次結果文件，依照其文件的相似情況，也就是文件初次結果排序進行調整；利用前幾篇文件具高相似度的資訊，調整排序於後的文件。本實驗實作「KNN 歸類模組二次排序」、「Label Propagation 歸類模組二次排序」、「PRF 查詢擴展」予以評估。

一、Label Propagation 歸類模組二次排序

使用迴圈計算 KNN 歸類法，將少量訓練文件集特徵再予以明確計算，實驗結果如下所示：

表 4-4：

Label Propagation 歸類二次排序實驗結果

欄位		Rigid	Relax
原架構	Title	0.2691	0.3229
	Description	0.2379	0.2912
使用後	Title	0.2520	0.3060
	imp	-6.4%	-5.2%
	Description	0.2359	0.2929
	imp	-0.8%	0.6%

使用歸類方式調整排序，也未能提升成效，即使查詢欄位 Description 有些許提升，但也達不到顯著情形，因此 Label Propagation 歸類二次排序在本實驗中視爲「無法提升成效的機制」，進一步對每一查詢問題觀察成長情況：

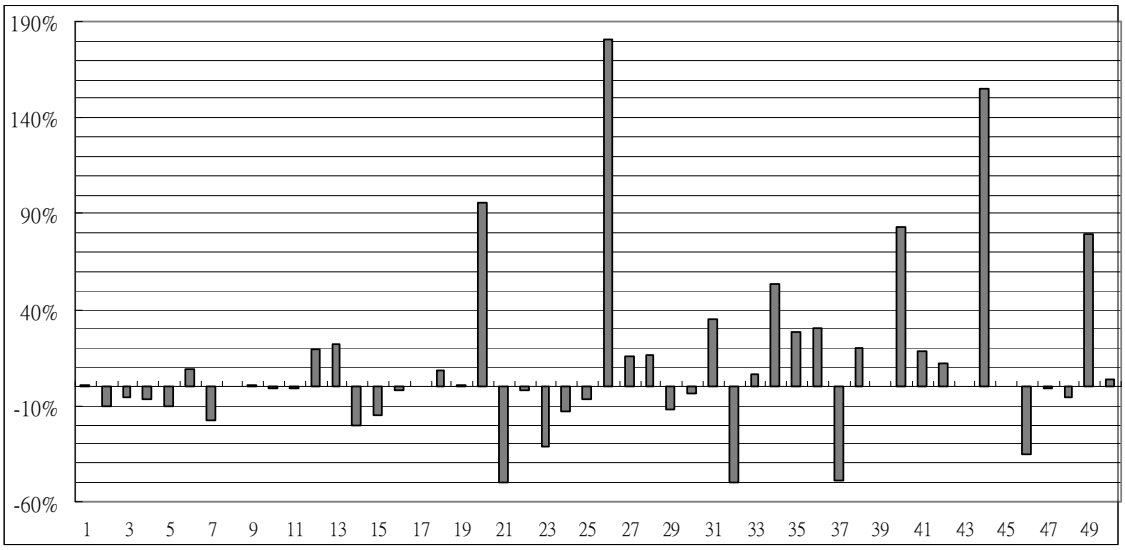


圖 4-7：Label Propagation 歸類二次排序各題查詢問題成效提升(imp)圖

(Description-run/rigid)

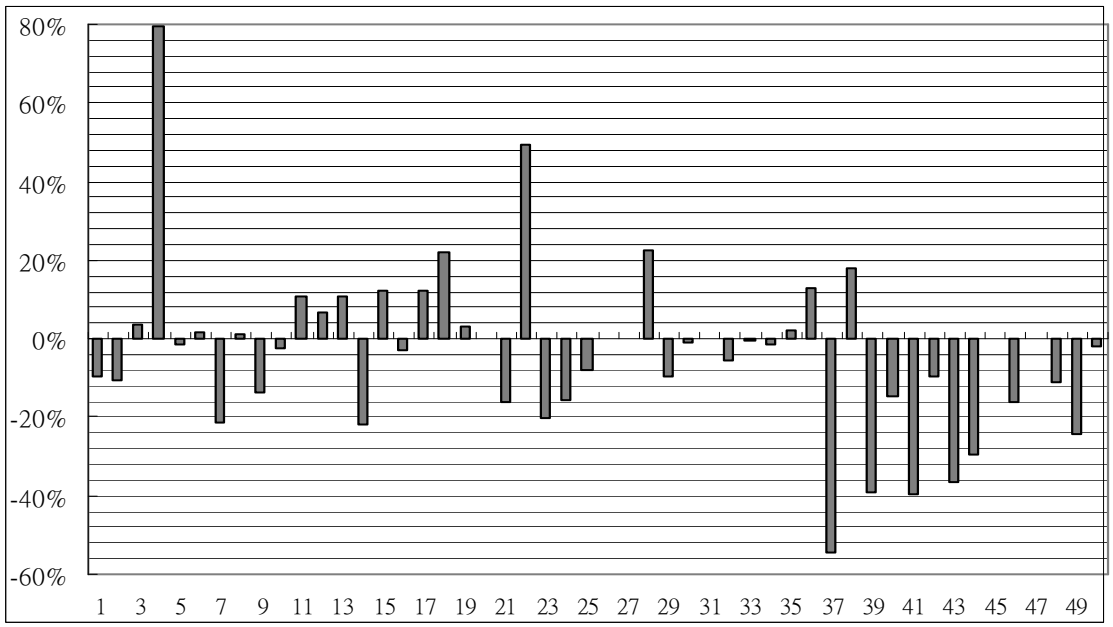


圖 4-8： Label Propagation 歸類二次排序各題查詢問題成效提升(imp)圖

(Title-run/rigid)

以 Description 查詢欄位來說，共有 23 題成效有所提升，雖然查詢問題 26 題成效提升較大，但由於原查詢 MAP 爲 0.0122，提升到 0.0343，對整體成效差異不大，不列爲評估。而查詢問題 44 題，初始查詢爲 0.0886，卻可以提升到 0.2256，可以算是成效提升最大的題型；成效下降程度方面，以查詢問題 21、32 題較爲嚴重。

67

Title 查詢欄位而言，共有 16 題成效有所提升，以查詢問題第 4 題成效提升較大(0.1109 -> 0.1992)，以查詢問題 37 題下降程度較大(0.0844 -> 0.0383)及 41 題(0.3482 -> 0.2108)。

二、KNN 歸類模組二次排序

本實驗以排序於前 10 篇文件，假設爲相關文件，而排序於後的 5 篇文件，假設爲不相關文件，使用 KNN 歸類模組，進行歸類，將歸類後的文件進行相似度調整。藉以希望能再次找到排序於後相似的文件，以提升成效，而歸類的模組使用 KNN，是爲了和 Label Propagation 有所比較。Label Propagation 和 KNN 的架構相似，差異在於 Label Propagation 的迴圈計算。實作後實驗結果如下所示：

表 4-5：

KNN 歸類文件二次排序實驗結果

欄位		Rigid	Relax
原架構	Title	0.2691	0.3229
	Description	0.2379	0.2912
使用後	Title	0.2429	0.3000
	imp	-9.7%	-7.1%
	Description	0.2336	0.2906
	imp	-1.8%	-0.2%

在此看出成效也是沒有提升的，多呈現下降的情形，其中以查詢欄位爲 Title 的使用 KNN 的方式下降情況較嚴重，和 Label Propagation 類似，但比較起來，以 Label Propagation 下降程度較小，應可視爲 Label Propagation 的計算方式有一定的成效，但在本實驗中相差不大。每一個查詢問題來看，如下圖所示：

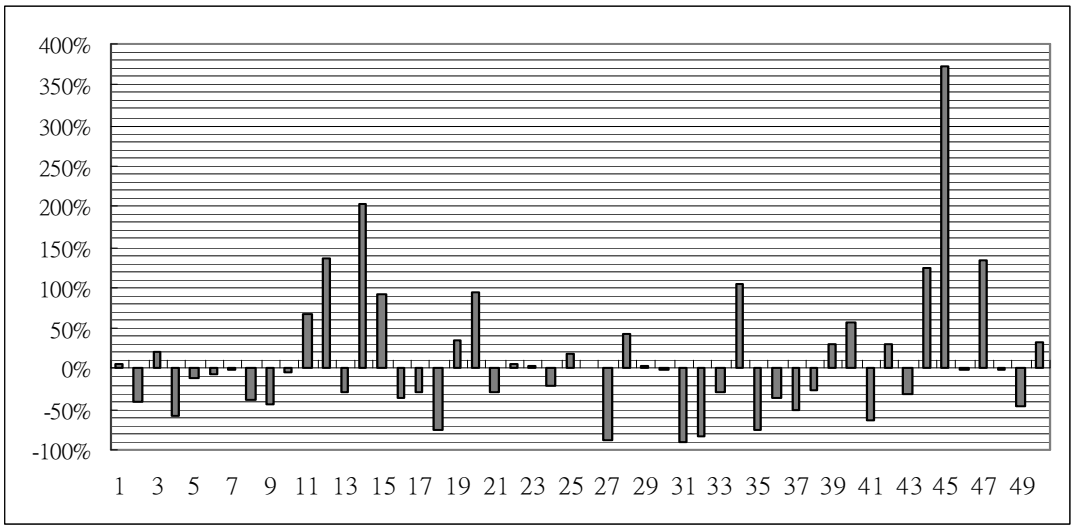


圖 4-9：KNN 歸類文件二次排序各題查詢問題成效提升(imp)圖
(Description-run/rigid)

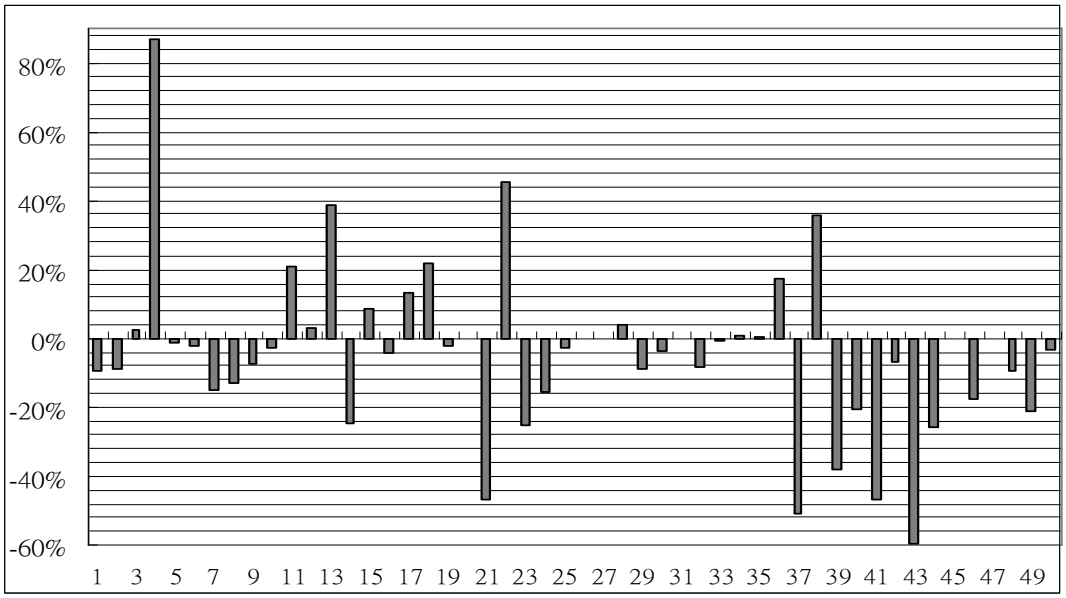


圖 4-10：KNN 歸類文件二次排序各題查詢問題成效提升(imp)圖(Title-run/rigid)

查詢 Description 欄位而言，查詢問題 45 題成長較大，但 MAP 從 0.0011->0.0052 不具代表性，因以第 14 題成效提升最大(0.1572 -> 0.4777)，雖以第 31 題下降幅度較大，但也是從 0.003 下降到 0.0022 差異不大，因此以第 27 題爲下降程度最大的題型(0.1689 -> 0.0213)。

查詢 Title 欄位而言，以第 4 題成效提升最多(0.1109-> 0.2073)，以第 43 題成

69

效下降最多(0.7556 -> 0.3048)，也可以看出和 Label Propagation 方式所分布的題型有所差異，如下表所示：

表 4-6：

KNN 及 Label Propagationt 查詢題型比較(Rigid)

成效		KNN		Label Propagation	
		Title	Description	Title	Description
提升最大(題號)	imp 差	4	31	4	44
	MAP 差	22(+0.1292)	14(+0.3205)	22(+0.1403)	44(+0.137)
下降最大(題號)	imp 差	43	27	37	21、32
	MAP 差	43(-0.4508)	18(-0.3201)	43(-0.2778)	23(-0.1635)

從 MAP 差異值初步判斷，可以看出 Label Propagation 對於不論是成效提升下降，都能較 KNN 保持在一定的範圍。另外，在查詢 Description 欄位時，這兩個方法所提升/減低成效的情況差異較大。

進一步觀察，可以發現即使兩兩文件相似度高，被歸類一起，但多爲一篇被判定爲相關，而另一篇卻沒有，如 Title 查詢詞「胚胎幹細胞」，如下表所示：

表 4-7：

歸類一起文件非同爲相關文件

相似度	共現詞	不相關文件	相關文件
0.760808	(體 外 試 驗, 造 肝 細 胞, 動 物 試 驗, 臍 帶 血, 臍 帶)	台大嘗試臍帶血培養造肝細胞 - 培養幹細胞 目前進行動物、體外試驗 近 ...	幹細胞造肝 台大展開動物試驗 - 若成功 肝患只要注射臍帶血 不需等器捐
0.748800	(葉 士 芃, 妹 妹, 造 血 幹 細 胞, 再 生 不 良 性 貧 血, 姊 妹)	姊姊貧血 妹妹捐造血幹細胞 - 二十六歲朱恆慧移植順利回院複檢 姊妹 ...	造血幹細胞移植 患者獲新生 - 姐妹情深
0.739733	(伊 薩 科 森, 巴 金 森 氏 症, gdnf, 麥 凱, 動 物)	巴金森氏症治療 動物實驗有進展	幹細胞治巴金森氏症 美動物實驗成功
0.701068	(引 導, 麵 包, 華 姆	人體組織是麵包 幹細胞就	幹細胞療效 可能還需

	斯, 麵 粉, 湯 姆 森)	是麵粉 - 幹細胞經引導可長成新而健康的細胞 ..	十年
0.729545	(幼 年 型 糖 尿 病, 以 色 列, 胰 島 素, 美 國 糖 尿 病 協 會, 耶 路 撒)	幼年型糖尿病 有疫苗防治 - 以國新突破 證實可減輕病情 避免遺傳 3 年後 ...	人類幹細胞製造胰島素 有譜 - 以色列研究露曙光 幼年型糖尿病患新生 ...
0.706214	(巴 金 森 氏 症, 伊 薩 科 森, rotenone, 殺 蟲 劑, 麥 凱)	殺蟲劑 可引發巴金森氏症	幹細胞治巴金森氏症 美動物實驗成功
0.709000	(輪 部, 角 膜, 眼 睛, 自 體 移 植, 上 皮)	隱形眼鏡族福音 - 角膜長期缺氧易病變 細胞自體移植成功率百分之百	羊膜培育眼睛輪部幹細胞 讓角膜再生 - 長庚醫院眼科部首創經兩年多 ...

大致而言，觀察文件標題，直覺上相似程度很高，但爲何「被歸類在一起，且相似度極高的文件，卻無法同被判定爲相關？」，如一案例文件相似度達 0.7397，內容如下。

表 4-8：

和判斷文件相似度高但被判斷爲非相關文件之內容

	判斷相關文件	判斷非相關文件
文件標題	幹細胞治巴金森氏症 美動物實驗成功	巴金森氏症治療 動物實驗有進展
文件內容	美聯社舊金山十六日電 研究人員十六日說，科學家快要用植入胚胎幹細胞的方法治療巴金森氏症了，不過這項新療法何時才能展開人體試驗，必須視懸而未決的一些政治議題何時有定案。哈佛醫學院的伊薩科森博士與美國衛生研究院的麥凱博士十六日說，他們兩人都已使用從實驗室	田思怡譯自法新社 美國、法國和瑞士研究人員已成功的在猴子實驗中，以基因治療法更換與巴金森氏症有關的腦部細胞。研究人員在猴子腦部直接注射一種特殊病毒，用來產生神經膠質衍生的親神經性病毒（glial-derived neurotrophic fac-tor virus,GDNF ）的基因。他們發現，

71

	中的動物胚胎取出的幹細胞，「治癒」了老鼠的巴金森氏症。伊薩科森在美國科學促進會舉行的全國會議中說，老鼠的胚胎細胞經過仔細處理，可以植入老鼠的腦部，取代被巴金森氏症殺死的細胞。伊薩科森說：「老鼠實驗顯示，這些細胞的功能恢復了。」麥凱和伊薩科森說，研究人員已經在摩拳擦掌，準備展開人體試驗，但是，在此之前，必須先解決美國社會與政治的議題。	GDNF 保護並強化那些因巴金森氏症惡化而死亡的腦部細胞，並刺激細胞產生多巴胺。多巴胺爲一種神經傳遞化學物質，喪失這種物質會導致巴金森氏症。

然而，此篇文件在人工判斷時，其限制規則爲：「有關胚胎幹細胞的應用方式、醫學貢獻，以及倫理爭議的介紹視爲相關。科學家的研究過程則視爲無關」，因此單用 Title 是難以分別此問題。

三、PRF 查詢擴展

(一) 實驗數據

PRF 查詢擴展，本研究使用前初次排序前六篇較相關的文件，每篇文件中最佳的前十五個詞，送入檢索系統。實驗結果如下所示：

表 4-9：

PRF 查詢擴展實驗結果

欄位		Rigid	Relax
原架構	Title	0.2691	0.3229
	Description	0.2379	0.2912

使用後	Title	0.315	0.3733
	imp	17.1%	15.6%
	Description	0.3228	0.3892
	imp	35.7%	33.7%

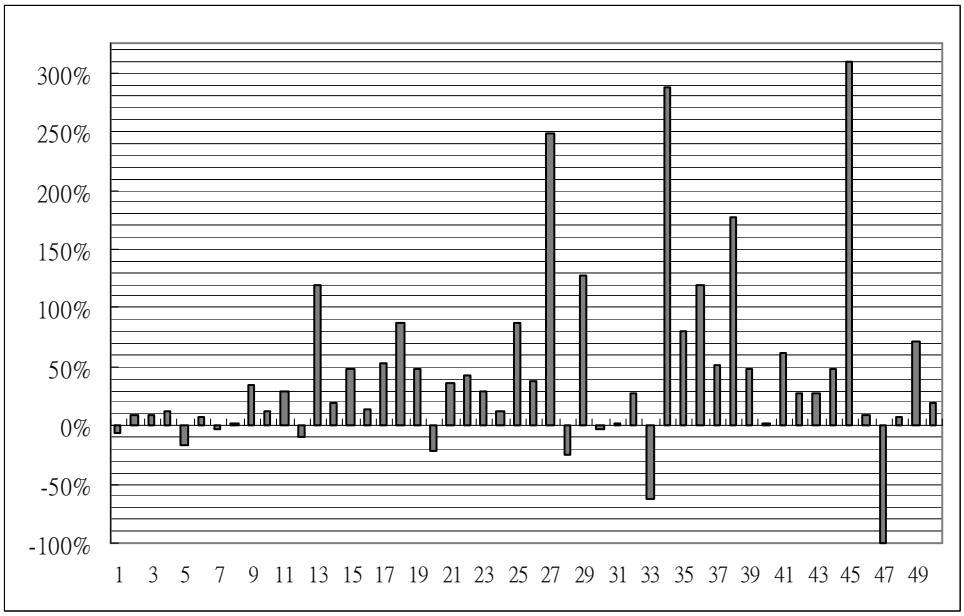


圖 4-11：PRF 查詢擴展各題查詢問題成效提升(imp)圖(Description-run/rigid)

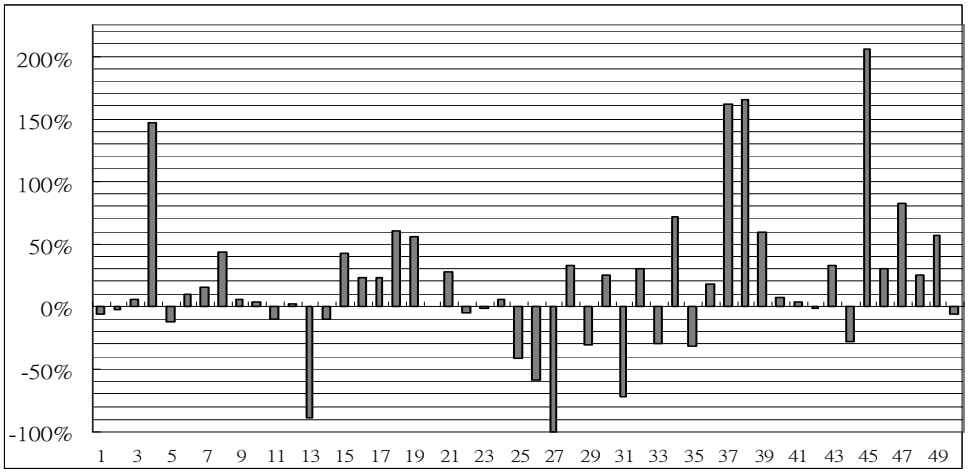


圖 4-12：PRF 查詢擴展各題查詢問題成效提升(imp)圖(Title-run/rigid)

查詢 Description 欄位方面，大部分的題目皆能有效的提升成效，只有 9 個題目降低了成效，而且下降的 MAP 皆控制在-0.06 而已，影響不大。單題成長幅度

73

方面，第 45、34 題由於原查詢分數較低，不具代表性(MAP_{45}:0.0011 -> 0.0045, MAP_{34}:0.0165->0.064)，因此以第 27 題成長最大，MAP 從 0.1689 提升到 0.5897。查詢 Title 欄位方面，共有 15 個題目產生成效下滑的情況，較爲嚴重的爲第 13 題，但 MAP 也只下降 0.0939 而已。成效提升以第 4 題效果最佳(MAP:+ 0.1630, imp:+ 147%)。

(二) 參數特性分析

從各方面的數據而言，以「PRF 查詢擴展」能有效且穩定的提升檢索成效。因此爲了分析目前使用 PRF 特性，以線性調整「PRF 查詢擴展」參數，試著分析參數和 PRF 的特性。實驗方式爲：

1. 控制變數「初次查詢前 6 篇文件」，將每篇可送出的詞，從 10、20、30 40、…200 進行調整，觀察和 MAP 是否存有正向關係。
2. 控制變數「每一篇送出 15 個詞」，調整需要的文件數，從 4、5、6、…10，進行調整，觀察和 MAP 是否存有正向關係。
3. 假設以上調整和 MAP 皆有正向關係，將從以上實驗挑選最佳組合驗證。
4. 將最佳組合驗證於其他文件集。

1. 初次查詢前 6 篇文件，每篇送出 X 詞

表 4-10：

前 6 篇 x 個最佳詞與 MAP 分佈情況

X	T-rigid	imp	T-relax	imp	D-rigid	imp	D-relax	imp
15	0.3126		0.3682		0.3163		0.3796	
10	0.3079	-1.50%	0.3683	0.03%	0.3128	-1.11%	0.3710	-2.27%
20	0.3166	1.28%	0.3690	0.22%	0.3170	0.22%	0.3835	1.03%
30	0.3149	0.74%	0.3731	1.33%	0.3237	2.34%	0.3899	2.71%
40	0.3177	1.63%	0.3758	2.06%	0.3324	5.09%	0.3941	3.82%
50	0.3183	1.82%	0.3756	2.01%	0.3344	5.72%	0.3943	3.87%

60	0.3237	3.55%	0.3795	3.07%	0.3357	6.13%	0.3942	3.85%
70	0.3233	3.42%	**0.3800**	**3.20%**	0.3366	6.42%	0.3955	4.19%
80	0.3219	2.98%	0.3797	3.12%	0.3389	7.15%	0.3963	4.40%
90	0.3222	3.07%	0.3795	3.07%	0.3402	7.56%	0.3980	4.85%
100	0.3222	3.07%	0.3790	2.93%	0.3395	7.33%	0.3961	4.35%
110	0.3259	4.25%	0.3797	3.12%	0.345	9.07%	**0.4012**	**5.69%**
120	0.3263	4.38%	0.3795	3.07%	**0.3455**	**9.23%**	0.4006	5.53%
130	0.3263	4.38%	0.3792	2.99%	0.3421	8.16%	0.3972	4.64%
140	0.3250	3.97%	0.3778	2.61%	0.3433	8.54%	0.3984	4.95%
150	0.3248	3.90%	0.3767	2.31%	0.3432	8.50%	0.3987	5.03%
160	0.3262	4.35%	0.3776	2.55%	0.3432	8.50%	0.3984	4.95%
170	0.3264	4.41%	0.3773	2.47%	0.343	8.44%	0.3979	4.82%
180	**0.3265**	**4.45%**	0.3772	2.44%	0.3423	8.22%	0.3973	4.66%
190	0.3263	4.38%	0.3771	2.42%	0.342	8.13%	0.3967	4.50%
200	0.3263	4.38%	0.3771	2.42%	0.3417	8.03%	0.3961	4.35%

送入的詞彙愈多，成效也不因此下滑，表現出 PRF 機制成效穩定。送入查詢詞愈多，雖然成長穩定，但電腦運算會較長，因此可以有所取捨。將成長程度繪製圖形，可以明確的了解成長情況。如下圖所示：

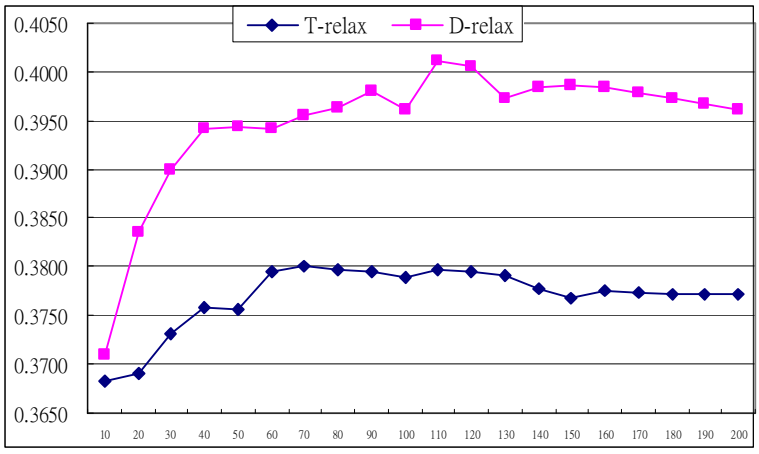


圖 4-13：前 6 篇 x 個最佳詞與 MAP (Relax)

75

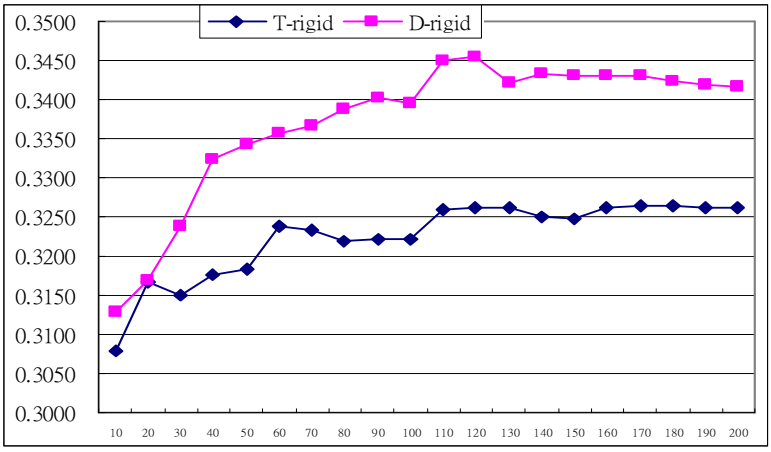


圖 4-14：前 6 篇 x 個最佳詞與 MAP(Rigid)

即使送入詞達到 200 個字時，也未能使成效明顯下降，整體而言，過去使用 15 個詞可能還有提升程度。總合以上兩圖，以成長度(imp)來計算如下圖所示：

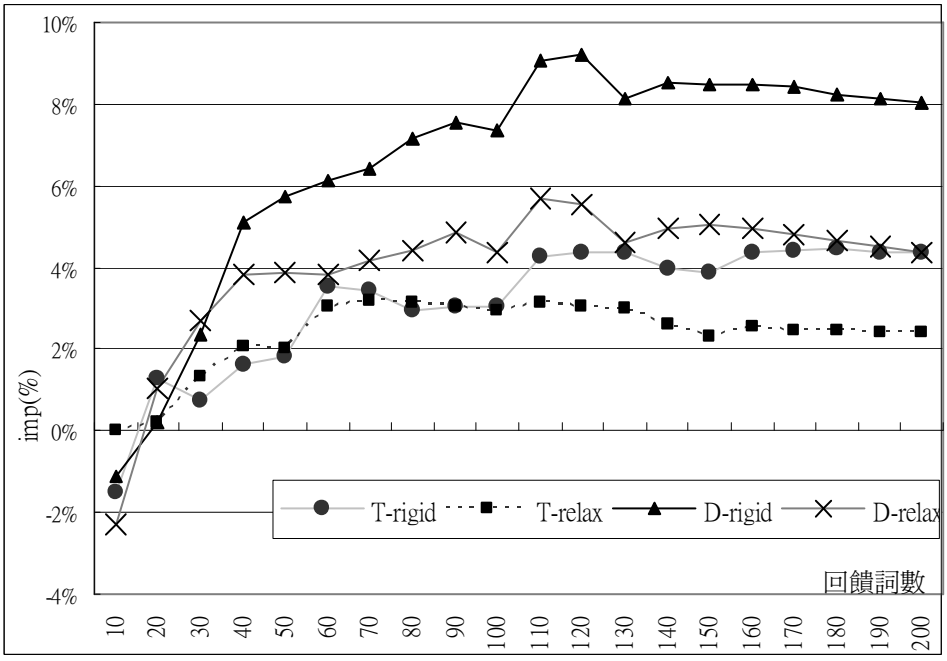


圖 4-14：成長度(imp)與回饋詞數

本研究結果，將曲線視爲二個部分：

1. Title-run 10~60 個詞；Description-run 10-40 個詞：增加幅度較高

2. Title-run 60 個詞之後；Description-run 40 個詞之後：成效穩定期

2. 初次查詢前 X 篇文件，每篇送出 15 詞

實驗結果如下表所示：

表 4-11：

最佳 15 個詞於前 X 篇文件 MAP 成效

X	T-Relax	T-Rigid	D-Relax	D-Rigid
4	0.3668	0.3057	0.3668	0.3057
5	0.3695	0.3082	0.3695	0.3082
6	0.3682	0.3126	0.3796*	0.3163
7	0.3714	0.3194	0.3714	0.3194
8	0.3771	0.3229	0.3771	0.3229
9	0.3776	0.3236*	0.3776	0.3236*
10	0.3782	0.3174	0.3782	0.3174
11	0.3822*	0.3197	0.378	0.3141
12	0.3776	0.3164	0.3782	0.3179
13	0.3784	0.3223	0.3752	0.314
14	0.3784	0.3231	0.3758	0.3158

註：以*表最大值，底線表原架構

差異不大，雖然篇數愈多，較 Title-run 可以提升些許成效，但分佈情況還是難以明確了解差異，如 Description-RUN 的 Relax 在第六篇時產生較不常態的增加，很有可能爲大部分的查詢問題在初次查詢時，第六篇文件詞較能檢索到正確文件。

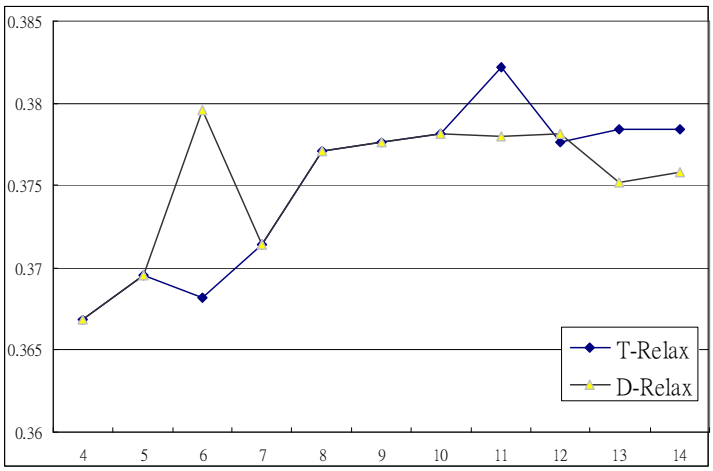


圖 4-15：最佳 15 個詞於前 X 篇文件 MAP(Relax)分佈圖

因此在本階段實驗上，未能明確看出使用多少篇文件回饋能大幅增加成效。

77

(三) NTCIR6 文件集實驗

表 4-12：

NTCIR6 文件集 PRF 穩定成長情況

單篇送入詞數	Title				Descrption			
	rigid		relax		rigid		relax	
	MAP	imp	MAP	imp	MAP	imp	MAP	imp
15	0.2283		0.3188		0.1988		0.2956	
30	0.2354	3%	0.3258	2%	0.2073	4%	0.307	4%
40	0.2317	1%	0.3225	1%	0.2058	4%	0.3042	3%
50	0.2346	3%	0.3252	2%	0.2045	3%	0.3056	3%
60	0.2351	3%	0.3243	2%	0.2054	3%	0.3054	3%
110	0.2303	1%	0.32	0%	0.2051	3%	0.3057	3%
120	0.2307	1%	0.3191	0%	0.2036	2%	0.3038	3%
130	0.2313	1%	0.3198	0%	0.2024	2%	0.3104	5%

即使查詢 Title 欄位成效成長不大，但送入的詞組較多，成效也能有所提升。

第三節 綜合評估

本研究數據和「提出該機制之研究團隊」(如：HKPU、I2R…)，結果差異較大，由於難以向該團隊要求其檢索系統使用情況，因此除了在該研究章節下，以細部觀察方式檢視問題所在。先行排除本研究實驗時程式是否有問題，在此，再進一步歸納其他可能的原因：

(1) 基礎檢索系統較適合進行該機制執行

由於研究團隊多在自行開發的檢索系統上，再進一步研究新的機制，而該機制也很有可能會較適合在該檢索系統上運行，也有可能因此能提升成效。雖然本研究已經盡可能客觀去實作該機制，也有可能如此問題發生。

(2) 其他參數調整以配合該機制執行

從本實驗觀察， I2R 的文件二次排序及 HKPU 標題二次排序，就公式而言，比較像是一個理想化的初步公式，因此常將相似度調整過當的情況發生。

以下部分再由另一個方向來評估機制可能的問題所在。

(3) 文件二次排序提升成效情況相較敏感

本研究完全取自於提出該機制團隊中的文件，評估、了解後，盡可能依其文獻敘述實作出來，在多次對照文獻無誤後，發現本實驗的情況和提出機制團隊的情況完全不同。既然完全客觀性的對照後，卻完全不同，那問題也可能發生於該機制較於敏感，易被小部分的差異所影響而產生完全不同的效果，例如：「初次查詢的相關文件排序情形」、「初始相似度差異」或「斷詞及詞彙選擇差異」等…。

雖然以上爲實驗分析後，得到一些數據及細部觀察來解釋此問題，但還是皆出於假設。

一、各檢索機制成效情況

本研究結果數據已經完成，分析以上兩節數據及分析，綜合比較於下：

表 4-13：

79

本實驗研究數據表

		Rigid	imp	Relax	imp
原架構	Title	0.2691		0.3229	
	Descript	0.2379		0.2912	
PRF 查詢擴展	Title	0.315	17.1%	0.3733	15.6%
	Descript	0.3228	35.7%	0.3892	33.7%
關鍵詞及文件特徵	Title	0.2486	-7.6%	0.3098	-4.1%
	Descript	0.2027	-14.8%	0.2606	-10.5%
文件標題	Title	0.1995	-25.9%	0.2512	-22.2%
	Descript	0.2027	-24.3%	0.2606	-26.2%
Label Propagation	Title	0.252	-6.4%	0.306	-5.2%
	Descript	0.2359	-0.8%	0.2929	0.6%
kNN	Title	0.2429	-9.7%	0.3	-7.1%
	Descript	0.2336	-1.8%	0.2906	-0.2%

繪成以下圖形，方便檢示：

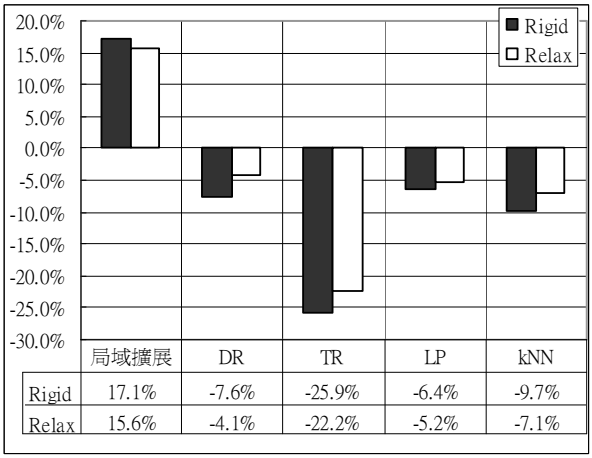
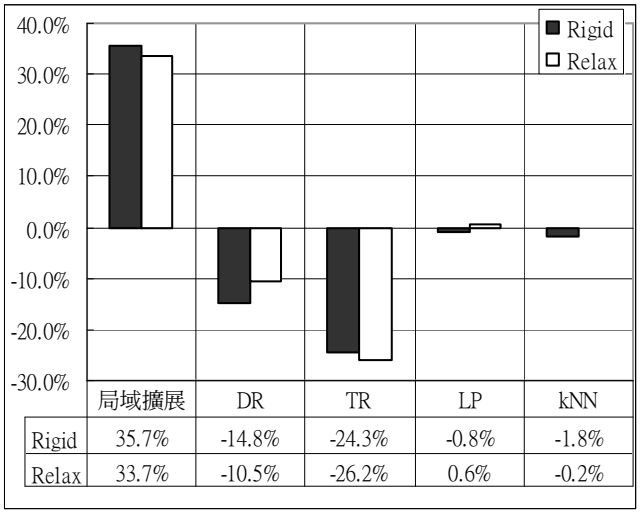


圖 4-16：Title run 的 imp 長條圖

上圖，DR 表示 I2R 文件二次排序(Document re-ranking)；TR 表 HKPU 標題二次排序(Title re-ranking)；LP 表示 Label Propagation 歸類演算法於文件二次排序；kNN 即爲 kNN 歸類演算法於文件二次排序。



	局域擴展	DR	TR	LP	kNN
Rigid	35.7%	-14.8%	-24.3%	-0.8%	-1.8%
Relax	33.7%	-10.5%	-26.2%	0.6%	-0.2%

圖 4-17：Description run 的 imp 長條圖

綜合以上結果，可知：

(1) 本實驗使用「PRF 查詢擴展」能大幅增加成效，而使用「文件標題二次排序」使文件成效下降最大。

(2) 文件二次排序機制皆無法有效的提升成效。

(3) 歸類二次排序比較方面，Label Propagation 較 KNN 穩定。

(4) 歸類二次排序，減少程度較不大。比較分析該模組特性，以視未來是否有研究發展空間。

81

二、歸類二次排序機制成效比較

了解以歸類進行二次排序機制是否有更進一步的研究意義，另一方面也是了解問題所在，本階段將用以下步驟查看機制的成效情形：

(1) 送入正確答案文件於「訓練文件組」及「回饋文件」中，讓該機制自行運算、歸類及查詢擴展，藉以了解理想情況下，該機制是否有效協助成效提升。

(2) 增加送入的正確文件篇數，進一步了解機制提升情況。

表 4-14：

各機制送入判斷正確文件數比較表

RunID	Title run		Description run	
	Rigid	Relax	Rigid	Relax
Bm25	0.2691	0.3229	0.2379	0.2912
Bm25+KNN(a)	0.2429	0.3	0.2336	0.2906
Bm25+LP(a)	0.252	0.306	0.2359	0.2929
Bm25+PRF	0.3126	0.3682	0.3163	0.3796
Bm25+KNN(1)	0.2690	0.3286	0.2540	0.3003
Bm25+KNN(2)	0.2983	0.3616	0.2752	0.3288
Bm25+KNN(3)	0.3261	0.3804	0.2974	0.3545
Bm25+KNN(4)	0.3337	0.3942	0.3102	0.3737
Bm25+KNN(5)	0.3369	0.3997	0.3215	0.3846
Bm25+LP(1)	0.2956	0.3498	0.2706	0.3163
Bm25+LP(2)	0.3189	0.3788	0.2863	0.3412
Bm25+LP(3)	0.3369	0.3915	0.3059	0.3617
Bm25+LP(4)	0.3415	0.4006	0.3108	0.3736
Bm25+LP(5)	0.3396	0.4049	0.3190	0.3828
Bm25+TRF(1)	0.3362	0.3893	0.3213	0.3742

Bm25+TRF(2)	0.3676	0.4243	0.3509	0.4067
Bm25+TRF(3)	0.3833	0.4461	0.3724	0.4363
Bm25+TRF(4)	0.3865	0.4480	0.3763	0.4458
Bm25+TRF(5)	0.3931	0.4542	0.3821	0.4479

註：(x)中，x=a 表示全自動檢索；x=1~5 表示送入正確文件 x 篇；TRF= true relevance feedback

分布情況如下圖所示：

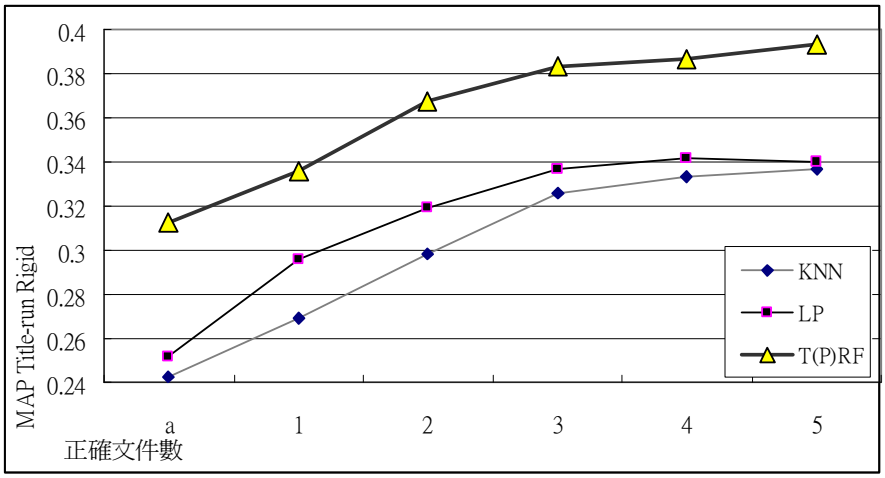


圖 4-18：送入正確文件數提升成效(Title-run/rigid)

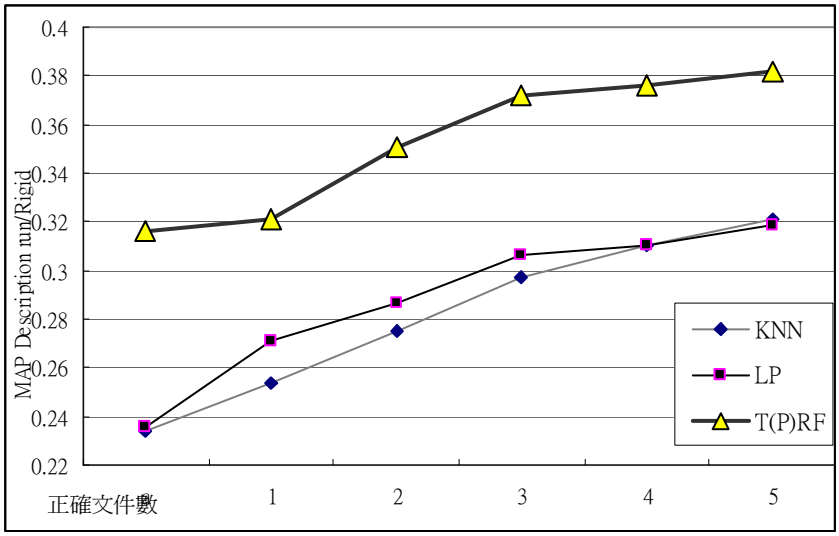


圖 4-19：送入正確文件數提升成效(Description-run/rigid)

83

觀察以上情況，可以發現：

(1) 以 TRF 查詢擴展所呈現的效果最佳，皆較另外二個機制要好。

(2) 在 Title-run 而言，須在「相關訓練文件組」中包含 2 篇以上的正確文件才能達到自動化的 PRF 水準；而 Description-run 就須包含 4~5 篇以上。

(3) 若 PRF 在有包含到 1-2 篇左右的文件，將能更加提升成效。(Tseng, et al., 2007)

第五章 結論與建議

本研究希望，以日本 NTCIR 檢索環境爲基礎，希望能評估出「研究成效佳」的機制、從其他研究團隊得到經驗及實際提升檢索系統成效，雖然最終實驗成效未如預期，但也得到了實驗結果及新的問題。最後，基於本研究結果之數據和分析，提出結論及建議，希望能供未來同質性研究作參考。

第一節 結論

(一) 文件二次排序提升成效情況相較敏感

文件二次排序方面，包含歸類法二次排序，在本研究實驗中，皆無法提升成效，而在該團隊的研究文獻中，反而能有效的提升成效，因此問題也可能發生於該機制較於敏感，易被少部分的差異所影響，而產生完全不同的效果，例如：「初次查詢的相關文件排序情形」、「初始相似度差異」或「斷詞及詞彙選擇差異」等…。

(二) PRF 查詢擴展能穩定的提升檢索成效

從實驗中，本研究以 PRF 查詢擴展和其他檢索方法比較，從各方面的比較結果，都以 PRF 表現情況較好，例如：「成效提升情況」，「各查詢問題成長情況」，「參數調整對穩定情況」，「理想查詢擴展檢視」等…。能看出理想的成效提升機制爲何，還更能看出 PRF 未來還有研究的空間。

(三) PRF 查詢擴展還有成效提升的空間

從最後第四章第三節的研究可以看出，如果 PRF 能再多回饋到 1~2 個以上「判斷相關」的文件，其成效提升程度會再更好，因此也可以成爲日後研究的借鏡。

85

第二節 建議

(一) 對於其他相關發表中的論文或會議文件所提出的機制，建議先行評估分析，再試著修正成對本身檢索系統合適的方式

從本研究的過程中可以了解，即使會議文件及其他技術文件中的實驗有較佳成效，而在複製該機制時，也未必能達到如此的成效，因此在使用該機制時，也需在進一步評估和分析，以修正對實驗檢索系統較佳的方式，不建議直接實際應用。

(二) 若檢索系統可以和使用者互動時，建議使用 PRF 查詢擴展

由本實驗可以得知，檢索系統使用 PRF 查詢擴展時，給予 1-2 篇正確文件，較其他檢索機制更能大幅提升成效，更能檢索到所需的文件，因此假使能和使用者有互動(例如：GOOGLE 查詢提示詞)時，使用 PRF 查詢擴展成效會較佳。

參考文獻

江玉婷(1999)。現行之重要資訊檢索測試集介紹/TREC。民國九十五年五月二日，取自：http://lips.lis.ntu.edu.tw/ytchiang/study/test/TREC.htm

江玉婷、陳光華(1998)。TREC 現況及其對資訊檢索研究之影響。國立臺灣大學圖書資訊學系，台北市。

陳光華（2001）。資訊檢索系統的評估－NTCIR 會議，*台灣大學圖書資訊學系四十週年系慶研討會*(頁 69-73)。

陳光華（2004）。資訊檢索的績效評估。*2004 年現代資訊組織與檢索研討會論文集*(頁 129-136)。

陳光華、莊雅棻（2001）。資訊檢索之中文詞彙擴展。*資訊傳播與圖書館學*，8-1，60-70。

陳光華、陳信希 (2004)。CIRB030 資訊檢索測試集簡介。*中華民國計算語言學學會通訊*。中華民國計算語言學學會，台北市。

陳光華、陳信希(2005)。跨語言資訊檢索與擷取測試集。民國九十六年六月二日，取自：http://www.csie.ntu.edu.tw/~ciet/form/paper/1.doc

陳致榮(2002)。<Organizing Knowledge>引言與摘要–第六章：前組合式標引與主題標目。民國九十五年五月二日，取自：http://research.pork.idv.tw/master/ok06.htm

曾元顯(1997)。關鍵詞自動擷取技術與相關詞回饋。民國九十五年五月二日，取自：http://www.lins.fju.edu.tw/~tseng/papers/feedback.htm

黃慕萱(1996)。*資訊檢索*。臺北市：臺灣學生書局。

葉至誠(2000)。*社會科學概論*。臺北市：揚智。

蔡育欽(2005)。查詢擴展之詞彙篩選應用於主題檢索之研究，碩士論文，私立輔仁

大學圖書資訊學系，台北縣。

Singhal, A., Salton, G., & Buckley C. (1996). Length Normalization in Degraded Text Collections, *Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval*(pp. 149-162).

Zhai, C., & Lafferty, J.(2001). A study of smoothing methods for language models applied to ad hoc information retrieval. *In Proceedings of SIGIR'01*(pp 334–342).

Guarino, N. (1997). Understanding, building and using ontologies. *International journal of human and computer studies*, 46(3/4), 219-310.

Fang, H., Tao, T., & Zhai, C. X. (2004). A Formal Study of Information Retrieval Heuristics.*Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 49-56). U.K: Sheffield.

Jaime, C., & Jade, G.(1998). The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval.*(pp. 335-336). Australia: Melbourne.

Chen, J., Rowena, Li, & Li, F. (2005). Chinese Information Retrieval Using Lemur: NTCIR-5 CIR Experiments at UNT. *Proceedings of NTCIR-5 Workshop Meeting*. Japan: Tokyo.

Min, J., Sun, L. & Zhang, J. (2005). ISCAS in English-Chinese CLIR at NTCIR-5. *Proceedings of NTCIR-5 Workshop Meeting*, Japan: Tokyo.

Yang, L., Ji, D. & Tang, L.(2004). Chinese Information Retrieval Based on Terms and Ontology. *Working Notes of NTCIR-4*, Japan: Tokyo.

Yang, L., Ji D., Zhou, G., Nie, Y., & Xiao, G. (2001). Document re-ranking using cluster validation and label propagation. *Proceedings of the 15th ACM international conference on Information and knowledge management CIKM '06*(pp. 690– 697).

Singhal, A.(2001). Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*(24(4), pp. 35–43).

Mitra, M., Singhal, A., &. Buckley, C.(1998). Improving Automatic Query Expansion. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp.206-214)

NTCIR (2006). Overview. Retrieved MAY 4, 2006, from http://research.nii.ac.jp/~ntcadm/outline/prop-en.html

Ricardo, B. Y., & Berthier, R. N.(1999). *Modern Information Retrieval*. New York: Addison Wesley.

Robertson, S. E. (1990). On term selection for query expansion. *Journal of Documentation*, 46, 359-364.

Robertson, S. E., & Sparck Jones, K.(1994). Simple, Proven Approaches to Text Retrieval.Computer Laboratory, University of Cambridge.

Robertson, S. E., & Walker, S. (1994). Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*(pp. 232-241). Ireland:Dublin.

Sakai, T., Kajiura, M., & Sumita, K. (1998). Generation and Evaluation of Search Queries using Boolean Expressions and Document Structure for Information Filtering (in Japanese). *IPSJ Journal*, 39(11), 3076–3083.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted Document Length Normalization. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrievl* (pp.21-29). Zurich: ACM SIGIR.

Fujtta S. (2005). A Decade after TREC-4 NTCIR-5 CLIR-J-J Experiments at Yahoo! Japan. *Proceedings of NTCIR-5 Workshop Meeting*, Japan: Tokyo.

TERC (2004). Overview. Retrieved MAY 4, 2006, from http://trec.nist.gov/overview.html

TERC (2007). TREC Tracks. Retrieved July 17, 2007, from http://trec.nist.gov/tracks.html

He, T., Qu, G., Tu, X., & JI, D.(2004). Chinese Information Retrieval Based on Related

Term Group, *Proceedings of NTCIR-5 Workshop Meeting*. Japan: Tokyo.

Xiao,Y., Luk, R.W.P., Wong, K.F., & Kwok, K.L. (2005). *Some Experiments with Blind Feedback and Re-ranking for Chinese Information Retrieval*. Proceedings of NTCIR-5 Workshop Meeting, Japan: Tokyo.

Yang, L, & Ji, D.(2005). I2R at NTCIR5 . *Proceedings of NTCIR-5 Workshop Meeting*, Japan: Tokyo.

Tseng, Y. H., Tsai, C. Y. & Chuang,C. J. (2007). On the Robustness of Document Re-Ranking Techniques: A Comparison of Label Propagation, KNN, and Relevance Feedback. *Proceedings of the Sixth NTCIR Workshop on Research in Information Access Technologies - Cross-Lingual Information Access*, Japan: Tokyo.

Tseng, Y. H., Tsai. Y. C., & Lin C. J. (2005). Comparison of Global Term Expansion Methods for Text Retrieval. *Proceedings of NTCIR-5 Workshop Meeting*, Japan: Tokyo.

附錄 A　NTCIR 5 CLIR 中文查詢問題

題號	查詢問題	
	Title	Description
001	時代華納，美國線上，合併案，後續影響	時代華納與美國線上合併案的後續影響。
002	祕魯總統，藤森，醜聞，賄選	祕魯總統藤森在 2000 年總統大選行賄曝光，並遭到國會罷免流亡海外之相關報導。
003	金大中，金正日，南北韓高峰會	兩韓領導人金大中與金正日於平壤舉行高峰會的相關報導。
004	美國防部長，柯恩，北京	美國國防部長柯恩於 2000 年 6 月訪問北京的相關報導。
005	G8，八國集團高峰會	2000 年在日本琉球名護市展開的 G8 高峰會之相關報導。
006	科斯克號，潛艇意外，國際救援行動	俄羅斯核子潛艇科斯克號意外沉沒及等待救援的相關報導。
007	李文和案，核武機密，國家安全	華裔科學家李文和涉嫌竊取美國洛薩拉摩斯實驗室核武機密的相關報導
008	我愛你，電腦病毒	電腦病毒「我愛你」造成個人電腦大癱瘓的相關報導。
009	地震，國際救援	全球各地發生大地震之災情與善後情形、國際救援協助之相關報導。
010	炭疽熱，細菌戰，恐怖攻擊	全球陷入炭疽熱病菌恐怖攻擊的疑雲以及後續反應之相關報導。
011	鈴木一朗，新人王，美國職棒大聯盟	鈴木一朗轉會至美國職棒大聯盟第一年的表現之相關報導。
012	卡普莉雅蒂，網球	美國網球女將卡普莉雅蒂成功復出拿下多項大賽冠軍並一度躍升 WTA 世界排名第一之相

91

		關報導。
013	神學士，塔利班，滅佛	阿富汗神學士政權摧毀境內大型佛像之相關報導
014	奈米技術	奈米技術的研究發展與應用之相關報導
015	EP-3 偵察機，殲八(F-8)戰機，美中關係	美國 EP-3 偵察機與中國殲八戰鬥機於海南島上空發生擦撞意外之相關報導。
016	歷史教科書爭議，二次世界大戰	日本文部省審核之二次世界大戰相關歷史教科書所引發的爭論之相關報導。
017	印度與巴基斯坦領土衝突，核武	印度與巴基斯坦的軍備競賽與領土衝突之相關報導。
018	菸草商，訴訟，賠償	菸草商遭受控告並求賠償金之相關報導。
019	超音速飛機，協和號，墜機	已 30 年未發生飛行意外的法國超音速飛機協和號墜機事故的原因與造成的損害。
020	變性，青蛙，魚	關於內分泌干擾物質的真相調查中，所發現的青蛙和魚令人憂慮的變化。
021	諾貝爾和平獎，金大中	金大中總統成爲亞洲第七位諾貝爾和平獎得主。他獲獎的原因。
022	狂牛症	不同國家對於防止狂牛症擴散所採取的緊急措施。
023	太空站，和平號，廢物儲存裝置的廢棄處理	預計於 2001 年 2 月被廢棄的和平號太空站，相關的廢棄決定和廢棄方法。
024	經濟艙，綜合症候群，航班	客機中經濟艙綜合症候群的損害。
025	老虎伍茲，運動明星	運動媒體或運動相關產業表彰老虎伍茲爲運動明星的報導。
026	捐獻，百萬富翁，遺產	像 Holingswiss 一樣，將所有財產捐獻，然後空手離開人世間的百萬富翁的一生。
027	長壽，秘訣，安東尼	像義大利的安尼歐・陶德一樣的人瑞的長壽秘

	歐・陶德	訣。
028	布卡，鳥人，退休	六度獲得世界錦標賽冠軍的 37 歲烏克蘭選手謝爾蓋・布卡的退休報導。
029	替代能源，空氣污染，電力	對環境有利，能產生電力的替代能源的開發情形。
030	太空觀光，國際太空站，丹尼斯・帝托	抵達國際太空站的首位太空觀光客丹尼斯・帝托或是其他未來的太空旅遊。
031	細微塵粒，心臟疾病	查詢因爲威脅城市居民的細微塵粒所引發的心臟相關疾病的高死亡率調查。
032	世界人口，人口預測	21 世紀世界人口變化情形。
033	愛滋病，基金，感染率	致力於降低愛滋病傳染率的特定基金或正在籌組的基金。
034	賓拉登，美國，軍事手段	美國追捕反美恐怖份子首領賓拉登的策略。
035	死刑，調查資料	討論死刑的調查或投票結果。
036	遠端遙控手術，機器人	使用機器人取代醫生實際接觸病人，遠端使用醫療系統或執行衛生保健等工作。
037	森內閣，支持度，愛媛丸	2001 年時，因森喜朗首相處理「愛媛丸」意外事件的方式，所造成的內閣支持度改變。
038	科索沃危機，北大西洋公約組織，聯合國	描述北大西洋公約組織在科索沃衝突的轟炸事件，以及聯合國所採取的行動。
039	Windows，Linux，競爭	微軟視窗和 Linux 在個人電腦作業系統上的競爭關係。
040	哈利波特，銷售量	超級暢銷書，哈利波特在全世界的銷售情形。
041	雪印乳業，乳製品，食物中毒	因雪印乳製品所造成的大規模食物污染案例的背景描述。
042	葛林斯班，貨幣政策	被認爲對於經濟復甦有貢獻的美國聯邦準備理事會主席葛林斯班所提出的貨幣政策。

93

043	START 2，俄羅斯，批准	俄羅斯批准 STAR2 的背景。
044	氣候異常，災害，造成	因氣候異常所帶來的災害。
045	人口問題，飢餓	和飢餓有關的人口問題的報導。
046	東帝汶，獨立，東帝汶人抗爭國家委員會	關於東帝汶獨立事件，東帝汶人抗爭國家委員會在聯合國東帝汶過渡行政當局開始運作後，所面臨的問題。
047	韓國大選，2000 年，大國家黨	韓國大國家黨在 2000 大選的結果。
048	基因改造食物，管理規則	世界各國爲保障食物安全，針對基因改造食物所制訂的管理規則。
049	野生動物，農作物，損害	世界各地有關野生動物損害農作物的報導。
050	人類基因，解碼，醫藥業	醫療及製藥業會因爲人類基因解碼計畫的進行而有所改變。